Advances in Experimental Medicine and Biology 736

Igor I. Goryanin
Andrew B. Goryachev *Editors*

Advances in Systems Biology

Springer

ADVANCES IN EXPERIMENTAL MEDICINE AND BIOLOGY

For further volumes:
http://www.springer.com/series/5584

Igor I. Goryanin • Andrew B. Goryachev
Editors

Advances in Systems Biology

Springer

Editors
Igor I. Goryanin
School of Informatics
The University of Edinburgh
Edinburgh, UK
Igor.Goryanin@ed.ac.uk

Andrew B. Goryachev
Centre for Systems Biology
School of Biological Sciences
The University of Edinburgh
Edinburgh, UK
andrew.goryachev@ed.ac.uk

ISSN 0065-2598
ISBN 978-1-4419-7209-5 e-ISBN 978-1-4419-7210-1
DOI 10.1007/978-1-4419-7210-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011943082

Preface

Systems biology takes a holistic view on biology and aims at elucidating design principles of whole biological systems rather than characterizing individual molecules or single events. It is generally believed that systems biology will transform biology from a descriptive to a predictive science, making it possible to understand, explain, and eventually engineer complex biological systems. In the past decades, we witnessed burgeoning development of various fields that mutually complement each other and together define the scope and methods of systems biology. This young and rapidly growing consortium of disciplines defies all attempts at rigid definition of its purpose and boundaries while continuing to evolve and develop new experimental tools and theoretical paradigms. Perhaps, the most definitive characteristic feature of systems biology is that it is a fundamentally interdisciplinary science that became a point of fusion of the traditional experimental biology with physics, chemistry, mathematics, computer science, and engineering. Inevitable cross-talk of distinct cultures, often a tumultuous and never an easy process, brought about the emergence of a new culture of modern quantitative biology.

The most recent advances and new developments in systems biology were presented and actively discussed at the 11th International Conference on Systems Biology which took place on 10–16 October, 2010 in Edinburgh. This meeting marked the tenth anniversary of the increasingly popular series of conferences initiated by Hiroaki Kitano in 2000 in Tokyo. The meeting in Edinburgh attracted the largest yet attendee number, which is sure to continue growing in the years to come. Reflecting the highly diverse interdisciplinary nature of systems biology, the scientific programme of the Conference featured eight plenary and 16 parallel sessions aiming at the fair representation of various contributing fields. As has become the tradition over the decade of ICSB conferences, particular attention was given to the developments in genomics, proteomics, metabolomics as well as mathematical modeling and computational tools. Special sessions were dedicated to the recent advances in neurobiology, biological rhythms and circadian clocks, and biological noise and cellular decision making. Strong emphasis was also given to the practical applications of systems biology in medicine, biotechnology, and

pharmaceutical industry. Following the trend of the previous meetings, ICSB 2010 witnessed continuously increasing coalescence of experimental and theoretical approaches that resulted in exciting, truly systems research projects presented at the Conference.

The present collection of articles has emerged from the contributions provided by the speakers of ICSB 2010 as well as by other leaders of systems biology who could not attend the meeting. As the biological systems themselves, this volume is the result of self-organization. Since each contributor chose the topic of their chapter independently from the others, the scope of this volume is a faithful and unbiased replica of the entire breadth and diversity of systems biology. At the same time, individual contributions naturally grouped together revealing the particularly exigent research directions that presently attract the most attention. These emergent clusters defined the sections of the present volume. Thus, traditionally strong interest remains focused on the identification, analysis, and modeling of networks that represent causative, correlative, and other relationships between various biological entities. Contributions by B. Andrews, J. Saez-Rodrigues, D. Armstrong, and their colleagues consider the use of the proteome-wide datasets as well as the development of high-throughput techniques for their acquisition. Chapters by B. Kholodenko and W. Kolch, E. Feliu, S. Schnell and their co-workers are devoted to the analysis and modeling of intracellular signaling networks. H. Kaltenbach and J. Stelling discuss in more abstract terms the theoretical aspects of modularity that is characteristic of biological networks.

Much interest is presently devoted to the understanding of cellular decision making, such as response and adaption to the environmental perturbations, cellular differentiation, and programmed cell death. Given the importance of these fundamental biological processes for the treatment of cancer and stem-cell-based regenerative technologies, to name just a few applications, this interest is well justified. Section 2 starts with a provocative discussion feature by D. Bray who posits that biological organisms, as simple as unicellular bacteria, carry acquired throughout the evolution information on optimal environmental conditions. The contributions by A. Levchenko, J. Fisher, D. Lutter, and others focus on cellular differentiation and apoptosis. Together they suggest that systems biology is finally getting into the position to tackle these exciting and exceptionally complex problems.

Section 3 considers spatial and temporal aspects of intracellular dynamics. Thus, D. Vavylonis and colleagues and A. Carlsson discuss systems properties of actin cytoskeleton, while M. Enculescu and M. Falke review modeling of morphodynamic phenotypes and dynamic regimes of cellular locomotion. More technically oriented contributions that present novel computational algorithms, software tools and theoretical methods are grouped into Sect. 4. Here E. Balsa-Canto, I. Sbalzarini, and their colleagues discuss global optimization and parameter identification in stochastic reaction networks. M. Blinov and I. Moraru present the rule-based modeling approaches that allow building larger models of complex reaction networks.

To conclude the volume, Sect. 5 discusses a broad spectrum of systems biology applications in medicine, biotechnology, and pharmaceutical industry. Discussion features by R. Phair, L. Kupfer, N. Benson, and their colleagues present the views from inside the industry on the advantages and pitfalls associated with the use of systems biology in drug design and development. Other contributors showcase practical applications of systems methods to the analysis of patient data and typical problems arising in biotechnology of microorganisms and livestock.

Finally, the Editors would like to express their sincere gratitude to Mrs. Fiona Clark who provided invaluable administrative support without which the effort of assembling this volume would be impossible.

Andrew B. Goryachev
Igor I. Goryanin

Contents

Contributors

Tobias Abenius Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, 412 96 Gothenburg, Sweden

Brenda Andrews The Donnelly Centre, Department of Molecular Genetics, University of Toronto, Toronto, Canada

Miguel Aon Johns Hopkins University, School of Medicine, 720 Rutland Avenue, 1059 Ross Building, Baltimore, MD 21205, USA

J. Douglas Armstrong School of Informatics, University of Edinburgh, Edinburgh, UK

Petri Auvinen Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Richard A. Baldock MRC Human Genetics Unit, MRC Institute of Genetic and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK

Eva Balsa-Canto (Bio)Process Engineering Group, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain

J.R. Banga (Bio)Process Engineering Group, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain

Matteo Barberis Humboldt University Berlin, Institute for Biology, Invalidenstr. 42, 10115 Berlin, Germany

Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Emmanuel Barillot U900 INSERM/Institut Curie/Ecole de Mines, Institut Curie, 26 rue d'Ulm, Paris 75005, France

Verena Becker Division Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany

Bioquant, Heidelberg University, Germany Present address: Department of Systems Biology, Harvard Medical School, Boston, MA, USA

N. Benson Modelling and Simulation, Department of Pharmacokinetics, Dynamics and Metabolism, Pfizer Worldwide Research, Pfizer Ltd., Sandwich CT13 9NJ, UK

Antje Beyer Department of Genetics, University of Cambridge, Cambridge, UK

Michael L. Blinov Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA

H.M.T. Boer Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands

Adaptation Physiology Group, Department of Animal Sciences, Wageningen University, Wageningen, The Netherlands

Dennis Bray Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK

Yifat Brill-Karniely Department of Chemistry, University of Cambridge, Cambridge, UK

Philipp Bruns Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum München, Munich, Germany

Department of Surgery, Technische Universität München, Munich, Germany

Albert Burger MRC Human Genetics Unit, MRC Institute of Genetic and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK

Department of Computer Science, Heriot-Watt University, Edinburgh EH14 4AS, UK

Laurence Calzone U900 INSERM/Institut Curie/Ecole de Mines, Institut Curie, 26 rue d'Ulm, Paris 75005, France

Anders E. Carlsson Department of Physics, Washington University, St. Louis, MO 63130, USA

Natalie L. Catlett Selventa, One Alewife Center, Cambridge, MA 02140, USA

Yolanda T. Chong The Donnelly Centre, Department of Molecular Genetics, University of Toronto, Toronto, Canada

Matthias Chung Department of Mathematics, Texas State University, 601 University Drive, San Marcos, TX 78666, USA

Allan Clark Centre for Systems Biology at Edinburgh, The University of Edinburgh, Edinburgh EH9 3JU, Scotland, UK

Jukka Corander Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Jole Costanza University of Catania, Viale A. Doria 6, Catania, CT 95125, Italy

Michael J. Cox The Donnelly Centre, Department of Molecular Genetics, University of Toronto, Toronto, Canada

L. Cucurull-Sanchez Modelling and Simulation, Department of Pharmacokinetics, Dynamics and Metabolism, Pfizer Worldwide Research, Pfizer Ltd., Sandwich, CT13 9NJ, UK

Tine Curk Department of Chemistry, University of Cambridge, Cambridge, UK

Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia

Renée Deehan Selventa, One Alewife Center, Cambridge, MA 02140, USA

O. Demin Institute for Systems Biology, Leninskie Gori, Moscow 11992, Russia

Jure Dobnikar Department of Chemistry, University of Cambridge, Cambridge, UK

Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia

Tyler Drake Department of Physics, Lehigh University, Bethlehem, PA 18015, USA

David A. Drubin Selventa, One Alewife Center, Cambridge, MA 02140, USA

Ralf Eberhard Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

J.A. Egea Department of Applied Mathematics and Statistics, Technical University of Cartagena (UPCT), Cartagena, Spain

Thomas Eissing Systems Biology and Computational Solutions, Bayer Technology Services GmbH, Building 9115, 51368 Leverkusen, Germany

Mihaela Enculescu Institute for Theoretical Physics, Technische Universität Berlin, Hardenbergstr. 36, 10623 Berlin, Germany

Martin Falcke Mathematical Cell Physiology, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany

Andrew P. Feinberg Department of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA

Elisenda Feliu Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

A. Fernandez-Villaverde (Bio)Process Engineering Group, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain

Jasmin Fisher Microsoft Research, Cambridge, UK

Simon Fourquet U900 INSERM/Institut Curie/Ecole de Mines, Institut Curie, 26 rue d'Ulm, Paris 75005, France

Brian P. Frushour Selventa, One Alewife Center, Cambridge, MA 02140, USA

David A. Fryburg Selventa, One Alewife Center, Cambridge, MA 02140, USA

Vashti Galpin Laboratory for Foundations of Computer Science, The University of Edinburgh, Edinburgh EH8 9AB, Scotland, UK

Stephen Gilmore Centre for Systems Biology at Edinburgh, The University of Edinburgh, Edinburgh EH9 3JU, Scotland, UK

Britta Göbel Graduate School for Computing in Medicine and Life Sciences, Institute of Mathematics and Image Computing, University of Lübeck, Maria-Goeppert-Strasse 1a, 23562 Lübeck, Germany

Peddinti V. Gopalacharyulu VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

David de Graaf Selventa One Alewife Center, Cambridge, MA 02140, USA

P. van der Graaf Pfizer, Pharmacometrics, Global Clinical Pharmacology, Walton Oaks, KT20 7NS, UK

Dario Greco Department of Bioscience and Nutrition, Karolinska Institutet, 141 83 Huddinge, Stockholm, Sweden

Maria Luisa Guerriero Centre for Systems Biology at Edinburgh, The University of Edinburgh, Edinburgh EH9 3JU, Scotland, UK

Alex Hajnal Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Michael O. Hengartner Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

G.M. de Hijas-Liste (Bio)Process Engineering Group, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain

Jane Hillston Centre for Systems Biology at Edinburgh, The University of Edinburgh, Edinburgh EH9 3JU, Scotland, UK

Aaron A. Van Hooser Selventa, One Alewife Center, Cambridge, MA 02140, USA

Pablo A. Iglesias Department of Electrical and Computer Engineering, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

Rebecka Jörnsten Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, 412 96 Gothenburg, Sweden

Hans-Michael Kaltenbach Department of Biosystems Science and Engineering, ETH Zurich, CH-4058 Basel, Switzerland

Atsushi Kamimura Institute of Industrial Science, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan

Boris Kholodenko Systems Biology Ireland, Conway Institute, University College Dublin (UCD), Belfield, Dublin 4, Ireland

I.N. Kiselev Institute of Systems Biology, Ltd, Novosibirsk, Russia

Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia

Teresia Kling Cancer Center Sahlgrenska, Institute of Medicine, Box 425, 415 30 Gothenburg, Sweden

Ursula Klingmüller Division Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany; Bioquant, Heidelberg University, Germany

Edda Klipp Theoretical Biophysics, Humboldt-Universität zu Berlin, Invalidenstr. 42, D-10115 Berlin, Germany

Michael Knudsen Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

Centre for Membrane Pumps in Cells and Disease (PUMPKIN), Aarhus University, Aarhus, Denmark

Tetsuya J. Kobayashi Institute of Industrial Science, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan

Walter Kolch Systems Biology Ireland, Conway Institute, University College Dublin (UCD), Belfield, Dublin 4, Ireland

Fedor A. Kolpakov Institute of Systems Biology, Ltd, Novosibirsk, Russia

Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia

Andeas Kremling Systems Biotechnology, Technical University of München, München, Germany

Lars Kuepfer Systems Biology and Computational Solutions, Bayer Technology Services GmbH, Building 9115, 51368 Leverkusen, Germany

Clemens Kühn Theoretical Biophysics, Humboldt-Universität zu Berlin, Invalidenstr. 42, D-10115 Berlin, Germany

E.O. Kutumova Institute of Systems Biology, Ltd, Novosibirsk, Russia

Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia

Daphna Laifenfeld Selventa, One Alewife Center, Cambridge, MA 02140, USA

I.N. Lavrik German Cancer Research Center (DKFZ), Heidelberg, Germany

Andre Levchenko Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

Antonio Di Lieto Neuroscience Centre, University of Helsinki, Helsinki, Finland

Erno Lindfors VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

Pietro Liò University of Cambridge, William Gates Bldg, 15 J J Thomson Avenue, Cambridge CB3 0FD, UK

Jörg Lippert Systems Biology and Computational Solutions, Bayer Technology Services GmbH, Building 9115, 51368 Leverkusen, Germany

David Lloyd Microbiology, School of Biosciences, Cardiff University, Main Building, P.O. Box 915 Cardiff CF10 3AT, Wales, UK

Jyrki Lötjönen VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

Dominik Lutter Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum München, Munich, Germany

Franziska Matthäus Center for Modeling and Simulation in the Biosciences (BIOMS), University of Heidelberg, Heidelberg, Germany

Jussi Mattila VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

Sofia D. Merajver Department of Internal Medicine and Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

Ion I. Moraru Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA

Christian L. Müller Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics, ETH Zurich, CH-8092 Zurich, Switzerland

Douglas B. Murray Institute for Advanced Biosciences, Keio University, Nipponkoku 403-1, Daihouji, Tsuruoka City, Yamagata 997-0017, Japan

Sven Nelander Cancer Center Sahlgrenska, Institute of Medicine, Box 425, 415 30 Gothenburg, Sweden

Giuseppe Nicosia University of Catania, Viale A. Doria 6, Catania, CT 95125, Italy

Matej Orešič VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

Alessio Papini University of Florence, Via La Pira, 4, Firenze, FI I-50121 Italy

Jennifer S. Park Selventa, One Alewife Center, Cambridge, MA 02140, USA

Antti Pesonen VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland

Robert D. Phair Integrative Bioinformatics Inc., Los Altos, CA 94024, USA

Nir Piterman Department of Computer Science, University of Leicester, Leicester, UK

Kui Qian Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Rajesh Ramaswamy Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics, ETH Zurich, CH-8092 Zurich, Switzerland

S. Röblitz Department of Numerical Analysis and Modeling, Computational Systems Biology Group, Zuse Institute Berlin (ZIB), Berlin, Germany

Julio Saez-Rodriguez EMBL-EBI and European Molecular Biology Laboratory (EMBL), Genome Biology Unit, D-69117 Heidelberg, Germany

José Sánchez Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, 412 96 Gothenburg, Sweden

Tapesh Santra Systems Biology Ireland, Conway Institute, University College Dublin (UCD), Belfield, Dublin 4, Ireland

Kalesh Sasidharan Institute for Advanced Biosciences, Keio University, Nipponkoku 403-1, Daihouji, Tsuruoka City, Yamagata 997-0017, Japan

Ivo F. Sbalzarini Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics, ETH Zurich, CH-8092 Zurich, Switzerland

Linnéa Schmidt Cancer Center Sahlgrenska, Institute of Medicine, Box 425, 415 30 Gothenburg, Sweden

Santiago Schnell Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

R.N. Sharipov Institute of Systems Biology, Ltd, Novosibirsk, Russia

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

S. Smirnov Institute for Systems Biology, Leninskie Gori, Moscow 11992, Russia

Oksana Sorokina School of Informatics, University of Edinburgh, Edinburgh, UK

Jörg Stelling Department of Biosystems Science and Engineering, ETH Zurich, CH-4058 Basel, Switzerland

C. Stötzel Department of Numerical Analysis and Modeling, Computational Systems Biology Group, Zuse Institute Berlin (ZIB), Berlin, Germany

Giovanni Stracquadanio The Johns Hopkins University, 217 Clark Hall, Baltimore, MD 21218, USA

Camille Terfve European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Fabian J. Theis Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum München, Munich, Germany

Jens Timmer BIOSS Centre for Biological Signalling Studies, Freiburg Institute for Advanced Studies, Institute of Physics, Center for Systems Biology, University of Freiburg, Freiburg, Germany

Masaru Tomita Institute for Advanced Biosciences, Keio University, Nipponkoku 403-1, Daihouji, Tsuruoka City, Yamagata 997-0017, Japan

Laurent Tournier INRA, Unit MIG (Mathématiques, Informatique et Génome), Domaine Vilvert, Jouy en Josas 78350, France

Renato Umeton Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Kiran Gireesan Vanaja Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

Dimitrios Vavylonis Department of Physics, Lehigh University, Bethlehem, PA 18015, USA

Carsten Wiuf Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

Centre for Membrane Pumps in Cells and Disease (PUMPKIN), Aarhus University, Aarhus, Denmark

H. Woelders Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands

Michelle L. Wynn Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

Eddy Yusuf Physics Department, Surya College of Education, Surya Research and Education (SURE) Center, Jln. Scientia Boulevard U7, Gading Serpong, Tangerang 15233, Indonesia

Andrei Zinovyev U900 INSERM/Institut Curie/Ecole de Mines, Institut Curie, 26 rue d'Ulm, Paris 75005, France

Part I
Multiscale Biological Networks: Identification, Modeling and Analysis

Chapter 1
Modular Analysis of Biological Networks

Hans-Michael Kaltenbach and Jörg Stelling

Abstract The analysis of complex biological networks has traditionally relied on decomposition into smaller, semi-autonomous units such as individual signaling pathways. With the increased scope of systems biology (models), rational approaches to modularization have become an important topic. With increasing acceptance of *de facto* modularity in biology, widely different definitions of what constitutes a module have sparked controversies. Here, we therefore review prominent classes of modular approaches based on formal network representations. Despite some promising research directions, several important theoretical challenges remain open on the way to formal, function-centered modular decompositions for dynamic biological networks.

1 Introduction

With the advent of high-throughput experimental techniques such as micro-arrays or mass-spectrometry, the complexity of biological networks became increasingly evident. Concomitantly, the search for "network design principles" became both feasible and necessary. The necessity stems from our inability to grasp and to meaningfully analyze networks of even moderate complexity without formal methods – based on mathematical modeling – and without some sort of "divide-and-conquer" approach to the analysis [18]. Note that the second aspect holds for formal and informal network analysis alike. While, despite inherent nonlinearities, the dynamics of small systems with a few state variables can sometimes still be successfully characterized, analyzing medium- and large-scale systems with potentially hundreds of states poses new challenges. As an example, a current model

H.-M. Kaltenbach (✉) • J. Stelling (✉)
Department of Biosystems Science and Engineering, ETH Zurich, CH-4058 Basel, Switzerland
e-mail: hans-michael.kaltenbach; joerg.stelling@bsse.ethz.ch

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_1,

Table 1.1 Overview of module definitions according to main characteristics and application areas

Class of approaches	Definition based on	Module overlaps	Function prediction	Network applications
Community detection	Structure	No	Guilt-by-association	Protein–protein interaction
Metabolic pathways	Function	Yes	Steady-state	Metabolic
Network motifs	Structure	Yes	Dynamics	Transcriptional, signaling
Retroactivity	Function	No	Dynamics	General
Monotone systems	Function	No	Dynamics	General

of the ErbB signaling pathway was taken that already comprises about 500 state variables [5]. Clearly, classical methods for the analysis of nonlinear systems such as phase-plane analysis will only work in the rarest of cases for such models.

Conceptual divisions of complex networks in biology are standard practice for many experimental or theoretical studies. This is consistent with a modular view on biological systems; they are constituted by semi-autonomous functional units performing specific functions. Much of the reasoning about biological entities (e.g., protein complexes) and functions (e.g., distinct signaling pathways) follows this notion. However, it was only recognized relatively recently that – similar to engineered systems – modularity could be a key to the quantitative understanding of large-scale networks [14, 22, 25].

Modularity has several potential implications for the systems analysis of biological networks:

- The decomposition of complex networks into manageable units, and their subsequent assembly, can allow us to comprehend large-scale systems.
- Corresponding modular concepts for mathematical modeling and formal analysis facilitate theoretical investigations in systems biology, for instance, in terms of parameter estimation.
- The large repertoire of available engineering methods and insights could lead to the identification of operating principles that are common to biological and engineering systems [7, 38].

Over the past years, we have seen the accumulation of *general* evidence for modularity in different areas of biology, ranging from molecular interaction networks inside cells to the structure of evolutionary processes [43]. Many *specific* aspects of the existence and implications of modularity in biology, however, are controversial [1, 27, 44]. These controversies are often rooted in different operational definitions of modules.

Here, we therefore review several promising approaches that were proposed to address the question of how to define and find suitable subsystems (or modules) that would allow a modular analysis of complex cellular networks. The overview of main concepts shown in Table 1.1 illustrates that the approaches differ in several dimensions such as the basis of the definition, and the scope of predictions and network applications.

As it is beyond the scope of this review to cover all module concepts that have been proposed, we focus on those approaches that start from a formal definition and operate on a formal representation – a mathematical model – of a cellular network. Moreover, we assume that it is the ultimate goal of modular approaches to achieve unique network decompositions that are based on function and that yield predictions on the system's dynamic behavior. Our discussion of existing methods will be guided by this principle. After introducing the formal basis for the types of mathematical models considered, we will present the module identification methods summarized in Table 1.1, and conclude with selected applications in system identification and analysis.

2 Models of Biochemical Networks

A biochemical network consists of a set of compounds or species whose connections and interactions can be captured by graph theory. Each species is then represented by a vertex or node in a graph, and a (directed or undirected) edge is drawn between species to denote an interaction.

One such representation is a *protein–protein interaction (PPI)* network (see [32] for a brief review of PPI network construction). Here, a representation by an undirected graph $G = (V, E)$ with vertex set V and edge set E is fairly straightforward. Each vertex $v \in V$ represents a protein, and an undirected edge $e = \{v_1, v_2\} \in E$ is drawn between a pair of proteins if experimental data suggests that these proteins interact. A small example is given in Fig. 1.1a. Several PPI datasets for the same organism can be combined into ever-larger sets of interactions, and PPI graphs often contain thousands of protein vertices. PPI graphs provide

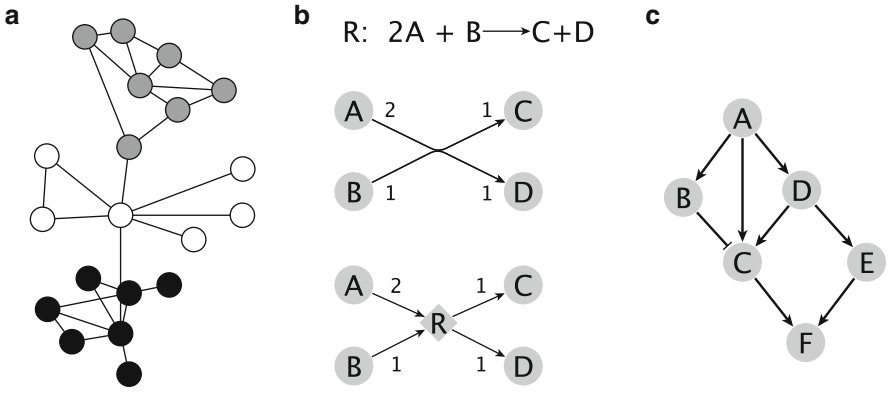


Fig. 1.1 (**a**) Graphical representation of a PPI network. Vertices in gray and black denote two potential modules with better intra- then inter-connectivity. (**b**) Representation of reaction $2A + B \rightarrow C + D$ as a hypergraph (*top*) and equivalent bipartite graph (*bottom*). Circle: species vertex, diamond: reaction vertex. (**c**) Signed interaction graph. Pointed arrow edge: positive, t-shaped edge: negative influence of one vertex on another vertex

a picture of all protein interactions: edges represent possible interactions, but dependencies between interactions or dependencies on experimental conditions are not represented.

Biochemical reaction networks admit a representation by a directed hypergraph or an equivalent directed bipartite graph. In the former, each vertex represents a species, and a directed edge is drawn from one set of species to another set of species if there is a biochemical reaction transforming the first into the latter set. In the bipartite representation, each reaction is additionally represented by a vertex and an edge is drawn from a species to a reaction vertex if that species is a substrate of the reaction. Conversely, an edge is drawn from a reaction to a species if the species is a product. Typically, edges in both representations are labeled with the stoichiometries of the species in the corresponding reaction, providing information in addition to the topology. An example is given in Fig. 1.1b, where the reaction $2A + B \rightarrow C + D$ is given in both representations.

Given a biochemical reaction network, the dynamics can be captured by a system of nonlinear ordinary differential equations (ODEs). For this, let $S_1, \ldots, S_n$ denote the chemical compounds or species and let r be the number of reactions in the reaction network. A reaction R_j is given by:

$$R_j : \sum_{i=1}^{n} a_{i,j} S_i \rightarrow \sum_{i=1}^{n} b_{i,j} S_i,$$

where $a_{i,j}$ is the *molecularity* of species S_i as a substrate in this reaction, $b_{i,j}$ is the molecularity of S_i as a product, and $N_{i,j} = b_{i,j} - a_{i,j}$ is the *stoichiometry* of the species in the reaction. Together, these stoichiometries form the $n \times r$ *stoichiometric matrix* N, with one row per species and one column per reaction.

Let $x_i \equiv x_i(t)$ be the concentration of species S_i at time t, called the states of the system and combined into the vector $x = (x_1, \ldots, x_n)^T$. Denote further by $v_j(x, p)$ the *rate of reaction* R_j as a function of the species concentrations and a vector of parameters p. These rates form the vector $v(x, p) = (v_1(x, p), \ldots, v_r(x, p))^T$ and they are often called the fluxes of the system for a particular (x, p). The dynamics of the network is then given by the set of n nonlinear ODEs as follows:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = N \cdot v(x, p). \tag{1.1}$$

While the stoichiometry (given by N) of a system of biochemical reactions is usually well characterized, crucial details of the rate law governing the change of a reaction rate as a function of the species concentrations (given by $v_j(\cdot)$) are often unknown. In particular, parameter values such as kinetic rate constants are notoriously difficult to get. Also, the algebraic form of the rate laws is often difficult to establish, especially in metabolic networks where several substrates and enzymes might simultaneously contribute to a reaction rate.

To take into consideration the (potential) dynamics of a network, the *species–species interaction graph* or *influence graph* can be constructed. This graph is

directed and signed; it has one vertex per species and a directed positive (negative) edge is drawn from S_i to S_j, if S_i has a positive (negative) influence on the concentration of S_j. The edge signs can either be established experimentally, as for network motifs (Sect. 5.1) or by the sign structure of the Jacobian matrix $N \cdot (\partial v / \partial x)$, as for monotone systems (Sect. 5.2). Figure 1.1c exemplifies an influence graph of six species. Thus, for the same biochemical network, different representations – with different levels of granularity – are possible, and it is to be expected that the choice of representation will influence the eventual modularization results.

3 Graphs and Community Detection

Once a graph for a network is constructed, its topology and in particular its connectivities can be analyzed using graph-based algorithms. This type of analysis requires only the graph itself. It uses neither the stoichiometric information provided by a reaction network, nor the dynamic or steady-state information provided by the ODE system of such a network.

One popular type of analysis aims at describing the overall organization of a graph in terms of topological statistics, such as the distribution of path lengths between vertices, and in particular the distribution of vertex degrees. The degree of a vertex v is the number of vertices $w \in V$ such that $\{v, w\} \in E$, that is, the number of its incident edges; vertices with high degree are sometimes called hubs. Graphs can be categorized by their vertex degree distribution. Of particular interest are scale-free graphs, in which the degrees follow a power–law distribution such that the probability of having degree d is proportional to d^{γ} [21]. Other approaches try to derive a more precise picture of the overall graph by extending the concept of the degree to so-called graphlets, which are essentially subgraphs of a given size. For each vertex, the number of graphlets it participates in is counted, and a graphlet distribution is computed. Comparisons between graphs can then be based on these distributions [29].

Densely connected subgraphs – sets of vertices that have more connection among each other than to the rest of the graph – are of particular interest in the topology of a graph. An example is given in Fig. 1.1a, where the shaded vertices belong to two potential modules. While various ways of defining "more densely connected" precisely were proposed, many algorithms to find the modules rely on methods from community detection [11]. Some of these algorithms try to find a minimal set of edges such that by removing these edges, the graph decomposes into a number of disjoint components, which are then identified as the modules. Other approaches use explicit or implicit measures for similarity of vertices and try to optimally cluster vertices into modules, such that members of a module are more similar to each other than to nonmodule vertices.

In the case of PPI networks, modules are usually identified with a particular biological function, with the reasoning that interacting proteins typically either stem from a protein complex or are otherwise simultaneously involved in the

same biological process. One way of finding a module's biological function is to analyze the gene ontology (GO) [6] terms associated with its proteins. If an enrichment of GO terms is found for a particular module, it is associated with the corresponding biological function. This annotation can then be transferred to proteins in the module that are not yet annotated with a function. While this method works reliably for detecting protein complexes in terms of modules, assigning a function to a larger module is often not straight-forward and various extensions for improving the biological relevance have been suggested [9, 27]. However, co-expression patterns of proteins and co-memberships are rarely correctly reflected in modules. In particular, not correcting for known protein complexes can introduce a severe bias in connectivities and complicates the analysis [44].

Metabolic network models contain additional information about stoichiometries of reactions and potentially kinetics of reaction rates, and several methods for topological modularization have also been proposed for these networks. While more densely connected subgraphs might not admit a straightforward biological interpretation, some methods rely on similar ideas to identify subgraphs that have only few connections to the remaining network and can potentially be identified with pathways of the network [13].

Several methods have also been proposed to decompose a metabolic network into hierarchies of modules. One attempt uses extensive simulations to cluster trajectories of (groups of) species by similarity and then iteratively assigns compounds and modules into other modules, until all modules are hierarchically nested [10].

While metabolic networks are usually described as hyper- or bipartite graphs, many decomposition methods work on a derived species–species influence graph instead. This causes particular problems, as different species-reaction schemes can lead to the same species influence description, which might lead to artifacts [24]. Note that many of the above approaches may suffer from the same problem because the underlying network model (usually, a simple graph) might not be appropriate to capture the network structure and its implicit constraints on network function.

4 Stoichiometric Network Analysis and Metabolic Pathways

Stoichiometric network analysis (SNA) operates on the stoichiometric matrix N and additionally incorporates other physico-chemical constraints on system behavior, such as reversibilities and capacities of reactions. It is thus based on a more detailed system model using hypergraphs that accounts for the coupling of educts and products in each reaction [19]. Moreover, the definition of pathways employs a functional criterion: each pathway has to define a feasible steady-state flux distribution in the network. Hence, any modular decomposition in SNA is limited to steady-state regimes, but it allows for function predictions in those regimes [39].

Extreme pathways (EPs) and elementary flux modes (EMs) are formally defined pathways that provide a functional decomposition of a network based on its stoichiometry. Briefly, each such pathway fulfills three criteria: (1) it allows for a

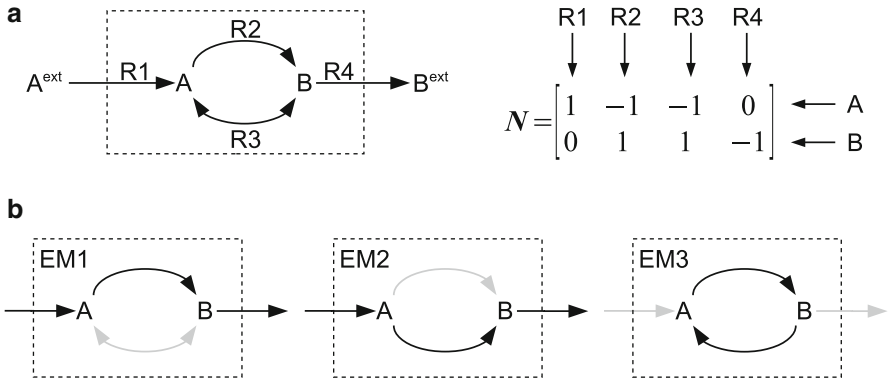


Fig. 1.2 (**a**) Example network with two internal metabolites (A, B) and two external metabolites (A^{ext}, B^{ext}), respectively, that is captured by the stoichiometric matrix N. Arrows denote reaction directions. (**b**) Decomposition of the network into three EMs EM1-3 where black/gray arrows indicate active/inactive reactions, respectively

steady-state flux distribution, (2) the fluxes are feasible, that is, no reaction directionality is violated, and (3) the pathways are minimal in the sense that a pathway cannot be represented by a combination of other pathways [39]. The small example network shown in Fig. 1.2 illustrates the principles of pathway analysis by EMs. Note that the EMs define functional regimes, not mutually exclusive subunits of the network; linear combinations of the EMs describe the entire space of valid steady-state flux distributions. However, the number of pathways – and their overlaps – explodes combinatorially with network size [20]. Hence, while incorporating a clear functional definition, it is a subsequent task to identify nonoverlapping modules from metabolic pathways. It is still a largely open question how to achieve such a decomposition and simple approaches such as (bi-)clustering of large pathways sets for realistic metabolic networks face substantial computational challenges.

Promising stoichiometry-based modularizations, however, have been proposed using the kernel matrix (an orthonormal basis of the right null-space of N). The concept presented by Poolman et al. [28] relies on the computation of reaction correlation coefficients, from which a distance matrix for reactions can be constructed. It is an extension of enzyme sub-sets, that is, fluxes in the network that are always fixed to a constant ratio, implying that they are 100% correlated [26]. Hierarchical clustering based on thus defined reaction distances is computationally feasible, and it yielded hierarchically nested modules in genome-scale metabolic networks. In contrast to those modules obtained from topology alone (Sect. 3), they incorporate functional criteria – but they are also less accessible to (biological) interpretation [28].

An alternative approach for module identification in metabolic networks starts from (predicted or experimentally determined) flux distributions in a network. Top–down partitioning of the metabolite interaction graph, where reaction edges are

weighted by their fluxes, yields the desired modules. The link to formal pathways is made by projecting potential partitions onto EMs to identify the most likely functional modules [47]. Thus, the approach combines modules in a functional regime with generally valid pathways. However, these predictions are restricted to operation of the network in steady-state. Also, theoretically derived criteria for the partition of functions in metabolic networks beyond pathways are still lacking; both concepts above rely on a heuristic step in identifying functional modules.

5 Modules in Dynamic Networks: Definition by Behavior

In principle, topological analysis of community structures and hierarchies as well as stoichiometric analysis of subnetworks leading to steady state can also be performed on fully dynamic networks given by (1.1). For an analysis of the dynamics of a larger network, however, it seems mandatory that modules have a prescribed or easily identified input–output description and therefore a clearly identified dynamics [1]. Two main approaches in this direction exist, namely, those focusing on local features of subnetworks, and others that attempt a global network decomposition based on well-defined behavior.

5.1 Local Structure: Network Motifs

One of the earliest attempts to capture dynamics of subnetworks identified several network motifs in models of transcriptional regulation [2]. A network motif in this sense is a small subnetwork that performs a particular dynamic function. Larger networks can then be analyzed by finding all contained network motifs and studying their relation.

An example of a network motif is the inconsistent feedforward mechanism, which was recognized as a subsystem able to generate pulses or to accelerate responses. Other examples include consistent feedforward schemes and densely overlapping regulons [2]. In [33], network motifs were identified in an *E. coli* network and a first attempt was made to use motifs in wiring diagrams to elucidate the structure of the overall network. Some motifs embedded in a larger context are given in Fig. 1.1c. Vertices A, B, C form an inconsistent feedforward motif. However, vertices A, C, D also give a consistent feedforward motif, sharing the edge $A \to C$ with the A, B, C motif. Both motifs are further embedded in larger feedforward motifs, ultimately ending in vertex F.

Note that, in principle, most network motifs do not have the same qualitative dynamics for all possible assignments of parameter values and for all possible realizations in terms of subgraphs. The bi-fan motif, for example, was found to exhibit various dynamics depending on its exact configuration [16]. However, in most cases, motifs found in real biological contexts show a unique dynamic function that can also be experimentally observed [2].

Originally developed for networks of transcriptional regulation, network motifs are also of particular interest for signaling networks, where they might be associated with clearly defined steps in the processing of an external signal. In contrast to regulatory motifs, these are often based on feedback rather than feedforward mechanisms. Important examples of subnetworks are reviewed in [40] and [41]. These include several response mechanisms that determine, for instance, how rapidly a system responds to a stimulus, as well as several feedback mechanism. It is long known in control engineering that the existence of certain subsystems is a necessary condition for particular dynamics: negative feedback, for example, is needed to generate oscillatory signals and positive feedback can be associated with an irreversible switch, that is, a bistable system. However, the potential reaction mechanisms implementing such subsystems are numerous and they are not easy to identify in a sufficiently complex network.

One unresolved problem associated with network motifs is prompted by their embedding in a larger context. On the one hand, motifs with specific local function do not necessarily have this same function when interconnected with a larger network that systematically feeds the motif's output back to its input. On the other hand, motifs in a larger context often overlap, such that sets of vertices and edges are part of more than one motif; an example of this was already given in Fig. 1.1c. Similar to the metabolic pathway-based module definitions, therefore, clear-cut criteria for module demarcation are still missing.

5.2 *Global Decomposition: Monotone Systems*

While network motifs present a bottom–up approach to finding dynamic modules in a given network, decomposing a network into modules by a top–down approach can also yield insight into the network's dynamics.

One recently proposed way to arrive at a decomposition of (1.1) into modules exploits the theory of monotone systems [15], extended to systems with inputs and outputs [34]. Such input–output systems extend (1.1) by a set of input functions u that allow an influence of the system by external signals, and an output function $y : \mathbb{R}^n \to \mathbb{R}^m$ that maps the state of the system to m numbers, which can in turn be fed into the input of another (sub)system. A system is monotone with respect to three partial orders (all denoted by $\prec$), defined on the inputs u, the outputs y, and the solutions $\phi(t; x_0, u)$ of the ODE system (1.1) with initial conditions x_0 and input $u(t)$, if the following monotonicity condition

$$u_1(t) \prec u_2(t),\ x_1 \prec x_2 \implies \phi(t; u_1, x_1) \prec \phi(t; u_2, x_2),\ y_1 \prec y_2$$

holds for all times $t \geq 0$, all inputs $u_{\{1,2\}}(\cdot)$, and all initial conditions $x_{\{1,2\}}$.

Monotonicity of a system allows statements on existence and stability of steady states. It guarantees that the system responds "well-behaved" in the presence of perturbations. Moreover, important special cases of monotonicity (for a certain class of partial orders) can be established from the topology of the bipartite network

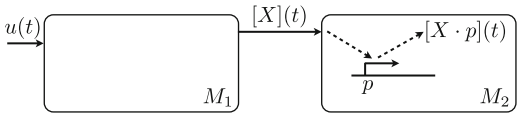


Fig. 1.3 Example of retroactivity: Protein X is the output of module M_1 and also forms a complex with a promoter in module M_2. With $[p] + [X \cdot p] = \text{const.}$, this changes the input-output characteristic of M_1. Adapted from [42]

graph alone, under very mild assumptions on the class of admitted reaction rate laws. Feedback configurations of monotone systems have also been successfully investigated; see [34] for details on monotone systems.

While most models of biological networks are not monotone, some are near-monotone in the sense that only a few edges in the graph need to be deleted to render them monotone. Algorithms for identifying such sets of edges have been proposed [8, 37] and they open the route to a decomposition of large systems into modules of monotone subsystems. The fact that these modules are also independent of particular assumptions on rate laws or kinetic parameters makes this approach particularly appealing.

5.3 Retroactivity

Another approach for decomposition of a larger network is to minimize the "retroactivity" between modules. Consider any module M_2 that is downstream of another module M_1: there are connections from M_1 to M_2, but not from M_2 back to M_1. Even in such a cascade configuration, with no apparent feedbacks between the modules, the downstream module might influence the dynamic properties of the upstream module. An example is given in Fig. 1.3, where an input $u(t)$ to module M_1 generates a protein species X, whose concentration is the input to module M_2. In M_2, the protein acts as a transcription factor that forms a complex $X \cdot p$ with a promoter p; importantly, the overall promoter concentration $[p](t) + [X \cdot p](t)$ stays constant. The input–output behavior of M_1 is then changed by the presence of M_2, as the stoichiometric interaction of X with p also changes the trajectory $[X](t)$. To quantitatively describe this influence, the concept of *retroactivity* was proposed together with an investigation of potential insulation mechanisms [42].

In terms of modular approaches, this concept can be exploited by using methods from community detection that simultaneously minimize the retroactivity between modules [31]. In principle, such modules could then be analyzed separately and only the respective upstream modules need to be taken into consideration for describing the dynamics of interconnected modules. By itself, retroactivity, however, does not provide a definition of modules, but minimization of retroactivity could be part of the objective in finding modular structures in biological networks.

5.4 Interfaces

Similar considerations apply to the topic of interfaces. Consider any two modules M_1 and M_2. Their *interface* is given by the quantities that need to be exchanged between them. For example, a particular reaction might be assigned to module M_1 and one of its substrates to module M_2. Then, the substrate's concentration needs to be exchanged between the modules as must the reaction rate (which is a quantity of M_1, but needed by M_2 to update the substrate concentration).

The problem of interfaces has some analogies to control and electrical engineering. In electronics, components used to design larger circuits have well-defined interfaces that prescribe the information carrier (electrons) as well as input/output impedances and the physical layout of contact wire. Biological networks on the other hand do not have a common information carrier, so a proper definition of interfaces is complicated by the fact that a concentration of some species X cannot be connected to a module that needs a reaction flux as input. While this is not a problem in analysis, it poses considerable problems designing novel biological circuits in synthetic biology [23].

The problem of impedances is in fact the starting point for studying the problem of retroactivity in biological systems. While it can be used to define modules in the first place, it might also be explicitly taken into account in the definition of the interfaces. For example, instead of connecting modules directly, the interface can be defined by pools of the interfacing species, which requires modules to be connected exclusively via these pools [23]. For signaling networks, this would also allow tuning the description detail by, for instance, having a pool of ATP and assigning either a dynamics to it or considering it constant. Modules of the network would then automatically benefit from either choice.

In control engineering, block-diagrams are an essential tool for describing larger systems as interconnected smaller systems with prescribed function. Here, single-input single-output (SISO) systems are well understood, even for nonlinear dynamics. However, analysis gets more involved the more inputs and outputs the systems have. Similar arguments hold for analyzing interconnected biological network modules. Thus, one objective of any top–down decomposition method should also be to minimize the module interfaces.

6 Applications: Modular Systems Identification and Analysis

Work on modular analysis of biological networks has often concentrated on identifying and characterizing the (types of) modules found in natural systems. However, at least in the two areas described below, modular analysis approaches have shown direct benefits for analyzing the system functions.

6.1 Control Systems

Decomposition of a given system into blocks with particular function is a standard practice for designing systems in control engineering. However, translating these techniques to biological systems is far from easy. For one, engineering systems are by default designed to provide a certain function irrespective of the system they are integrated into, for example, by providing insulation to avoid retroactivity. In addition, many well-established results from engineering such as the internal model principle [12] are in theory also applicable to biological models, but they require subtle but nontrivial extensions. Nevertheless, examples of successful application of control engineering methods exist, the most prominent one probably being the analysis of the chemotaxis system in *E. coli* [22, 46]. Starting from fundamental concepts in engineering, the perfect adaptation observed in the chemotaxis system could be shown to require an integral feedback loop, which was then identified in the reaction network. Here, an implicit modularization of the system helped in understanding the design principles of the whole system.

Other examples include the successful extension of the internal model principle to biological signaling pathways, in which plant (the controlled system) and controller cannot reasonably be separated, and a main feature is the detection of a signal [35]. A first review of attempts to bring control theoretic concepts into biological systems analysis is given in [36], including the internal model principle, monotone systems, and retroactivity.

6.2 Modular Response Analysis

Another example of modular analysis using concepts from control engineering is the development of metabolic control analysis (MCA) and, more recently, its application to the modular analysis of signaling pathways. MCA was originally developed to study enzymatic reactions and to analyze the control of fluxes in metabolic networks [45]. In essence, it is a sensitivity analysis of the reaction rates with respect to parameters and species concentrations. MCA derives much of its power from the fact that the underlying network puts additional constraints on the sensitivities. Those constraints do not occur for general physical or engineering systems and they lead to, for example, the celebrated summation theorems [30].

While originally designed to analyze control of fluxes in steady-state, MCA has been adapted to be applicable to more transient system behaviors as they occur in signaling networks. Here, efforts to apply MCA techniques by so-called modular response analysis [3, 4, 17] are of particular interest. In one application, responses of modules with respect to perturbations are exploited to identify connections in

an unknown network. Furthermore, systems of different modules of organization – combined models of metabolism, regulation, and signaling – can be analyzed by MCA methods because connections from one module to the next are often not stoichiometric, that is, by mass flow, but purely at the level of flux regulation. For example, while the concentration of an enzyme might be changed by a regulatory network, it only acts as a modifier for metabolic reactions, and the metabolic network has no direct influence on the enzyme concentration. This would be an extreme example of insulation where the two systems do not have retroactivity.

7 Conclusions and Perspectives

With constantly increasing size and scope of models for biological networks, computational systems biology faces new challenges to provide adequate methods for analysis of such networks. For now, methods based on divide-and-conquer strategies seem imperative. In this paper, we focused on networks that admit a formal representation and presented several decomposition approaches that lead to modules with rationally defined properties.

Methods based on topology alone work reasonably well for PPI networks, but the input–output dynamics needs to be taken into consideration when modularizing metabolic or signaling networks with inherent dynamics. Then, methods that simultaneously provide modules with specific "well-behaved" input–output dynamics and minimal interfaces between these modules are of particular interest. In engineering, modules or subsystems with specific dynamic behavior are manifold and commonly used. For biological networks, however, the task is mostly analysis rather than synthesis, which often precludes an unique decomposition into modules. In particular, small subnetworks with specific properties, such as network motifs or extreme modes, often overlap, which complicates the analysis.

In order to be able to analyze larger biochemical reaction networks, such as metabolic or signaling networks, new and improved methods are needed. While there are several examples of a successful analysis using standard tools from control engineering, differences between the analysis of engineered and biological systems are often subtle, yet important. For example, there is no clear cut between the plant and the controller in a biological system, and parts of a network might be used by several response mechanisms. Further, both the structure and the specific parameter values of biochemical reaction network models are often incomplete and uncertain. Therefore, any modularization method needs to be robust with respect to small perturbations in the network structure or parameters.

Acknowledgment Financial support by the EU FP7 project UNICELLSYS is gratefully acknowledged.

References

1. Alexander RP, Kim PM, Emonet T, Gerstein MB (2009) Understanding modularity in molecular networks requires dynamics. Sci Signal 2(81):44
2. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8(6): 450–461
3. Bruggeman FJ, Snoep JL, Westerhoff HV (2008) Control, responses and modularity of cellular regulatory networks: A control analysis perspective. IET Syst Biol 2(6):397–410
4. Bruggeman FJ, Westerhoff HV, Hoek JB, Kholodenko BN (2002) Modular response analysis of cellular regulatory networks. J Theor Biol 218(4):507–520.
5. Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK (2009) Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol 5:239
6. Consortium, TGO (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29
7. Csete M, Doyle J (2002) Reverse engineering of biological complexity. Science 295: 1664–1669
8. DasGupta B, Enciso GA, Sontag E, Zhang Y (2007) Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. Biosystems 90(1):161–178
9. Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol 1:24
10. Ederer M, Sauter T, Bullinger E, Gilles ED, Allgower F (2003) An Approach for Dividing Models of Biological Reaction Networks into Functional Units. Simulation 79(12):703–716
11. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):175–174
12. Francis B, Wonham W (1976) The internal model principle of control theory. Automatica 12:457–465
13. Gagneur J, Jackson DB, Casari G (2003) Hierarchical analysis of dependency in metabolic networks. Bioinformatics 19(8):1027–1034
14. Hartwell L, Hopfield J, Leibler S, Murray A (1999) From molecular to modular cell biology. Nature 402 (Suppl.):C47–C52
15. Hirsch M, Smith H (2006) Monotone dynamical systems. Handbook Differen Equat Ord Differen Equat 2:239–357
16. Ingram P, Stumpf M, Stark J (2006) Network motifs: structure does not determine function. BMC Genom 7(1):108
17. Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag ED, Westerhoff HV, Hoek JB (2002) Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. Proc Natl Acad Sci USA 99(20):12841–12846
18. Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
19. Klamt S, Haus UU, Theis F (2009) Hypergraphs and cellular networks. PLoS Comput Biol 5(5):e1000385
20. Klamt S, Stelling J (2002) Combinatorial complexity of pathway analysis in metabolic networks. Mol Biol Rep 29:233–236
21. Koonin EV, Wolf YI, Karev GP, Almaas E, Barabasi A.L (2006) Power laws in biological networks. In: Power Laws, Scale-Free Networks and Genome Biology, Molecular Biology Intelligence Unit, Springer USA, 1–11
22. Lauffenburger DA (2000) Cell signaling pathways as control modules: complexity for simplicity? Proc Natl Acad Sci USA 97(10):5031–5033
23. Marchisio MA, Stelling J (2008) Computational design of synthetic gene circuits with composable parts. Bioinformatics 24(17):1903–1910
24. Montanez R, Medina MA, Sole RV, Rodrigues-Caso C (2010) When metabolism meets topology: Reconciling metaboltie and reaction networks. BioEssays 32:246–256
25. Nurse P (2003) Understanding cells. Nature 424:883
26. Pfeiffer T, Sanchez-Valdenebro I, Nuno J, Montero F, Schuster S (1999) METATOOL: For studying metabolic networks. Bioinformatics 15:251–257

27. Pinkert S, Schultz J, Reichardt J (2010) Protein interaction networks – More than mere modules. PLoS Comput Biol 6(1):e1000659
28. Poolman MG, Sebu C, Pidcock MK, Fell DA (2007) Modular decomposition of metabolic systems via null-space analysis. J Theor Biol 249(4):691–705
29. Przulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23(2):e177–e183
30. Reder, C (1988) Metabolic control theory: A structural approach. J Theor Biol 135:175–201
31. Saez-Rodriguez J, Gayer S, Ginkel M, Gilles E.D (2008) Automatic decomposition of kinetic models of signaling networks minimizing the retroactivity among modules. Bioinformatics 24(16):i213–i219
32. Seebacher J, Gavin AC (2011) SnapShot: Protein–protein interaction networks. Cell 144(6):1000.e1
33. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1): 64–68
34. Sontag E (2007) Monotone and near-monotone biochemical networks. Lecture Notes in Control and Information Sciences, vol. 357, pp. 79–122
35. Sontag ED (2003) Adaptation and regulation with signal detection implies internal model. Syst Contr Lett 50(2):119–126
36. Sontag ED (2004) Some new directions in control theory inspired by systems biology. Syst Biol 1(1):9–18
37. Soranzo N, Ramezani F, Iacono G, Altafini C (2010) Graph-theoretical decompositions of large-scale biological networks. Automatica, conditionally accepted.
38. Stelling J, Kremling A, Ginkel M, Bettenbrock K, Gilles E (2001) Towards a Virtual Biological Laboratory. In: Kitano H (ed) Foundations of Systems Biology, MIT Press, Cambridge, MA, pp. 189–212
39. Terzer M, Maynard ND, Covert, MW, Stelling J (2009) Genome-scale metabolic networks. Wiley Interdiscip Rev Syst Biol Med 1(3):285–297
40. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles, and blinkers: Dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15(2):221–231
41. Tyson JJ, Novák B (2010) Functional motifs in biochemical reaction networks. Annu Rev Phys Chem 61:219–240
42. Vecchio DD, Ninfa AJ, Sontag ED (2008) Modular cell biology: retroactivity and insulation. Mol Syst Biol 4:161
43. Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8(12):921–931
44. Wang Z, Zhang J (2007) In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 3(6):e107
45. Westerhoff HV, Kolodkin A, Conradie R., Wilkinson SJ, Bruggeman FJ, Krab K, van Schuppen JH, Hardin H, Bakker BM, Moné MJ, Rybakova KN, Eijken M, van Leeuwen HJP, Snoep JL (2009) Systems biology towards life in silico: Mathematics of the control of living cells. J Math Biol 58(1–2):7–34
46. Yi TM, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc Natl Acad Sci USA 97(9):4649–4653
47. Yoon J, Si Y, Nolan R, Lee K (2007) Modular decomposition of metabolic reaction networks based on flux analysis and pathway projection. Bioinformatics 23(18):2433–2440

Chapter 2
Modeling Signaling Networks Using High-throughput Phospho-proteomics

Camille Terfve and Julio Saez-Rodriguez

Abstract Cellular communication and information processing is performed by complex, dynamic, and context specific signaling networks. Mathematical modeling is a very useful tool to make sense of this complexity. Building a model relies on two main ingredients: data and an adequate model formalism. In the case of signaling networks, we build mainly upon data at the proteome level, in particular about the phosphorylation of proteins. In this chapter we review recent developments in both data acquisition and computational analysis. We describe two approaches, antibody based technologies and mass spectrometry (MS), along with their main features and limitations. We then go on to describe some model formalisms that have been applied to such high-throughput phospho-proteomics data sets. We consider a variety of formalisms from clustering and data mining approaches to differential equation-based mechanistic models, rule-based, and logic based models, and on through Bayesian network inference and linear regressions.

1 Introduction

Whatever their nature, identity, and environment, cells are continuously exposed to signals, whether reflecting their internal state, or emerging from growth factors, neighboring cells or the extracellular matrix. All these signals need to be received, interpreted and possibly transmitted or propagated, in an integrated manner so

C. Terfve
European Bioinformatics Institute (EMBL–EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
e-mail: terfve@ebi.ac.uk

J. Saez-Rodriguez (✉)
EMBL–EBI and European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstrasse 1, D-69117 Heidelberg, Germany
e-mail: saezrodriguez@ebi.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_2,

as to produce the appropriate response. This information processing is performed through the use of highly dynamic and context specific networks assembled from a multitude of signaling molecules [32]. Given their fundamental role in cellular function and intercellular coordination, deregulation of signaling networks is often involved in the development of diseases [3, 27]. Furthermore, although development of resistance to drugs can happen through accumulation of mutations, it seems that another underlying mechanism can be the rewiring and adaptation of the signaling network. Combinatorial genetic perturbations in yeast suggest that signaling networks are extremely adaptive to such perturbations [32]. Studying signaling networks as a whole, in physiological and disease contexts, is therefore essential to understand how cells function and respond to their environment and how this process is deregulated in diseases, to potentially provide new venues for therapies.

Understanding how the elements that make up signaling systems are organized and function together to allow the cell to respond to a perturbation is a challenge. This is only the beginning that is to be investigated [28]. Therefore, it has been argued that mathematical modeling is necessary to make sense of the sheer amount of elements that enter into play [1, 13, 28, 32, 40]. A key point when modeling is to be aware of the assumptions made in building the model (level of detail, scope, etc.), and to interpret the model outputs correctly [5, 65]. Indeed, there are many ways to model biochemical and in particular signaling networks, and the choice of a particular formalism depends on the system under investigation, the data at hand, and the question to be answered.

Ideally, one builds a model based on good quality data of the system under study; in the case of signal transduction data is at the proteome level, since proteins are the ultimate agents of cellular activity. Events occurring at the transcriptional levels can however also be included in models of signaling systems (see [34] for a review on cellular regulatory networks encompassing regulation at different levels). Proteomics is the field of biology that studies the expression, modification, conformation, and activity of proteins in a system [2, 30]. This is challenging for many reasons, mainly because screening of the entire proteome (as can be done for the genome or transcriptome) is impractical with current technology and because the proteome cannot be defined using a list of proteins. Indeed, the wide varieties of post-translational modifications (PTMs) and their combinations lead to a combinatorial explosion of the number of states that need to be assessed. To add to the complexity of this picture, protein abundance in a single cell population frequently spans more than 6 orders of magnitude [2], which is broader than the dynamic range of any routine proteomics technology currently available. In addition, limited information can be inferred from investigations at the mRNA level since transcripts levels are poorly correlated with protein abundance [23, 73]. An ideal technology to probe the proteome of human cells in a signaling framework would have to be able to measure accurately the concentrations of more than 30,000 different proteins and their splice variants, each possibly subject to a variety of post-translational modifications (e.g., an estimated 100,000 sites for phosphorylation alone) and should be able to measure all this in a time-dependent,

cell and compartment specific manner, under various conditions such as genomic background and stimulations by drugs [62]. Although no single platform is currently capable of such an achievement, some progress has been made with regards to capabilities of medium to high-throughput proteomics technologies, which we will discuss further in this chapter. In particular, we will discuss advances in antibody based and mass spectrometry (MS) based proteomics workflow that have allowed and will allow informative modeling of signaling networks.

Signal propagation involves changes at three levels: regulated PTMs, protein–protein interactions (often owing to PTMs), and changes in the expression level of proteins. These three levels are coordinated through dynamic regulation which is often spatially segregated [10]. Amongst these three levels, we focus on PTMs because they are the most immediate events, often trigger the other events and can thus often be used as a proxy for those other events. However, depending on the time scale of the process studied, it is possible that expression and degradation events would play an important role in the dynamics of the system. One should therefore be cautious and, if possible, also be able to measure the abundance of proteins.

Over 200 types of PTMs have been reported and that number is still growing [23]. In particular, phosphorylation and associated players (kinases, phosphatases, and phospho-binding domain containing proteins) play a very important role in signaling since this PTM can control the formation of multiprotein complexes, the dynamic localization of proteins, as well as their stability and enzymatic activity, and about 30% of proteins are phosphorylated at any one time [13]. This points to the importance of phosphorylation in the context of signaling, and combined with the availability of assays to measure phosphorylation of proteins on a large scale, this makes phospho-proteomics a common focus when looking at signaling.

The phosphorylation state of a protein reflects the result of the action of kinase/phosphatase reaction pairs. In some cases, e.g., phosphorylation of the activation loop of some protein kinases, this is correlated with the activity status of the protein. However, this is not true in the general case, and care should be taken when interpreting phosphorylation status [13]. Relatively recent data suggests that most in vivo phosphorylation sites have not even been detected yet [54]. In addition, although we can now sequence a full genome, signaling data are necessarily incomplete [20] because characterizing a "full" signaling network would necessitate to either characterize the full state of the cell (including contextual information and state of modification/interactivity of every agent) or to be able to delineate where the signaling network starts and ends.

PTMs not only have individual roles but can also function in combinations to precisely regulate molecular interactions, protein activity and stability, in a context specific manner [29]. Therefore, interpreting phosphorylation in a signaling context is likely to prove very challenging, and prioritizing functionally important phosphorylation sites for experimental investigation is going to be crucial [67], as will be the identification of kinases involved in particular modifications [13].

This chapter is outlined as follows: in the first part, we will discuss available experimental platforms and explain their particular features and limitations. In the second section, we will briefly discuss methods for building signaling networks that

have been applied to high-throughput phospho-proteomics data sets. This section is not exhaustive neither in terms of application examples nor in terms of specific methodologies, but it rather aims at describing modeling frameworks that have been applied to phospho-proteomics data by showing a couple of examples, and discuss the advantages and disadvantages of each method. Most of the methods presented have been applied in many other contexts (often prior to signaling networks), in particular to the much more mature field of modeling of gene regulatory networks [4, 31, 44].

2 Phospho-proteomics Data Collection

Data collection methods for medium-/high-throughput proteomics can be roughly divided into two categories: those that do not make any assumption about the sample composition (e.g., shotgun MS), and those that measure a predetermined set of proteins (e.g., affinity based approaches) [2, 62]. Affinity based technologies most commonly make use of antibodies, and those methods will be the subject of the first part of this section. The second part of this section will examine the principles of common shotgun MS and will take a closer look at targeted MS as a potential alternative to antibody based approaches to generate large data sets for the development of systems biology models. The choice of a method to use ultimately depends on the material and expertise available, and the number of experiments that can be performed often results from a balance between the time and cost per experiment.

2.1 Antibody-based Methods

All antibody based methods build upon the same principle: the interaction of a target protein with an antibody, an interaction that should happen with both high-affinity and selectivity. Therefore, all of these methods suffer from the same limitation: the data is only as good as the antibodies are, and investigators are therefore limited by the availability of high-quality antibodies [2]. However, new multiplexing technologies offer the ability to analyze hundreds to thousands of samples a day, thereby allowing assays on multiple time points, and across multiple conditions of interest (which is not yet possible with MS due to the labor intensive process of analyzing more than a few conditions), although the total number of signals measured rarely exceeds a few dozens [3, 75] (see Fig. 2.1 for an overview of multiplexing capacities of the methods examined in this chapter). Such data sets usually need to be normalized and quality controlled (e.g., assessing reproducibility, detecting outliers, etc.). There are a number of computational tools to do so and to connect processed data to modeling tools. This is however outside the scope of this chapter.

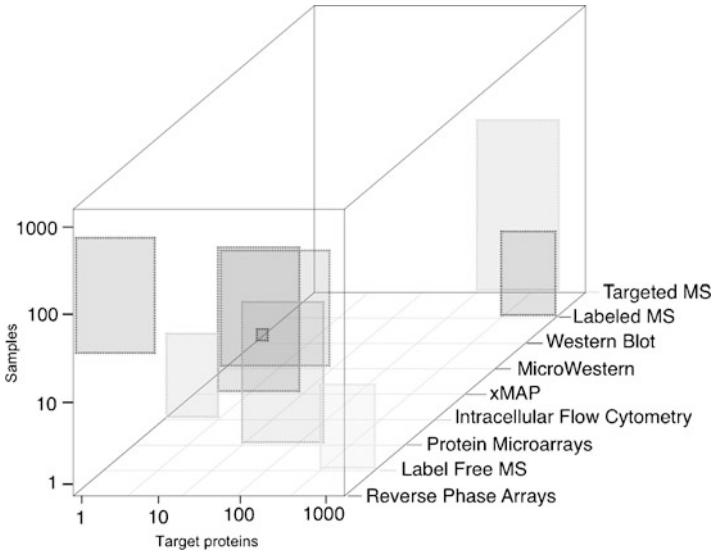


Fig. 2.1 Overview of multiplexing capacities of data collection methods. This figure displays approximations for the ranges of numbers of samples and proteins that can be interrogated using various phospho-proteomics platforms. These numbers are not intended as absolute statements on the capabilities of these methods and in particular they may change slightly depending on the actual protocol used, experimental setting and proteins targeted. This figure is inspired from [2]

The combination of fluorescently labeled antibody recognition and single cell measurement capacity of fluorescence activated cell sorting (FACS) [59] seems to be a promising technology due to its single cell nature. Another promising technology is the microwestern array developed by Ciaccio et al. [11], which in addition to antibody based recognition provides an extra separation step by electrophoresis. Other technologies, which we will not discuss in this chapter, such as high-throughput microscopy [24] associated with immunofluorescence, and mass cytometry [68], are being developed and could potentially be applied to the generation of high-throughput phospho-proteomics data sets adapted for modeling. To date, most commonly used antibody-based technologies are protein arrays, reverse-phase protein arrays, and the bead based xMAP technology from Luminex [62].

2.1.1 Intracellular Multicolor Flow Cytometry

Intracellular multicolor flow cytometry allows the simultaneous measurement of multiple phosphorylated proteins and phospholipids in large populations of cells, on a single cell basis [59]. The principle is simple, the cells are fixed and incubated with fluorescently labeled antibodies, and then are subjected to FACS which quantitatively measures the targets' expression or modification level. The main limitation of this technology is the availability of suitable reagents, i.e., antibodies compatible

with flow cytometry. Furthermore, this technology only allows a relatively small number of proteins to be examined simultaneously (up to a dozen). The ability to barcode cell populations before protein labeling potentially allows this technology to be applied to the processing of multiple samples/conditions in parallel [38].

2.1.2 Microwestern Arrays

Microwestern arrays build upon the well established western blot technology, which enables quantitative and sensitive analysis of protein abundance and modification after electrophoretic separation, while a high-throughput capacity is achieved by applying the protocol to microarrayed cell lysates. The main advantage of adding the electrophoretic separation step to the workflow is that it allows for a reduction in sample complexity, whereas the antibody detection step results in signal amplitude proportional to the abundance of immobilized protein. The signal localized at a physical position on the membrane can be related to molecular size standards, so the antibody cross reactivity problem associated with most other technologies can be controlled to some extent [11]. This method showed, in the proof of principle study, a linear relationship between antigen concentration and signal intensity over from 2 to 3 orders of magnitude [11]. The main advantages of this method over classical protein arrays are an increased specificity owing to the electrophoretic separation step, low sample requirements (compared to technologies such as xMAP) and the wide availability of reagents since antibodies developed for the classical western blot should be applicable to this method.

2.1.3 Array and Bead-based Methods

All other methods described here (reverse phase arrays, protein arrays, and xMAP technology) rely on the same principles and in particular are composed of three main ingredients: (1) an identification system which is required for multiplexing (i.e., a physical support with unique identity, whether a location on a 2D arrangement or unique physical properties of beads in suspension), (2) a capture system (to immobilize the protein(s) of interest, whether directly on the support as in reverse phase arrays or through interaction with antibodies as in protein arrays), and (3) a detection system (to produce a signal that is ideally linearly proportional to the amount of captured target protein, typically fluorescent-labeled detection or enzymatic-labeled detection such as a biotinylated secondary antibody bound by a streptavidin-linked peroxydase) [2].

In protein microarrays, the captured antibody is covalently bound to a slide in an ordered manner, and the slide is incubated with the sample. For detection, either the sample itself is chemically labeled with a fluorophore, or it is detected by a labeled secondary antibody (sandwich assay). This technology can measure up to hundreds of proteins but the number of samples is somewhat limited. Direct labeling allows for the simultaneous measurement of multiple analytes and only requires one

high-quality antibody per target protein, but due to uneven labeling of all proteins and chemical alterations this method can be rich in false positive and display a high-background. Sandwich assays on the other hand provides a more accurate and specific detection, but require two high-quality antibodies [75]. This is not a trivial problem for microarrays as, contrary to antibodies for western blots that detect denatured proteins, antibodies for such array technologies must be able to recognize the substrate in native state but immobilized on a slide, which can impose steric constraints on the interactions.

Reverse phase arrays are similar in principle but in this case the lysate itself is spotted on the support and therefore multiple lysates (dozens to hundreds) can be processed on a single slide. One can then either incubate the entire slide with one antibody or create physical compartments within which distinct primary antibodies can be used. A labeled secondary antibody then binds the captured antibody. This technology only requires one specific antibody for detection of each protein but it is therefore highly dependent on the selectivity of this antibody, and this added to the presence of all cellular proteins bound to a slide is bringing up issues of cross reactivity that have been reported to cause substantial noise [2,11,75]. Therefore, the accuracy of reverse phase arrays tends to be lower than that of protein microarrays, specifically when sandwich assays are used [58].

The xMAP technology is conceptually similar to protein microarrays except that rather than being localized on particular spots on a support, specific antibodies are associated with microspheres in suspension that are internally dyed to generate different spectral signatures. This technique theoretically supports the analysis of up to 100 analytes per well, since the beads can be multiplexed and incubated with a single sample. For detection, a mixture of biotinylated antibodies is added, and a fluorescently labeled molecule binds the detection antibody. Quantification is obtained by a flow cytometer based instrument capable of reading the beads' spectral signature and the fluorescence intensity simultaneously. Having beads in suspension rather than planar microarrays allow for faster reaction kinetics and high-surface to volume ratio, and consequently better washes and homogeneous chemical reactions resulting in an increase in the signal to noise ratio [2,71]. A disadvantage of this approach compared to protein arrays is that it requires considerably more cell material and the cost of detection is approximately 30 times higher per protein detected [11].

2.2 Mass Spectrometry

MS is an analytical technique that determines the mass to charge ratio of charged analytes, thereby providing a means to identify chemical compounds. Applied to proteomics, it allows systematic protein identification and quantification (provided that an appropriate protocol is used) from complex samples using a combination of liquid chromatography separation of peptides generated by digestion, followed by their analysis by tandem MS (a protocol called shotgun LC–MS/MS) [23].

2.2.1 Shotgun MS/MS

A classical shotgun MS workflow proceeds as follows: the protein samples are digested with trypsin and the lysate is fractionated by reversed-phase liquid chromatography, which is used to separate the complex mixture of peptides on the basis of their hydrophobicity. Other types of chromatography such as strong cation exchange are also commonly used, where the peptides are separated based on their charge. The peptides in fractions eluting from the chromatography columns are then vaporized and ionized, typically by subjecting the solution to an electric potential, which causes the formation of a spray and the desolvation and ionization of the peptides (a technique called electrospray ionization) [23]. In the MS stage, the mass to charge ratio of all ions is determined, then the first mass analyzer selects ions for collison induced dissociation, where neutral gas molecules are used to fragment the peptide. The resulting fragment ions are measured in the second mass analyzer of the tandem MS [10, 23]. The precursor ion intensities measured at the MS stage can be used for peptide quantification, and the MS/MS fragment ion information can be used to identify the peptide through its sequence, by comparing the experimental MS/MS fragmentation pattern to theoretical counterparts derived from a database of sequences from in silico digested proteins. Subsequent protein identification can be obtained through a database search [23]. For a review about how to obtain and interpret sequence information from tandem MS experiments, we refer the interested reader to reference [66].

2.2.2 Data Processing Challenges

The problem of assigning sequences to MS spectra is not a trivial one, and each identification should be carefully assessed for its statistical significance [8]. Most of the algorithms performing this task report one or more peptide spectrum match (PSM) scores that reflect the quality of the match between the experimental and computed theoretical peptide spectrum. Statistics associated with these scores are typically obtained by searching the data against a target/decoy database, i.e., in addition to search through real sequences, the search is also performed against a randomized, shuffled, or reversed database. This gives an approximation of the FDR (expected proportion of false assignments among a selected set of predictions) by counting the number of matches in the target (presumably mainly true positives), and decoy (presumably mainly false positives) databases that satisfy a score criteria. Some algorithms supplement this information by implementing methods to improve the discrimination between correct and incorrect PSMs, for example, by building classifiers that also make use of other features reported by the search algorithm, such as charge state, difference in score to the second best hit, etc., which are often used by experts to manually validate the PSMs [8].

After having identified the peptide present in the sample with a certain level of confidence, another problem arises before the data can be readily interpreted: the protein assignment problem, i.e., identifying the protein composition of the sample

from which the peptide sequences result [52]. Indeed, the same peptide sequence can match multiple different proteins, making both identification and quantification challenging. This problem can be partially alleviated when the sample complexity is further reduced prior to digestion and LC–MS/MS using techniques such as 2D gels, which can provide additional information such as molecular weight and the isoelectric point of the protein. The issue is particularly challenging in the case of higher eukaryote organisms since these organisms present a certain degree of sequence redundancy [52].

Distinguishing between different proteins of similar sequence is of course increasingly difficult when the sequence coverage decreases (i.e., the fraction of the protein sequence that is covered by identified peptides). Unfortunately, the sequence coverage observed in shotgun MS proteomics experiments is typically quite low. Several factors contribute to this, such as, the size of the proteins to be identified, enzymatic digestion constraints, and the detection mass range of the instrument. Furthermore, some unexpected PTMs can lower the chances of a peptide being observed, and low abundance or poorly ionizing peptides are also less likely to be selected for MS/MS sequencing [52]. For more information about this topic, we refer the reader to the following review [52].

2.2.3 Quantitative MS

Regarding the quantification of proteins using MS, two main approaches can be applied: differential isotope labeling and label free quantification. Differential isotope labeling builds on the hypothesis that when measuring two analytes of identical chemical composition but different stable isotope composition, their relative signal intensity represents their relative abundance in the sample. There are two main ways to do this: in vitro labeling or in vivo incorporation of isotope-labeled amino acids through metabolic labeling (stable isotope labeling with amino acids in cell culture, SILAC). For in vitro labeling, the two samples are prepared separately and the protein or peptide solutions are individually labeled with heavy or light version of tagging reagents. The recently introduced iTRAQ technology allows peptide labeling with isobaric tags, as the name indicates, keeps the mass of differentially labeled precursors constant, i.e., appearing as a single peak in the MS1 spectrum. Quantification occurs in the MS/MS spectrum by comparing peak areas of sample-specific reporter ions [23, 58]. Compared to isotope labeling techniques which only allow up to typically three samples to be compared simultaneously, the iTRAQ labeling protocol can compare up to eight samples in a single LC–MS/MS run. A very similar idea is implemented in the Tandem Mass Tags protocol [70]. SILAC is an in vivo labeling method where different populations of cells are grown in presence of media containing light or heavy isotope versions of lysine or arginine most commonly [23], although other amino acids have been used (e.g., leucine [55]), and labeling of living animals such as rats with N15 has also been reported [22]. Since the labeling occurs very early on in the protocol, this method avoids many

of the errors and biases than can be introduced in the sample processing. However, this method is limited to cells or organisms that can be metabolically labeled, i.e., typically cell cultures and not primary samples [23], and it is quite a complex and time consuming protocol, which limits its implementation to laboratories with significant infrastructure [13].

Label free quantification by peptide precursor ion intensities is based on the alignment of high-mass accuracy MS1 (i.e., precursor ions) spectra obtained from separate LC–MS/MS experiments. Peptides are identified and aligned based on their specific retention time and mass to charge ratio. The relative abundance changes are calculated from the aligned spectra on the basis of the signal intensities of extracted ion chromatograms. Another label free method, spectral count, relies on the assumption that the rate at which a precursor ion is selected for fragmentation is correlated to its abundance. The spectral counts from peptides mapping to the same protein are then averaged into a protein abundance index. This method depends on the quality of the MS/MS peptide identification and protein assignment, and although it works relatively well for abundant proteins, it is often problematic for small and low abundance proteins [23]. In general, label free techniques provide a less accurate quantification than stable isotope label methods [58].

2.2.4 MS for PTMs

Because the addition of a PTM to a protein causes a defined mass change, MS can measure and localize modifications with a single amino acid resolution. However, PTM analysis poses specific challenges beyond those described above: modified peptides are often present at low amounts, can lead to more complicated MS/MS spectra and increase the database search space [10]. Therefore, it is usually necessary to enrich the sample for the modification of interest in order to increase the dynamic range and sensitivity. Depending on the PTM, this can be done by derivatization of the PTMs and chemical solid phase capture, or more commonly, for phosphorylation using metal affinity chromatography, titanium dioxide chromatography, or antibodies specific for a modification [23]. Ideally, one would hope to obtain all modified peptides and only those but in practice all modified peptides will be enriched to a certain degree with respect to the starting mixture, with an enrichment factor that can range from only several folds for some modifications to over a hundred fold for phosphorylations [10].

When looking at PTMs, two different tasks are performed: the identification of the peptide bearing the PTM, and the unambiguous localization of the PTM-bearing amino acid on this peptide [10]. Neither of these tasks is trivial, and although in principle any PTM can be detected provided that it leads to a modification in the mass of the peptide, in practice a full mapping of the PTMs of a protein requires full sequence coverage (i.e., detection of all the peptides of the protein). This is not straightforward as typically only a subset of the peptides generated by proteolytic digestion of a protein are detected, unless optimization strategies are used.

2.2.5 Limitations of the Shotgun MS/MS Approach

Although shotgun MS/MS approaches offer a coverage of the proteome that no other technology can currently approach (i.e., about 7,000 proteins can be quantified in an experiment) [10], the technology also shows several limitations. A first limitation is that this depth of analysis typically comes at a high cost in terms of time of experiment (i.e., experimental time typically in days), which limits the ability to interrogate multiple conditions/samples. For this reason, classical shotgun proteomics workflows are better qualified as "high-content" than "high-throughput" experiments. Other fundamental limitations are extreme redundancy and under sampling associated with the method, which result in a saturation effect, i.e., the number of proteins currently identified by shotgun MS is well below the complete proteome [41, 56]. Indeed, since ions are selected at random for fragmentation and MS/MS analysis, the most highly expressed proteins are identified multiple times at the cost of proteins expressed at low level, which dramatically limits the dynamic range of shotgun MS approaches [41]. A typical shotgun MS experiment offers a dynamic range of detection of 3–4 orders of magnitude, whereas it is estimated that the concentration of proteins can vary up to 10 orders of magnitude in human body fluids [23]. Furthermore, owing to the high-redundancy and extreme complexity of the sample, the full spectrum of peptides present is largely under sampled, which in turn means that repeated analyzes of the same or similar biological samples can show distressingly little overlap of identified proteins [41] since each experiment will sample only a subset of the proteins and not necessarily the same subset in each repeat [56]. To some extent, these problems can be overcome by extensive fractionation and multiple enrichment steps, but this requires an additional non-negligible amount of both experimental and computational work [56].

2.2.6 Targeted MS/MS

One way around the limitations of shotgun MS is to adopt a strategy where the mass spectrometer is tuned to analyze specific proteotypic peptides, i.e., peptides that are observable by MS and uniquely identify a target protein. This approach, termed target-driven MS, starts from a list of proteins of interest and carefully selects target peptides for their high propensity to be identified by MS and to uniquely identify a protein or protein isoform of interest. These proteotypic peptides can be identified experimentally (by searching through repositories of observed proteins) or computationally (by predicting them, if the protein has not been previously observed). This type of workflow is called selected/multiple reaction monitoring (S/MRM) and is typically carried out in triple quadrupole type mass spectrometers [41]. The specific proteotypic peptides will be selected in the first quadrupole, then fragmented by collision induced dissociation in the second quadrupole and a second mass filter in the third quadrupole allows for the filtering of the corresponding fragment ions. The identification and quantification of proteotypic peptides is based

on the mass to charge ratios of the precursor and fragment ions pair, which are referred to as "transitions" and are highly specific for a particular peptide [41]. Single reaction monitoring refers to the case where one transition is observed for each peptide, whereas in multiple reaction monitoring, multiple transitions are monitored [41]. In combination with isotope labeling, this technology allows for very accurate, sensitive, and reproducible quantification of the proteotypic peptides that are analyzed. If one can provide the approximate retention time information, then the time of detection of specific transitions can be restricted, therefore allowing for detection of multiple peptides per measurement, a technology referred to as scheduled selected reaction monitoring [41].

Applied to a study of selected proteins in the yeast *Saccharomyces cerevisiae* [56], this technology has been shown to be able to detect and accurately quantify yeast proteins expressed over the full range of cell abundance, from less than 50 copies per cell to over a million copies per cell, without additional fractionation or enrichment steps. This study also demonstrated the capacity of this workflow to comprehensively monitor more than a hundred proteins in a 1 h MS run, which then opens new possibilities for investigating a system under different conditions and replicates. A bottleneck of this workflow, however, is the validation of the SRM transitions that constitute the final mass spectrometric assay in the particular mass spectrometer used for the experiment [56]. Therefore, although targeted MS offers the most sensitive MS detection capabilities to date [23], and unprecedented sample multiplexing capabilities, setting up, optimizing, and validating an assay is relatively time consuming [23, 56]. The accurate mass tag strategy, which is based on the definition (using tandem MS) of peptides whose masses are characteristic of a protein and which can then be detected and quantified by a single MS, can also be used to perform higher throughput targeted MS analyzes [64]. However, this technique suffers from the same drawbacks in terms of time to set up the assay.

3 Computational Analysis of Large Scale Phospho-proteomics Data Sets

Having overcome or mitigated all the challenges mentioned above to collected a good quality high-throughput data set, one faces the challenge of interpreting it, which is not a straightforward task and is practically impossible based on inspection and intuition alone. However, mathematical analysis can provide invaluable help in extracting information, that is, not readily apparent. Various approaches to do so are available, and some of them will be described in this section.

We will start by describing applications of methods derived from machine learning and statistics (such as supervised and unsupervised learning, enrichment analysis, etc.). These methods are mainly used for hypothesis generation (i.e., providing leads for areas of further investigation), and usually generate limited

explanatory or mechanistic insights, but they are relatively straightforward to apply to large and noisy data sets. These methods are also generally unbiased (i.e., hypothesis free) and in this sense are a good starting point in an analysis because they provide a good first overview of the data [43], and do not rely on extensive a priori expert knowledge which might not even be available for the system under investigation.

Another set of approaches that is frequently applied in the same context as the above, is the mapping of data (e.g., differential expression/modification, phenotypic data, etc.) to known or derived "pathway maps." All of these approaches are very familiar to the field of functional genomics and in that sense these methods are quite mature and well known. These types of analyzes have recently been applied quite extensively to investigate large scale phospho-proteomics MS experiments in various settings, which is what we will discuss in the first part of this section. We will refer to these methods as "descriptive" approaches.

Although the methods mentioned above have the potential to generate useful hypotheses, they do not address a fundamental functional characteristic of signaling systems, which is the ability to process information (input) and produce a response (output). To study this process, we need to generate more detailed and hopefully more realistic models of what happens in the cell when a signal is processed. Such models include, but are not limited to, partial least square regression, ordinary and partial differential equations (ODE/PDE), Bayesian networks, rule-based, and logic-based models. We will refer to these formalisms as "predictive approaches". These models are predictive in the sense that given a set of conditions that was not present in the data used to build the model, they should be able to predict the behavior of the system. These methods usually generate explanatory and mechanistic hypotheses (although the actual mechanisms are described with broadly variable levels of details and therefore so are the insights generated, so care should be taken when interpreting them). There are many ways to look at and classify different types of modeling approaches, and all of them are somewhat artificial. The distinction that we make between "descriptive" and "predictive" models is only made for organizational purposes and is not intended as an absolute or universal classification (Fig. 2.2).

3.1 "Descriptive" Approaches

In this section, we will describe some "data-driven" approaches to signaling networks that have been applied to MS and affinity based large scale phospho-proteomics data sets, and briefly mention some of the insights that have been extracted from these analyzes. Whereas affinity-based data sets are now extensively used to generate complex quantitative models, MS proteomics data sets are mainly still at the stage where descriptive investigations are a necessary first step to make sense of the wealth of information that is generated.

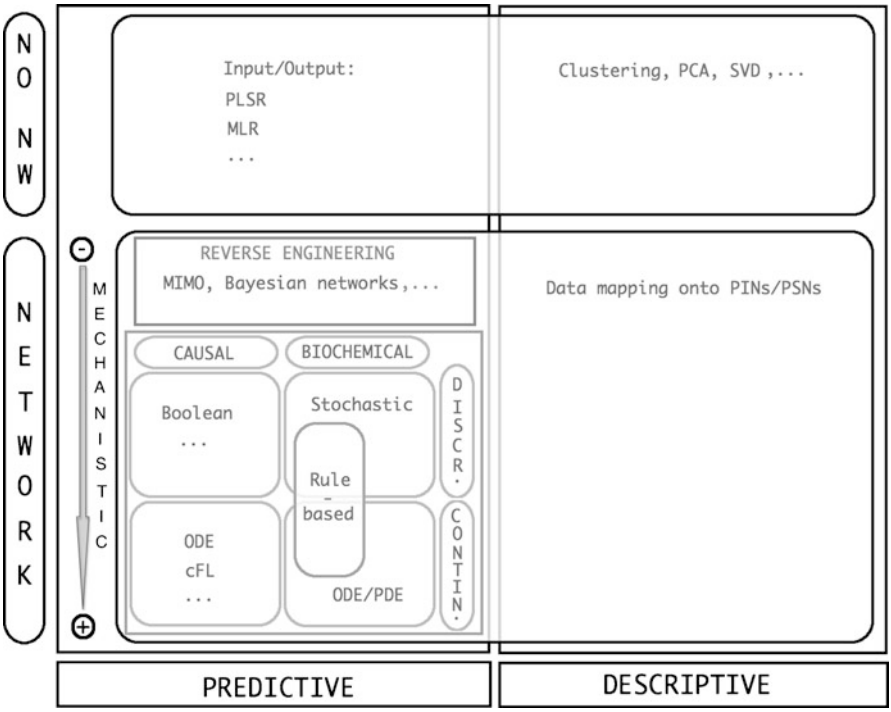


Fig. 2.2 Model formalisms that have been applied to signaling networks. Although there are many ways to look at and classify these methods, we chose to distinguish between "descriptive" and "predictive" approaches, on the basis that the latter allow prediction of a system's response given a set of conditions and data on predictor variables. Within these categories, we further distinguish between methods that rely on or produce a network (i.e., a graph composed of nodes and edges), and methods that do not. In the group of predictive network-based approaches we can further distinguish between approaches that do not necessarily rely on previous knowledge about the system's connectivity (reverse engineering), and methods that rely on some sort of previous knowledge. The latter group can be further divided into 4 categories depending on their discrete/continuous nature and on their causal/biochemical character. Causal approaches only seek to determine relationships between species (such as protein A activates protein B) whereas biochemical approaches include some degree of mechanistic description of the reactions at play. PLSR = partial least square regression, MLR = multiple linear regression, PCA = principal component analysis, SVD = singular value decomposition, MIMO = multiple inputs multiple outputs models, ODE = ordinary differential equations, PDE = partial differential equations, cFL = constrained fuzzy logic, PINs = protein interaction networks, and PSNs = protein signaling networks

3.1.1 Global Investigations of the Phospho-proteome

In view of the highly complex task of making sense of high-throughput phospho-proteomics data in a signaling context, several tools have been developed specifically for this type of data, such as PTMscout [50], NetworKIN [40] or the PHOSIDA [21] database, amongst others. PTMscout is a web-based interface for viewing,

manipulating, and analyzing high-throughput PTM data. Analysis capabilities focus on hypothesis generation through subset selection and enrichment analysis based on annotations (such as GO, Pfam, local sequence features, etc.) or user-defined criteria on dynamic profiles [50]. This tool also provides help in the assignment of peptides to proteins by providing orthogonal information such as annotations and mRNA expression when available.

NetworKIN [40] is an algorithm for prediction of kinases from experimentally determined phosphorylation sites, that integrates sequence specificity with contextual information extracted from resources such as interaction and pathway databases, literature mining, mRNA expression studies, etc. The improved accuracy of this prediction algorithm compared to non-contextual versions indicates that effects such as subcellular compartmentalization, anchoring proteins, temporal, and cell specific expression, etc., play a crucial role in determining kinase-substrate specificity. This in turn points again to the fact that signaling is a highly context specific concept, and that a network level understanding of kinase activity is likely to be necessary even when it comes to understanding single molecular events.

PHOSIDA [21] is a phosphorylation site database for large scale and high confidence quantitative phospho-proteomics experiments that allows the retrieval and analysis of such data, and includes information on evolutionary conservation as well as a phosphorylation site predictor. Other databases, such as the manually curated phosphorylation site database PhosphoSite [26], offer additional information such as association with diseases and sequence logos. The databases mentioned above include some type of analysis tools, but there are also other data repositories that can be valuable resources for proteomics investigations, such as the PRIDE [72] and Phospho.ELM [15] databases.

An interesting perspective on the global function and investigation of phosphosignaling was recently provided by Bodenmiller [6]. In this study, 97 kinases and 27 phosphatases in yeast were systematically knocked out or inhibited, followed by phosphopeptide enrichment and label free LC–MS/MS identification of more than 1,000 phosphopeptides showing a significant change in abundance compared to a wild type situation. Analysis of the direct versus indirect effects of these deletions led to the observation that not a single kinase showed exclusively direct effects. Furthermore, analysis of growth speed and morphological features of each deletion strain revealed that the phenotype strength was not necessarily reflected in the magnitude of the effect on the phospho-proteome. Together, these observations reinforce the view that signaling has to be very flexible and redundant to allow the cell to respond to a changing environment, and point to the fact that modulating any branch of a network might not be possible without system-wide adaptations.

3.1.2 Analysis of Pathway Utilization Downstream of Receptors

The study performed by Olsen [54] set the stage for MS analysis of signaling by pointing both at the complexity of the problem and the sparseness of our knowledge of the involvement of phosphorylation in signaling. Using a strategy

combining phosphopeptide enrichment, high-accuracy identification by LC–MS/MS and SILAC, they were able to quantify dynamic changes in phosphopeptides levels at 6,600 sites on 2,244 proteins upon stimulation of HeLa cells with EGF for different times (from 0 to 20 min). In addition to this, this study also includes some spatial information since nuclear and cytosolic fractions of each condition were obtained and analyzed. Using a cutoff of a minimal 2-fold change upon stimulation, the authors determined the sites that were dynamically regulated and performed fuzzy *c*-means clustering (where each point belongs to clusters with a certain degree, depending on its distance to the centroids of the clusters) to identify groups of sites with similar dynamic profiles. The main conclusions of this study included the observation that most in vivo occurring phosphorylation sites had probably not been detected before, and that groups of phosphosites from the clustering analysis do contain functionally related members.

Another important result of this analysis was obtained by looking at phosphosites which map to the same protein. Indeed, the authors observed that 77% of proteins that had a regulated phosphopeptide also had at least one other site whose regulation profile was different (either unchanging or belonging to a different cluster of the above analysis). This underscores the fact that when looking at the degree of phosphorylation of proteins we should always measure site specific events if we want to obtain accurate and functionally relevant information. This also points to the complexity of interpreting phosphorylation data since it seems that phosphorylation can serve different functions at different sites in the same protein. Finally, Olsen et al. [54] noted that only a subset of the proteins found to be dynamically regulated by EGF signaling were known to be involved in growth-factor signaling, which points to potential gaps in our knowledge of even well-studied pathways.

A similar system was investigated by Huang et al. [27] with different goals and methods, with the objective of determining differences in the signaling downstream of a truncated extracellular mutant of the EGF receptor (EGFRvIII, frequently found in glioblastoma multiforme), compared to the wild type EGFR, and depending on the level of expression of the mutant receptor. The workflow of this analysis was as follows: transduced U87MG glioblastoma cell lines expressing differential levels of EGFRvIII were isolated by FACS, peptides from these cell lines were then isolated, stable isotope labeled and mixed. Next, tyrosine phosphorylated peptides were immunoprecipitated and further enriched by IMAC, and finally analyzed by LC–MS/MS. Quantitative phosphorylation profiles were generated for 99 sites on 69 proteins, which were mapped to canonical EGFR signaling cascades. This indicated that signaling downstream of EGFRvIII and wild type EGFR favour different routes, and this is also dependent on the level of expression of EGFRvIII, e.g., cells that highly overexpress EGFRvIII preferentially use the PI3K pathway over the MAPK and STAT3 pathways. Using a self-organizing map, the authors also identified phosphotyrosine sites with similar profiles, which led to the identification of a cluster of sites that significantly increased as a function of EGFRvIII expression. Examination of the members of this cluster led to the hypothesis that the EGFRvIII receptor was constitutively activating the cMet pathway. Finally, quantification of the phosphorylation sites on the receptor itself pointed to differences in regulation

between wild type and truncated receptors. Altogether these observations indicate that although phosphorylation of the EGFRvIII might not be qualitatively different from the wild type situation, quantitative differences at each individual site might have functional implications reflected in different utilization of downstream pathways and therefore different biological responses [27]. This in turn means that a quantitatively accurate model of this system is likely to prove very useful.

The paper by Krueger et al. [37] aimed at determining the tyrosine phosphoproteome of the insulin signaling pathway by stimulating SILAC labeled differentiated brown adipocytes with insulin for various times, then immunoprecipitating phosphotyrosine containing peptides and analyzing them by LC–MS/MS. Thirty three proteins were identified to be significantly regulated, which was confirmed by western blot. By looking at the dynamic profiles and fold activation of the proteins in this candidate list, they were able to generate hypotheses for new insulin induced candidate effectors and to link them with branches of the insulin pathway.

Matsuoka et al. [45] also used MS to investigate phosphorylation events downstream of a cellular signal, this time concentrating on the landscape of the DNA damage response (DDR) mediated by the ATM and ATR kinases. Briefly, they mixed and immunoprecipitated peptides from two SILAC labeled populations of HEK 293T cells, one having been exposed to ionizing radiations, using antibodies to phospho-SQ or phospho-TQ (ATM and ATR recognise Ser–Gln and Thr–Gln motifs). The samples were then subjected to LC–MS/MS and 905 phosphorylation sites on 700 proteins were shown to display a more than four fold increase following DNA damage by ionizing radiation. This list of proteins was then examined manually, and mined for enriched GO annotations and functional modules using the softwares from Ingenuity. This showed an enrichment for proteins involved in nucleic acid metabolism, and revealed many clusters of proteins previously known to be interacting, but not necessarily known to be involved in the DDR. A subset of the proteins in this list, that were not previously know to be involved in the DDR, was also examined for functional involvement in this response using siRNAs. Although the approach applied here cannot formally distinguish between direct targets of ATM and ATR kinases and targets of kinases with similar specificity, all identified phospho-sites are likely to be regulated by the DDR, and their belonging to a large number of interconnected functional modules suggests an impact of the DDR on cellular physiology that is far broader than expected [45].

More recently, phospho-proteomics was again used to generate qualitative hypotheses using pathway enrichment, this time with the objective to investigate signaling events downstream of the mutant protein NPM-ALK, which is common in positive anaplastic large cell lymphomas [76]. GP293 cells were transfected with either NPM-ALK or a NPM-ALK mutant with decreased tyrosine kinase activity (used as a negative control), the phosphopeptides were then purified and subjected to LC–MS/MS. This led to the identification of 506 phosphoproteins present only in NPM-ALK expressing cells, from which a pathway enrichment analysis was performed (using a Fisher exact *t*-test). The samples were also hybridized to antibody arrays and differential phosphorylation was used as a basis for pathway

enrichment. Both methods resulted in a substantially overlapping list of enriched pathways, from which the authors chose to focus on the TNF/Fas/TRAIL pathway, performing various validations of the involvement of this pathway (comparison with a list of previously generated potential binding partners of NPM-ALK, western blot quantification of three proteins in this pathway, and siRNA knock down of two of those, combined with a viability assay of the knock down of TRAP1 upon drug treatments). This study again underlines the ability of MS data to generate qualitative hypothesis regarding signal transduction.

3.1.3 Analysis of Reciprocal Signaling in Cell–Cell Communication

The approach adopted by Jorgensen et al. [33] is slightly different and addresses a fundamental biological fact, that is, signaling usually happens within the context of tissues and often involves multiple populations of cells. This is particularly important when the signaling is initiated by cell–cell contact, as in the case of the ephrin–EphR interaction. In such cases, the signaling typically involves the reciprocal exchange of distinct information between the interacting cells, leading to mutually coordinated alterations in their respective behaviors. Therefore, stimulating such systems with soluble versions of the ligands is an artificial setting that might provide only limited understanding (e.g., signaling between EphR and ephrin expressing cells might be influenced by interactions with adhesion molecules).

Therefore, in this study, EphB2 and ephrin-B1 expressing populations of HEK293 cells were SILAC labeled and co-cultured for 10 min, then lysed and mixed with non-stimulated EphB2 expressing cells as a reference, before phosphotyrosine peptide isolation and LC–MS analysis. This led to the identification of 442 sites on 304 proteins that significantly decreased or increased in abundance upon stimulation, in one or both cell types, revealing common and cell specific modes of regulation. The authors then turned to a siRNA screen in which monitoring of the cell sorting response when mixing the two cell populations (when mixed, EphB2, and ephrin-B1 expressing cell populations form distinct colonies with well defined boundaries) allowed them to propose a list of proteins involved in this phenotypic response. Using the NetworKIN [40] and NetPhorest [46] tools, a network was constructed based on the prediction of kinases, phosphatases, and phospho-binding modules for each phosphotyrosine that was found to be modulated upon cell–cell contact. These predictions were then pruned based on criteria from the MS and siRNA analyses, and other information such as protein interactions. The obtained network was then represented in a cell-population specific way using the modulation of phosphotyrosine sites determined by the MS analysis. Finally, the MS experiment was repeated using a variant of ephrin-B1 that lacked the cytoplasmic tail, thereby impairing its ability to relay the signal inside the ephrin-B1 expressing cells, but not its ability to interact with EphB2. A significantly different response was observed in the EphB2 expressing cells in this case compared to when the full ephrin-B1 was used, thereby confirming that there is a bidirectional signaling process at play in the

system. This study not only demonstrates the power of MS to investigate complex signaling systems, but also points to the limitations of the in vitro systems in which we commonly conduct our investigations.

3.2 "Predictive" Approaches

In this section, we will describe more detailed and predictive approaches to modeling of signaling networks that have been applied to proteomics data sets, mostly acquired using affinity based technologies. By "predictive" we mean that these models are often capable of computing the expected state or evolution of the system when under particular conditions (e.g., when applying an inhibitor against one of the species in the model). We will start with simple (linear) regression based models that can predict some variables based on linear combinations of other measurements. We will then briefly touch on other correlation based methods. This will be followed by the presentation of ordinary differential equations (ODEs) as a natural way to describe processes where species of interest are changing as a function of time in a quantifiable manner [5]. Then, in light of the extraordinary combinatorial complexity that often arises in signaling systems, we will discuss alternative methods to model detailed signaling networks, such as logic-based and rule-based approaches. Finally, we will discuss the role of previous expert knowledge in the inference process and briefly present Bayesian networks as a strong statistical approach to deal with this.

3.2.1 Input/Output Regression Based Approaches

Two linear regression based approaches will be described here, partial least squares regression (PLSR) and multiple linear regression (MLR). In PLSR, the data are separated into a set of inputs and a set of outputs, which are then reduced to their principal components and a linear solution is identified that relates the inputs to the outputs. PLSR can be used to determine which inputs display the biggest correlation with outputs for example, and it can also be used to predict the outputs from inputs measured in new experiments. MLR is similar to PLSR but the linear solution is computed directly between the measured variables, without dimensional reduction, which makes its results easier to interpret [2]. However, both models suffer from the same limitation: being linear models, they cannot capture coupled effects and nonlinear phenomena such as saturation, switch like effects, etc. [51].

MLR can and has been used to reconstruct network topology from experimental data, for example, in the study by Alexopoulos et al. [3]. In this paper, primary hepatocytes and HepG2 liver cancer cells were exposed to multiple conditions made of combinations of one of 7 growth factors or cytokines, in the presence or absence of 7 small molecule kinase inhibitors. The level or state of modification of 17 intracellular proteins and 50 secreted peptides were measured using a

sandwich immune assay with the xMAP platform. MLR was then performed to relate signals to cytokine secretion, and to relate cues and inhibitors to signals. The regression weights were then used to draw connections between ligands and readouts which allowed the comparison of immediate-early signaling downstream of 7 transmembrane receptors in normal and transformed hepatocytes [3]. Edges selected based on greatest differential regression weights between hepatocytes and HepG2 cells were selected for further experimental investigation. From this analysis, the authors were able to conclude that the magnitude of responses to stimulations (whether reflected in the intracellular signals or in cytokines secretion) were vastly different between the two cell types and that even when both cell types are responding to the same ligand, the extent to which specific downstream pathways are activated is very different.

In the work by Gaudet et al. [20], PLSR was used to extract information from a vast compendium of data acquired from multiple assays such as kinase activity, quantitative immunoblotting, and antibody microarrays. Briefly, HT-29 cells were treated with TNFα, in combination with EGF or insulin, and 19 protein signals were measured over 24 h, along with 4 different measurements of apoptotic response measured by flow cytometry. PLSR is then used to relate signaling data to apoptotic responses. The authors showed that the model derived from the full compendium and a set of metrics derived from the time course data performed extremely well when assessed by leave one out cross validation and independent validation on a new data set. The authors also showed that models built on single protein measurements were poorly predictive, and more surprisingly that models built on measurements of multiple signals from single types of assays were also inferior. Furthermore, models built from the raw measurements only performed poorly on the validation data set, whereas models built only from the derived metrics capturing the time dependent profiles of the signals performed as well as the full model. This points to the fact that time-dependent information is crucial to the predictive power of the model. Finally, they showed that models based on data obtained with cells exposed to multiple combinations of cytokines are less sensitive to experimental noise. In the related study by Janes et al. [29], the contribution of single proteins to the apoptotic response was investigated, and the proteins JNK1, MK2, and ERK were found to provide the most information for prediction of the apoptosis status, based on the average information contained in their derived metrics. The authors also noted that prediction efficiency was maximal with 4–5 signals, and that a model derived from signals measured only in the first 4 h after stimulation (before the onset of apoptosis) were already sufficient to predict the apoptotic signature.

3.2.2 Network Inference

Many methods to build models of signaling networks rely to some extent on previous knowledge about the system under investigation (e.g., a fully detailed mechanistic description of the process at hand in mechanistic models, or a simple description of the logic interactions involved in logic-based models). Building such

models involves a literature (or database) search which is not only usually heavy in terms of workload, but it is also error prone because many molecular events are context specific and the context of an interaction is not necessarily reported. This also biases investigations towards well studied systems. Some formalisms however allow the reconstruction of signaling networks entirely from data, without relying on any type of mechanistic knowledge.

The regression based methods presented above require only very limited prior knowledge, i.e., determining which variables are dependent and which ones are assumed to be explanatory. These methods do not rely on any graph (network) structure, but only predict some variables based on their statistical dependency with others. Interestingly, in the context of the DREAM initiative (www.the-dream-project.org) when confronted with the challenge of predicting unseen measurements of proteins and/or cytokines for combinations of stimuli and inhibitors of a signaling pathway (based on measurements of the same players under different combinations of the same stimuli/inhibitors), methods that performed the best used a statistical approach that did not rely explicitly on an underlying signaling network [57]. Duvenaud et al. [16] also reported that functional causal models that predict the effects of actions on the system (as conditional density models) without relying on any graph tend to perform well or better than methods for learning conditional density models based on graphs. There are many other correlation based methods that can be applied to signaling networks, and many of those have been developed for gene regulatory networks (see the following for reviews [4, 44]). One should be aware, however, when interpreting such analyzes that a correlation does not necessarily mean a causal link, and that correlations can encompass both direct and indirect interactions.

In Ciaccio et al. [11], for example, the algorithm ARACNe [42], which was originally developed for microarray expression profiles, is applied to the analysis of a data set on 91 phosphosites on 67 proteins at 6 time points after stimulation with 5 EGF concentrations, obtained using microwestern arrays. The algorithm uses information theoric approaches to prune indirect interactions inferred by co-expression methods. In Santos et al. [63], an approach called modular response analysis [36] is used to determine the MAPK network architecture in the context of NGF and EGF stimulation. This method is a sensitivity analysis based process relying on measuring network responses to successive small perturbations (here implemented by RNAi), at steady state conditions. Network connections are inferred by computing local response coefficients, which estimate the sensitivity of one module of the network to perturbation of another module, in isolation of the total network [63]. Although in this analysis the system studied is much smaller than those interrogated using high-throughput proteomics, similar approaches could be used to study larger systems.

In the work by Nelander et al. [51], a methodology is proposed to derive network models from time courses of evolution of molecular species upon perturbations. This works builds upon the type of models called multiple inputs multiple outputs (MIMO) models where the time dependent evolution of activities of the system's components (outputs) are described by differential equations as nonlinear functions

(transfer functions) of themselves and a vector of perturbations (inputs). Within the nonlinear transfer function the dependencies between elements of the system are described as linear combinations of the components. The coefficients of these linear dependencies can be interpreted as strength of interactions between the nodes, assuming that they reflect underlying causal relationships between the components, thereby making it possible to derive a node–edge representation of the inferred system (where an edge is present when this strength of interaction is above a certain level) [51]. This representation (as any purely data-driven) has the disadvantage that the nodes in the model are the perturbed and observed molecular species only, which might not be identifiable as single molecular species, and it ignores any unperturbed and unobserved species that might be involved in the connectivity structure of these nodes [51].

3.2.3 Bayesian Network Inference

Bayesian networks have the natural ability to accommodate previous knowledge to a chosen extent. Depending on the level of information that one wants to put in the prior of the models (see below), one can make the inference process entirely independent of any prior knowledge (flat prior) or bias the inference towards models that are casted "more likely" based on a priori expert knowledge. A Bayesian network consists of a directed acyclic graph with vertices representing the molecular species to be modeled as random variables, edges describing conditional independencies between those variables, and parameters describing the conditional distributions implied by the graph (e.g., when the states are discrete, this typically takes the form of a probability for a target node to take each of its possible states given all possible combinations of states of its parents nodes). The graph structure implies that each variable is conditionally independent of all non-children nodes given its immediate parent nodes [49]. Bayesian network inference aims at making inferences regarding the structure of the graph using Bayes' theorem, which states that the posterior probability of a graph (probability of a graph given the data) is given up to proportionality (i.e., ignoring a normalizing constant when comparing structures obtained from the same set of data and distributional assumptions) by the product of the marginal likelihood (probability of the data given the graph) and the prior distribution over directed acyclic graphs (i.e., how likely is each individual graph structure) [49]. Using certain distributional assumptions, the posterior probability of graphs can be computed up to proportionality, which is enough to compare graphs in a search procedure, in order to find a graph structure that is optimal under the statistical model at hand.

Given their solid basis in statistics, Bayesian networks are naturally able to handle stochastic aspects of biological processes and noisy measurements [31]. However, this comes at a high cost in terms of data requirements. Such an approach is, however, ideal when the data at hand is cell specific and therefore each measurement includes data about a whole population of cells at the single cell level, as is the case in the study by Sachs et al. [59] where intracellular multicolor

flow cytometry is used. In this paper, Bayesian network inference is used to investigate signaling networks of human primary naive CD4+ T cells, downstream of CD3, CD28, and LFA-1 activation, based on measurements of 11 phosphorylated proteins and phospholipids [59]. Similarly in [11] Bayesian networks are used to model the dependencies between 67 proteins (measured by microwestern arrays at 6 time points) after stimulation with EGF at 5 different concentrations. In this case, in order to have enough samples for each measurement, each time point and concentration of the stimulus is used as an independent sample, yielding 20 samples per measurement. One of the strength of Bayesian networks in this context is that, using carefully chosen prior distributions on graphs, it is possible to include information on network features such as particular edges, types of edges, degree distribution, and sparsity. In practice this means that not every possible graph is considered equally plausible, and that we can bias the search towards graphs that we consider a priori more likely [49]. This in turn has the advantage of constraining the space of possible graphs to search, which makes the inference process more efficient, while maintaining the Bayesian networks' natural ability to deal with noise and stochasticity. The expert knowledge involved in specifying those priors can be as detailed as specifying a particular edge to be very likely or as vague as specifying that ligands should generally interact with receptors and not effectors [49].

When interpreting Bayesian networks it is important to be aware that many Bayesian networks can represent the same statements of conditional independence, i.e., the inference process can be unable to distinguish among a series of graph with the same undirected graph but in which some edges might have different directions [44]. However, perturbing the states of measured molecules with molecular interventions can help resolving this problem by providing information on the causal relationships between nodes [59]. Furthermore, a limitation of Bayesian networks is that they are constrained to be acyclic, which means that feedback loops for example cannot be uncovered. However this limitation can be overcome by using dynamic Bayesian networks [59].

3.2.4 Reaction-based Models

All of the models described above infer a topology as statistical dependencies between variables, not mechanistic links. If some mechanistic knowledge about the topology of the system is available, then other methods can be applied that incorporate this information. An extreme case compared to network inference is the application of ODE/PDE models where detailed knowledge about the the biochemistry (reactions) of the system is written down as a set of differential equations.

Biochemical (also called physicochemical [1]) models describe the temporal evolution of individual biomolecular species as functions of their rates of production and consumption in terms of mass action kinetics, which is an empirical law expressing the rates of reactions as proportional to the concentrations of their reactants [1]. In the simplest case one uses ODEs, and spatial heterogeneity (i.e., changes

in the location of species) is represented by compartmentalization, where each compartment is assumed to be perfectly well mixed (i.e., instantaneous transport inside a compartment, leading to homogeneous concentrations of all species across the whole compartment). Partial differential equations (PDEs) arise when the spatial dimension is explicitly modeled, i.e., spatial gradients are now included in the representation. Building an ODE/PDE model involves three main steps, often applied in an iterative way: model development (write down biological knowledge in terms of rate of change equations), parameter estimation (determining the values of unknown parameters), and model validation (comparing model predictions to independent experimental data) [5].

When designing the models, two critical decisions need to be made: what is the scope of the model and at which level of detail will the system be described [9]. Defining the scope involves determining how much of the system needs to be modeled in order to achieve the goal of the modeling process, and deciding on the level of detail involves choosing a level of representation of the molecular species and complexes (i.e., do we want to represent all modifications and interactions explicitly). The latter point is especially challenging because biological species are often capable of assembling into multi-component complexes, undergoing multiple PTMs, and segregating into various sub-cellular compartments and locally concentrated areas, and we often do not know how to interpret these events in terms of signals. This latter problem is referred to as "combinatorial complexity" and is what quickly makes ODE and PDE models untractable [7, 12, 17, 18]. Another common problem with ODE/PDE models is parameter estimation, which involves determining the range of parameter values over which the model closely reproduces the experimental data [1]. Problems arise in this process when the model reproduces the experimental behavior equally well over a large range of parameter values, therefore making those parameters unidentifiable.

Some common simplifying assumptions are made to overcome the problem of combinatorial complexity, such as ignoring intermediate states of assembly when they are fast, or lumping together biochemical forms that are thought to be equivalent. However, these remain assumptions, and just as any other assumption made in building the model (e.g., well mixed compartments, etc.) it is very important to be aware that the equations obtained are only valid given all of the assumptions made, and so each assumption and the implications thereof should be discussed, in light of explicitly stated design goals [1,5]. It is also important to note that ODE and PDE models are deterministic continuum approximations of what happens in the system [9]. When limited number of molecules are involved in a process (e.g., small compartments or slow reactions), then stochastic effects may become important and a deterministic approximation might not be able to accurately represent the evolution of the system [1].

Despite these complications, ODE and PDE models can be used to generate valuable insights into biological questions. In Birtwistle and Kholodenko [5], for example, the authors describe how simple and more complex ODE and PDE models can be used to gain insights into the role of endocytosis in signaling. In the paper by Chen et al. [9], for example, a detailed model of ERK and Akt

regulation by two ErbB ligands and four ErbB receptors during the immediate-early phase of ligand stimulated cell signaling is built, parameterized, and analyzed. This model includes 28 proteins, but accounting for protein–protein interactions, PTMs, and compartmentalization generated an additional 471 species, requiring 499 differential equations, 201 unique reaction rates, and 28 non-zero initial conditions. This leads to a complex parameter optimization problem, despite some parameters being measured or extracted from the literature. Other parameters were estimated from the data by minimizing the difference between experimental and simulated data [9].

This model was found to be unidentifiable, with some parameters being quite constrained across similarly performing models (in terms of fit to data), and some parameters spanning the entire range of values allowed in the search. However, the authors were still able to perform a sensitivity analysis of the partially calibrated models (i.e., an investigation of which parameters have the largest influence over a chosen observable, when varied), as well as a dose responsiveness analysis, and to extract useful predictions from those analyzes. For example, they showed that the calibrated $v_{\max}$ for the PP2A compartment targeting pRAF and pMEK was markedly different from the compartment targeting pAkt, which led them to hypothesize that dephosphorylation of Raf, MEK, and Akt occurs at different rates, and that this presumably involves different PP2A-containing complexes. The sensitivity analysis also yielded valuable insights, such as which parameters have the biggest influence on EGF- or HRG-stimulated pERK across multiple partially calibrated models, and the observation that parameter sensitivity critically depends on the observable that is chosen [9]. This shows that, provided that care is taken in interpreting the results of an analysis, and that parameter uncertainty is considered in this process, even partially calibrated models can provide valuable insights.

3.2.5 Rule-based Models

A formalism that naturally describes the mechanisms of signaling systems despite their associated combinatorial complexity is the principle of rule-based model. A rule-based description of a system allows a rich variety of knowledge about this system to be expressed in a single formalism (see [25] for a review) [31]. Briefly, the system is described as a set of agents which have labeled sites that can each have an internal state, typically used to denote PTM status. The agents are acted on by rules, which provide descriptions on how they interact, with common interactions consisting of binding/unbinding of agents, modification of the state of a site, and deletion/creation of an agent. The left hand side of a rule specifies a condition that applies on a pattern of agents and their site values, whereas the right hand side specifies actions on agents mentioned on the left. Only the information that is triggering the accomplishment of the rule needs to be specified on the condition side of the statement [14].

Simulation of a rule-based model can be performed by the repeated process of matching the facts (patterns of states of agents) against the condition part of

the rules and carrying out the action part of the rules where the condition part is satisfied [31]. A control strategy is used to determine the order in which the rules are applied, which typically takes the form of a rule-based version of Gillepsie's algorithm [14, 31, 65]. Popular languages to write and simulate such models are Kappa [14], and BioNetGen [17] which is extended in the software NFsim [65]. Differential equations can be also derived from the rules; if all possible species are described the number of states increases exponentially due to the combinatorial explosion, but methods exist to simplify them at least to some degree [7, 12, 18].

3.2.6 Logic-based Models

Whether as reactions or rules, building a biochemical model requires a lot of mechanistic information about the system, and the resulting models are difficult to simulate. This limits their applicability to relatively small and well studied systems. However, data generated by high-throughput methods typically provide wide scope information which leads to the need for formalisms capable of handling big networks for which only limited mechanistic knowledge is available. A suitable formalism to model large networks for which some mechanistic knowledge is available are logic-based models, that include dependencies between components, while ignoring the molecular details [31, 47, 74]. In logic-based models dependencies between nodes are specified in terms of gates, which are associated with truth tables that describe output states for all possible combinations of input states [47]. If two proteins A and B have a positive effect on the activation of a third one C, the corresponding gate can be either an OR (either A OR B activates C) or AND (only A AND B together activate C).

The simplest type of logic model is a Boolean model, in which each state is either on or off (1 or 0). Following the pioneering work by Kauffman [35], Boolean logic models have been used extensively to model genetic regulatory and signaling networks [31, 47, 74]. This formalism allows one to compute the state of activity of each node of a graph given different inputs or initial states. Cause–effect relationships in biological pathways can often be found in the literature, and in databases such as reactome (http://www.reactome.org) or panther (http://www.pantherdb.org). However, these resources rarely include specific gates, nor cell-type specific information. This problem can be overcome by using signaling data to train a Boolean model from a generic prior knowledge network derived from the literature or databases [61]. By pruning the network, one obtains models with a much higher predictive power, that are specific to the data (and thus cell-type) they have been trained to. Thus, by leveraging prior knowledge and dedicated signaling data, one can model relatively large networks with relatively sparse data, and because one includes intermediates (not just perturbed or observed variables), the mechanistic insight is higher than in purely data driven models. Thanks to their simplicity, these models can easily accommodate $\sim$100 nodes and be trained to phospho-proteomics sets of $\sim$1000 data points [61].

A main limitation of the boolean logic approach is that all species are considered either on or off, and the model is therefore not able to account for intermediate levels of activation. Fortunately, several logic-based extensions provide a means to do so, such as multi-state discrete models and fuzzy logic [47]. In multi-state discrete models, additional levels between 0 and 1 are specified, whereas in fuzzy logic a set of user-defined functions are used to transform discrete logic conditions into relationships between continuous inputs and outputs. An extension of the approach from [61] was recently proposed that allows the training to data of a fuzzy logic model obtained from previous knowledge [48]. The approach is termed "constrained fuzzy-logic" because the set of relationships between model species is limited, thereby making it possible to train both the topology of the network and the particular quantitative relationship involve at each gate, and allows to model features not captured by Boolean logic [48]. However, this ability comes together with an increase in complexity that renders the approach more difficult to apply to large networks (above a few dozen species) [48].

Both of the approaches described above compute a steady state of the logic model. However, logic-based models can integrate the notion of time, with various degrees of detail. To compute a trajectory of the system, the status of nodes are updated (as functions of the sate of their input nodes) at each (time) step according to two main updating schemes: synchronous, where all nodes are updated simultaneously with a new state depending on the state of each node's inputs at the previous time step, and asynchronous, where nodes are updated in random order with a new state depending on the state of some input nodes at the previous and some at the current time step [47]. Mixed asynchronous schemes allow some nodes to be updated before others, making it possible to model separate time scales. Logic models can also be converted into ODEs, making both species and time continuous, albeit at the cost of increased complexity [47].

Finally, logic-models can be extended to incorporate probabilistic interactions, thereby incorporating uncertainty in biological knowledge and/or stochasticity of the system [47]. Logic-based models can also be implemented in a Bayesian framework (see [19] and [39] for more information).

4 Summary

Modeling is an invaluable tool to make sense of large and/or complex systems from a functional perspective. Signal processing involves regulations on the proteome at three highly regulated and coordinated levels: regulated post translational modifications (PTMs), protein–protein interactions, and changes in expression levels. The PTM level is what we focus on here because it is the most immediate one and often triggers changes at the other levels, and in particular we concentrate on phosphorylation as a major regulator of protein function and activity. Many proteins are modified at many sites in a highly dynamic and context dependent manner, and combinations of modifications can have various functional consequences that we

are only beginning to unravel. Therefore, interpreting PTM data from a signaling perspective is still a significant challenge, and investigations in this area are likely to benefit from modeling approaches.

A modeling process requires a data set and an appropriate modeling framework. Tables 2.1 and 2.2 summarize the main features of particular applications of modeling pipelines that have been mentioned throughout this chapter. Antibody based approaches allow the quantitative measurement of protein or protein modification levels using technologies such as protein arrays, reverse phase arrays, xMAP, intracellular multicolor flow cytometry, and microwestern arrays. All of these platforms have different limitations in terms of samples and targets multiplexing, signal to noise ratios, and dynamic range. However, they are all based on the recognition of a target by a specific antibody and therefore all suffer from the same limitation: the availability and quality of antibodies. The ability of these methods to be applied to many samples in parallel is a significant asset because it allows for multiple perturbation experiments that inform the network inference process. Compared to MS approaches, antibody based technologies offer limited protein coverage but are more easily scalable to large number of samples [2, 62], and in general require a smaller amount of sample [11].

In contrast, the classical mass spectrometry workflow (shotgun MS) is a non-biased approach (i.e., it is not aimed at particular proteins) that allows the detection of many more proteins in a single experiment. The unit that is identified in LC–MS/MS workflows is a peptide, and peptide mapping to proteins is a non-trivial problem, especially in higher eukaryote organisms where a high level of sequence redundancy can be expected, thus the importance of rigorous statistical approaches for assessing protein identification. Shotgun MS is inherently biased towards peptides that are highly abundant and easy to detect, and selection of peptides for MS/MS is a random process that undermines the reproducibility of shotgun LC–MS/MS approaches. Shotgun MS is somewhat limited with regards to the number of samples that can be processed. Targeted MS is likely overcome some of these limitations, since this technology has the ability to be highly quantitative and reproducible, with an unprecedented dynamic range and the ability to investigate many conditions. However, this method requires a long time for workflow optimization, which means that a significant workload investment has to be done before being able to collect data. With the advances in instrumentation and the emergence of targeted proteomics workflows, MS now has the potential to represent a viable and a more powerful, fully quantitative alternative to antibody based methods. Therefore we expect MS to play a crucial role in the field of modeling of signal transduction networks in the future, provided that modeling frameworks are adapted to the particular features of such data sets.

Whatever the method used, it is important to systematically document and report the pipeline that is used from the data collection to modeling (e.g., normalization in the case of antibody based methods, peptide identification, protein inference, and quantification in the case of MS). Ideally, both the raw and processed data should be available alongside detailed methods for any reinvestigation or even reinterpretation of results. This is particularly challenging in the case of MS since

Table 2.1 Overview of applications mentioned in this chapter using mass spectrometry

Paper	Goal	Technology	Model	Conditions	Readouts
Bodenmiller et al. [6]	Determine influence on yeast ppome of systematic deletion of kinases and ppase	TiO_2 ppide enrichment, LC–MS/MS	DE, correlation impact on ppome/orthogonal phenotype	124 deletion/analog-sensitive strains + wild type = 125	8814 peptides, 1026 proteins
Olsen et al. [54]	Phosphorylation events downstream of EGF stimulation	SILAC, TiO_2 ppide enrichment, LC–MS/MS	D.E., clustering	5 tp (stimulation time), 2 subcellular fractions	6,600 sites, 2,244 proteins
Huang et al. [27]	Differences in signaling downstream of EGFR mutant	Stable isotope labeling, TyrP-IP, IMAC, LC–MS/MS	Mapping on pathway, clustering	4 cell lines	99 sites, 69 proteins
Krueger et al. [37]	Define tyrosine-ppome of insulin signaling pathway	SILAC, TyrP-IP, LC–MS/MS	DE, mapping to pathway	5 tp (stimulation time)	33 proteins
Matsuoka et al. [45]	DDR mediated by ATM and ATR	SILAC, phosphoSQ and phosphoTQ-IP, LC–MS/MS	DE, functional annotation, enrichment	2	905 sites, 700 proteins

(continued)

Table 2.1 (continued)

Paper	Goal	Technology	Model	Conditions	Readouts
Wu et al. [76]	Signaling downstream of NMP-ALK	IMAC, TyrP-IP + IMAC, LC–MS/MS	DE, pathway enrichment	2	4798 sites, 1548 proteins in the control; 5340 sites, 1758 proteins in the NMP-ALK cell line (506 unique to this cell line)
Jorgensen et al. [33]	Bidirectional signaling in EphB2 ephrin-B1 expressing cells	SILAC, TyrP-IP, LC–MS/MS	DE, kinase/ppase/phosphobinding proteins prediction, map to obtained network	3 cell lines	442 sites, 304 proteins

DE = differential expression (or regulation, in the case of PTMs),
TyrP-IP = phosphotyrosine immunoprecipitation, tp = time points,
pp = phosphoprotein, ppome = phosphoproteome,
ppase = phosphatase, ppide = phosphopeptide

Table 2.2 Overview of applications mentioned in this chapter using affinity based methods

Paper	Goal	Technology	Model	Conditions	Readouts
Alexopoulos et al. [3]	Difference in network topology between primary hepatocytes and hepatocellular carcinoma cell lines	xMAP	MLR	64 comb. of 7 growth factors/ck and 7 small molecule kinase inhibitors	50 ck, 17 pp
Gaudet et al. [20]; Janes et al. [29]	Link apoptotic response to signaling data, investigate aspects of data requirements	Kinase activity assays, quantitative immunoblotting, antibody microarrays, flow cytometry	PLSR	10 comb. of 3 stimuli, 13 tp	19 signaling proteins + 4 apoptotic markers
Ciaccio et al. [11]	Signaling downstream of EGF stimulation	microwestern arrays	ARACNe (correlation + information theory) and Bayesian network inference	5 EGF concentrations, 6 tp	91 phosphosites, 67 proteins
Santos et al. [63]	Architecture of MAPK network in response to NGF or EGF stimulation	Quantitative western blots	Modular-response analysis (sensitivity analysis to small perturbations)	3 siRNAs, 2 stimuli, 2 tp	3 proteins
Nelander et al. [51]	Test method on EGFR/MAPK and PI3K/AKT pathways in a breast cancer cell line	Western blots	MIMO	21 comb. of EGF + 6 inhibitors	3 proteins
Sachs et al. [59]	Test method on intracellular signaling networks of human primary naive CD4+ T cells, downstream of CD3, CD28, and LFA-1 activation	Multicolor flow cytometry (single cell)	Bayesian network inference	9 stimulators/inhibitors	11 pp and phospholipids

(continued)

Table 2.2 (continued)

Paper	Goal	Technology	Model	Conditions	Readouts
Chen et al. [9]	Quantify signal flow through ErbB-activated pathways	Multiple kinetic parameters from literature, quantitative immunoblotting,	ODEs	10 tp, 2 ligands, 3 cell lines	3 proteins
Saez-Rodriguez et al. [61]; Morris et al. [48]	Test method, understand cooperative/antagonistic interactions among ligands in hepatocellular carcinoma cells	xMAP	Boolean logic [61]; constrained fuzzy logic [48]	64 comb. of 7 ck and 7 small molecule kinase inhibitors	16 proteins

tp = time points, ck = cytokine,
pp = phosphoprotein, comb. = combinations

reporting raw data involves finding an appropriate way to store thousands of spectra (and accompanying LC retention times, and metadata) for each experiment [53]. The workflow from experiments to models can encompass multiple steps and a number of tools are available to develop data processing pipelines while maintaining the consistency of the workflow and keeping data provenance [60, 62], allowing connection with multiple modeling methods. Equally important is the development of and compliancy to standards for capturing, representing, annotating, and reporting the data and models. This should facilitate effective quality assessment, promote transparency, and enhance accessibility [69].

"Descriptive" modeling approaches mainly rely on methods from statistics and machine learning, and include for example differential expression analysis usually followed by clustering or mapping onto known networks. "Predictive" models are capable of providing estimates of the behavior of a system under a set of conditions that were not used to build the model. A very simple way to do so is using regression approaches such as PLSR, which links linearly correlated variables but do not provide mechanistic information, and can only capture linear phenomena. More detailed models can be built that include mechanistic and/or causal relationships between elements of the system that can be represented by a graph ("wiring diagram"), such as differential equations, logic-based, rule-based, or Bayesian network models. Models built upon proteomics shotgun MS data sets that have been reported so far generally belong to the descriptive category. In addition to context specific knowledge, large scale phospho-proteomics analysis by MS have generated valuable insights into the dynamics and characteristics of phosphorylation networks in signaling, which are opening new avenues for investigation.

Affinity based approaches on the other hand have produced data that have allowed extensive modeling of various (mainly well studied) systems. Although biochemical descriptions based on differential equations can provide a detailed and accurate description of signal transduction, they suffer from limitations when handling large systems, in particular due to the combinatorial complexity arising from signaling systems. Coarse graining of the system and simplifying assumptions provide ways around this but ODE/PDE models are still limited to systems of a couple of dozens of nodes. Rule-based models handle the combinatorial complexity by defining sets of rules that apply on biomolecular patterns without having to account for the full context of those patterns. This has the advantage of representing mechanistic knowledge (and assumptions) in an intuitive and explicit way, and allowing heterogeneous types of information to be incorporated into the model. However, rule-based models can only be applied to well studied systems because they rely entirely on an accumulated knowledge.

Methods that represent the system with lower level of details can provide alternatives to model bigger, not-so-well-known systems. Logic-based approaches for example represent only logical relationships between nodes in a network, and are therefore conceptually simple, computationally cheap, and causally correct [74]. Boolean logic models are limited to on/off representations of systems, but extensions such as fuzzy logic overcome this problem, albeit at the cost of increased complexity and therefore limitations in the size of the system that can be

interrogated. Finally, Bayesian networks provide a strongly statistically grounded alternative to infer signaling networks when little information about the system is available (although various levels of previous knowledge can be incorporated in the inference process). Bayesian networks can handle noise and stochasticity in the data in a natural way, but require rich data sets, which has limited their application so far to relatively small systems.

Whatever the biological question, it is very important to ask oneself the following questions before building a model: what is the scope and level of detail that I can and should model in order to (1) account for the limitations of my dataset and (2) reach the goal of my analysis. An adequate solution relies on choosing a formalism with the right level of detail to answer our question, and which yields the most interpretable results for the problem under investigation [47]. When interpreting the results of a model, it is also very important to be aware that a model is only as true as its assumptions, and that every methodology has limitations inherent to the way that they build, represent, and simulate the system.

References

1. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. Nat Cell Biol 8(11):1195–1203, DOI 10.1038/ncb1497, URL http://dx.doi.org/10.1038/ncb1497
2. Alexopoulos LG, Saez-Rodriguez J, Espelin CW (2009) High-throughput protein-based technologies and computational models for drug development, efficacy, and toxicity. John Wiley and Sons, Inc., New Jersey, pp 29–52. DOI 10.1002/9780470431818.ch2, URL http://dx.doi.org/10.1002/9780470431818.ch2
3. Alexopoulos LG, Saez-Rodriguez J, Cosgrove BD, Lauffenburger DA, Sorger PK (2010) Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. Mol Cell Proteom MCP 9(9):1849–1865, DOI 10.1074/mcp.M110.000406, URL http://www.ncbi.nlm.nih.gov/pubmed/20460255, PMID: 20460255
4. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007) How to infer gene networks from expression profiles. Mol Syst Biol 3:78, DOI 10.1038/msb4100120, URL http://www.ncbi.nlm.nih.gov/pubmed/17299415, PMID: 17299415
5. Birtwistle MR, Kholodenko BN (2009) Endocytosis and signalling: a meeting with mathematics. Mol Oncol 3(4):308–320, DOI 10.1016/j.molonc.2009.05.009, URL http://www.ncbi.nlm.nih.gov/pubmed/19596615, PMID: 19596615
6. Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, Pedrioli PG, Gerrits B, Picotti P, Lam H, Vitek O, Brusniak M, Roschitzki B, Zhang C, Shokat KM, Schlapbach R, Colman-Lerner A, Nolan GP, Nesvizhskii AI, Peter M, Loewith R, von Mering C, Aebersold R (2010) Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci Signal 3(153):rs4, DOI 10.1126/scisignal.2001182, URL http://www.ncbi.nlm.nih.gov/pubmed/21177495, PMID: 21177495
7. Borisov NM, Markevich NI, Hoek JB, Kholodenko BN (2005) Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity. Biophys J 89(2):951–966, DOI 10.1529/biophysj.105.060533, URL http://www.ncbi.nlm.nih.gov/pubmed/15923229, PMID: 15923229

8. Brosch M, Choudhary J (2010) Scoring and validation of tandem MS peptide identification methods. Meth Mol Biol (Clifton, NJ) 604:43–53, DOI 10.1007/978-1-60761-444-9_4, URL http://www.ncbi.nlm.nih.gov/pubmed/20013363, PMID: 20013363
9. Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK (2009) Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol 5:239, DOI 10.1038/msb.2008.74, URL http://www.ncbi.nlm.nih.gov/pubmed/19156131, PMID: 19156131
10. Choudhary C, Mann M (2010) Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol 11(6):427–439, DOI 10.1038/nrm2900, URL http://dx.doi.org/10.1038/nrm2900
11. Ciaccio MF, Wagner JP, Chuu C, Lauffenburger DA, Jones RB (2010) Systems analysis of EGF receptor signaling dynamics with microwestern arrays. Nat Meth 7(2):148–155, DOI 10.1038/nmeth.1418, URL http://dx.doi.org/10.1038/nmeth.1418
12. Conzelmann H, Saez-Rodriguez J, Sauter T, Kholodenko BN, Gilles ED (2006) A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. BMC Bioinformatics 7:34, DOI 10.1186/1471-2105-7-34, URL http://www.ncbi.nlm.nih.gov/pubmed/16430778, PMID: 16430778
13. Cutillas P, Jorgensen C (2011) Biological signalling activity measurements using mass spectrometry. Biochem J 434(2):189–199, DOI 10.1042/BJ20101974, URL http://www.biochemj.org/bj/434/bj4340189.htm
14. Danos V, Feret J, Fontana W, Harmer R, Krivine J, Biosystems P, Suprieure EN, Polytechnique E (2007) Rule-based modelling of cellular signalling. Proc of the 18th Int Conf on Concurrency Theory (CONCUR07), Lecture Notes in Computer Science 4703:17–41, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.228
15. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson T (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5(1):79, DOI 10.1186/1471-2105-5-79, URL http://www.biomedcentral.com/1471-2105/5/79
16. Duvenaud D, Eaton D, Murphy K, Schmidt M (2009) Causal learning without DAGs. JMLR J Mach Learn Res URL http://jmlr.csail.mit.edu/proceedings/papers/v6/duvenaud10a/duvenaud10a.pdf
17. Faeder JR, Blinov ML, Hlavacek WS (2009) Rule-based modeling of biochemical systems with BioNetGen. Meth Mol Biol (Clifton, NJ) 500:113–167, DOI 10.1007/978-1-59745-525-1_5, URL http://www.ncbi.nlm.nih.gov/pubmed/19399430, PMID: 19399430
18. Feret J, Danos V, Krivine J, Harmer R, Fontana W (2009) Internal coarse-graining of molecular systems. Proc Natl Acad Sci USA 106(16):6453–6458, DOI 10.1073/pnas.0809908106, URL http://www.ncbi.nlm.nih.gov/pubmed/19346467, PMID: 19346467
19. Gat-Viks I, Shamir R (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast. Genome Res 17(3):358–367, DOI 10.1101/gr.5750507, URL http://www.ncbi.nlm.nih.gov/pubmed/17267811, PMID: 17267811
20. Gaudet S, Janes KA, Albeck JG, Pace EA, Lauffenburger DA, Sorger PK (2005) A compendium of signals and responses triggered by prodeath and prosurvival cytokines. Mol Cell Proteom MCP 4(10):1569–1590, DOI 10.1074/mcp.M500158-MCP200, URL http://www.ncbi.nlm.nih.gov/pubmed/16030008, PMID: 16030008
21. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8(11):R250, DOI 10.1186/gb-2007-8-11-r250, URL http://www.ncbi.nlm.nih.gov/pubmed/18039369, PMID: 18039369
22. Gouw JW, Krijgsveld J, Heck AJR (2010) Quantitative proteomics by metabolic labeling of model organisms. Mol Cell Proteom MCP 9(1):11–24, DOI 10.1074/mcp.R900001-MCP200, URL http://www.ncbi.nlm.nih.gov/pubmed/19955089, PMID: 19955089
23. Gstaiger M, Aebersold R (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet 10(9):617–627, DOI 10.1038/nrg2633, URL http://dx.doi.org/10.1038/nrg2633

24. Harrison C (2008) High-content screening: integrating information. Nat Rev Drug Discov 7(2):121, DOI 10.1038/nrd2522, URL http://dx.doi.org/10.1038/nrd2522
25. Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W (2006) Rules for modeling signal-transduction systems. Science's STKE: Signal Transduct Knowl Environ 2006(344):re6, DOI 10.1126/stke.3442006re6, URL http://www.ncbi.nlm.nih.gov/pubmed/16849649, PMID: 16849649
26. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 4(6):1551–1561, DOI 10.1002/pmic.200300772, URL http://www.ncbi.nlm.nih.gov/pubmed/15174125, PMID: 15174125
27. Huang PH, Mukasa A, Bonavia R, Flynn RA, Brewer ZE, Cavenee WK, Furnari FB, White FM (2007) Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. Proc Natl Acad Sci USA 104(31):12,867–12,872, DOI 10.1073/pnas.0705158104, URL http://www.ncbi.nlm.nih.gov/pubmed/17646646, PMID: 17646646
28. Hyduke DR, Palsson B (2010) Towards genome-scale signalling-network reconstructions. Nat Rev Genet 11(4):297–307, DOI 10.1038/nrg2750, URL http://dx.doi.org/10.1038/nrg2750
29. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB (2005) A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. Science (New York, NY) 310(5754):1646–1653, DOI 10.1126/science.1116598, URL http://www.ncbi.nlm.nih.gov/pubmed/16339439, PMID: 16339439
30. Jensen ON (2006) Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 7(6):391–403, DOI 10.1038/nrm1939, URL http://www.ncbi.nlm.nih.gov/pubmed/16723975, PMID: 16723975
31. de Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol: J Comput Mol Cell Biol 9(1):67–103, DOI 10.1089/10665270252833208, URL http://www.ncbi.nlm.nih.gov/pubmed/11911796, PMID: 11911796
32. Jorgensen C, Linding R (2010) Simplistic pathways or complex networks? Curr Opin Genet Dev 20(1):15–22, DOI 10.1016/j.gde.2009.12.003, URL http://www.ncbi.nlm.nih.gov/pubmed/20096559, PMID: 20096559
33. Jorgensen C, Sherman A, Chen GI, Pasculescu A, Poliakov A, Hsiung M, Larsen B, Wilkinson DG, Linding R, Pawson T (2009) Cell-specific information processing in segregating populations of eph receptor ephrin-expressing cells. Science (New York, NY) 326(5959):1502–1509, DOI 10.1126/science.1176615, URL http://www.ncbi.nlm.nih.gov/pubmed/20007894, PMID: 20007894
34. Joughin BA, Cheung E, Karuturi RKM, Saez-Rodriguez J , Lauffenburger DA, Liu ET (2010) Cellular regulatory networks, systems biomedicine – Chapter 4. Academic Press, San Diego, pp 57–108, DOI 10.1016/B978-0-12-372550-9.00004-3, URL http://www.sciencedirect.com/science/article/pii/B9780123725509000043
35. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224(5215):177–178, URL http://www.ncbi.nlm.nih.gov/pubmed/5343519, PMID: 5343519
36. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7(3):165–176, DOI 10.1038/nrm1838, URL http://www.ncbi.nlm.nih.gov/pubmed/16482094, PMID: 16482094
37. Krueger M, Kratchmarova I, Blagoev B, Tseng Y, Kahn CR, Mann M (2008) Dissection of the insulin signaling pathway via quantitative phosphoproteomics. Proc Natl Acad Sci USA 105(7):2451–2456, DOI 10.1073/pnas.0711713105, URL http://www.ncbi.nlm.nih.gov/pubmed/18268350, PMID: 18268350
38. Krutzik PO, Clutter MR, Trejo A, Nolan GP (2011) Fluorescent cell barcoding for multiplex flow cytometry. In: Robinson JP, Darzynkiewicz Z, Dobrucki J, Hyun WC, Nolan JP, Orfao A, Rabinovitch PS (eds) Current protocols in cytometry. John Wiley & Sons, Inc., Hoboken, NJ, USA, URL http://www.currentprotocols.com/protocol/cy0631

39. Li P, Zhang C, Perkins EJ, Gong P, Deng Y (2007) Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 8 Suppl 7:S13, DOI 10.1186/1471-2105-8-S7-S13, URL http://www.ncbi.nlm.nih.gov/pubmed/18047712, PMID: 18047712
40. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MATM, Jorgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K, Metalnikov P, Nguyen V, Pasculescu A, Jin J, Park JG, Samson LD, Woodgett JR, Russell RB, Bork P, Yaffe MB, Pawson T (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129(7):1415–1426, DOI 10.1016/j.cell.2007.05.052, URL http://www.ncbi.nlm.nih.gov/pubmed/17570479, PMID: 17570479
41. Malmstrm J, Lee H, Aebersold R (2007) Advances in proteomic workflows for systems biology. Curr Opin Biotechnol 18(4):378–384, DOI 10.1016/j.copbio.2007.07.005, URL http://www.ncbi.nlm.nih.gov/pubmed/17698335, PMID: 17698335
42. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7, DOI 10.1186/1471-2105-7-S1-S7, URL http://www.ncbi.nlm.nih.gov/pubmed/16723010, PMID: 16723010
43. Markowetz F (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. PLoS Comput Biol 6(2):e1000,655, DOI 10.1371/journal.pcbi.1000655, URL http://www.ncbi.nlm.nih.gov/pubmed/20195495, PMID: 20195495
44. Markowetz F, Spang R (2007) Inferring cellular networks – A review. BMC Bioinformatics 8 Suppl 6:S5, DOI 10.1186/1471-2105-8-S6-S5, URL http://www.ncbi.nlm.nih.gov/pubmed/17903286, PMID: 17903286
45. Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, Hurov KE, Luo J, Bakalarski CE, Zhao Z, Solimini N, Lerenthal Y, Shiloh Y, Gygi SP, Elledge SJ (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science (New York, NY) 316(5828):1160–1166, DOI 10.1126/science.1140321, URL http://www.ncbi.nlm.nih.gov/pubmed/17525332, PMID: 17525332
46. Miller ML, Jensen LJ, Diella F, Jorgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R (2008) Linear motif atlas for phosphorylation-dependent signaling. Sci Signal 1(35):ra2, DOI 10.1126/scisignal.1159433, URL http://www.ncbi.nlm.nih.gov/pubmed/18765831, PMID: 18765831
47. Morris MK, Saez-Rodriguez J, Sorger PK, Lauffenburger DA (2010) Logic-based models for the analysis of cell signaling networks. Biochemistry 49(15):3216–3224, DOI 10.1021/bi902202q, URL http://www.ncbi.nlm.nih.gov/pubmed/20225868, PMID: 20225868
48. Morris MK, Saez-Rodriguez J, Clarke DC, Sorger PK, Lauffenburger DA (2011) Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli. PLoS Comput Biol 7(3):e1001,099, DOI 10.1371/journal.pcbi.1001099, URL http://dx.doi.org/10.1371/journal.pcbi.1001099
49. Mukherjee S, Speed TP (2008) Network inference using informative priors. Proc Natl Acad Sci 105(38):14,313–14,318, DOI 10.1073/pnas.0802272105, URL http://www.pnas.org/content/105/38/14313.abstract
50. Naegle KM, Gymrek M, Joughin BA, Wagner JP, Welsch RE, Yaffe MB, Lauffenburger DA, White FM (2010) PTMScout, a web resource for analysis of high throughput post-translational proteomics studies. Mol Cell Proteom MCP 9(11):2558–2570, DOI 10.1074/mcp.M110.001206, URL http://www.ncbi.nlm.nih.gov/pubmed/20631208, PMID: 20631208
51. Nelander S, Wang W, Nilsson B, She Q, Pratilas C, Rosen N, Gennemark P, Sander C (2008) Models from experiments: combinatorial drug perturbations of cancer cells. Mol Syst Biol 4:216, DOI 10.1038/msb.2008.53, URL http://www.ncbi.nlm.nih.gov/pubmed/18766176, PMID: 18766176
52. Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteom MCP 4(10):1419–1440, DOI 10.1074/mcp.R500012-MCP200, URL http://www.ncbi.nlm.nih.gov/pubmed/16009968, PMID: 16009968

53. Olsen JV, Mann M (2011) Effective representation and storage of mass Spectrometry-Based proteomic data sets for the scientific community. Sci Signal 4(160):pe7, DOI 10.1126/scisignal.2001839, URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;4/160/pe7
54. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127(3):635–648, DOI 10.1016/j.cell.2006.09.026, URL http://www.ncbi.nlm.nih.gov/pubmed/17081983, PMID: 17081983
55. Ong S, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteom 1(5):376 –386, DOI 10.1074/mcp.M200025-MCP200, URL http://www.mcponline.org/content/1/5/376.abstract
56. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of s. cerevisiae by targeted proteomics. Cell 138(4):795–806, DOI 10.1016/j.cell.2009.05.051, URL http://www.ncbi.nlm.nih.gov/pubmed/19664813, PMID: 19664813
57. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PloS One 5(2):e9202, DOI 10.1371/journal.pone.0009202, URL http://www.ncbi.nlm.nih.gov/pubmed/20186320, PMID: 20186320
58. Rosario AMD, White FM (2010) Quantifying oncogenic phosphotyrosine signaling networks through systems biology. Curr Opin Genet Dev 20(1):23–30, DOI 10.1016/j.gde.2009.12.005, URL http://www.ncbi.nlm.nih.gov/pubmed/20074929, PMID: 20074929
59. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal Protein-Signaling networks derived from multiparameter Single-Cell data. Science 308(5721):523 –529, DOI 10.1126/science.1105809, URL http://www.sciencemag.org/content/308/5721/523.abstract
60. Saez-Rodriguez J, Goldsipe A, Muhlich J, Alexopoulos LG, Millard B, Lauffenburger DA, Sorger PK (2008) Flexible informatics for linking experimental data to mathematical models via DataRail. Bioinformatics 24(6):840 –847, DOI 10.1093/bioinformatics/btn018, URL http://bioinformatics.oxfordjournals.org/content/24/6/840.abstract
61. Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, Klamt S, Sorger PK (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Mol Syst Biol 5:331, DOI 10.1038/msb.2009.87, URL http://www.ncbi.nlm.nih.gov/pubmed/19953085, PMID: 19953085
62. Saez-Rodriguez J, Alexopoulos LG, Stolovitzky G (2011) Setting the standards for signal transduction research. Sci Signal 4(160):pe10, DOI 10.1126/scisignal.2001844, URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;4/160/pe10
63. Santos SDM, Verveer PJ, Bastiaens PIH (2007) Growth factor-induced MAPK network topology shapes erk response determining PC-12 cell fate. Nat Cell Biol 9(3):324–330, DOI 10.1038/ncb1543, URL http://www.ncbi.nlm.nih.gov/pubmed/17310240, PMID: 17310240
64. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra TD, Udseth HR (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics 2(5):513–523, DOI 10.1002/1615-9861(200205)2:5⟨513::AID-PROT513⟩3.0.CO;2-W, URL http://www.ncbi.nlm.nih.gov/pubmed/11987125, PMID: 11987125
65. Sneddon MW, Faeder JR, Emonet T (2011) Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. Nat Meth 8(2):177–183, DOI 10.1038/nmeth.1546, URL http://www.ncbi.nlm.nih.gov/pubmed/21186362, PMID: 21186362
66. Steen H, Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 5(9):699–711, DOI 10.1038/nrm1468, URL http://www.ncbi.nlm.nih.gov/pubmed/15340378, PMID: 15340378
67. Tan CSH, Jrgensen C, Linding R (2010) Roles of "junk phosphorylation" in modulating biomolecular association of phosphorylated proteins? Cell Cycle (Georgetown, Tex) 9(7):1276–1280, URL http://www.ncbi.nlm.nih.gov/pubmed/20234177, PMID: 20234177

68. Tanner SD, Ornatsky O, Bandura DR, Baranov VI (2007) Multiplex bio-assay with inductively coupled plasma mass spectrometry: towards a massively multivariate single-cell technology. Spectrochim Acta B 62(3):188–195, DOI 10.1016/j.sab.2007.01.008, URL http://www.sciencedirect.com/science/article/B6THN-4N0HJDH-1/2/05bbd4e8f7b003df4d258be40015b7ba
69. Taylor CF, Field D, Sansone S, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz P, Bogue M, Booth T, Brazma A, Brinkman RR, Clark AM, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Novere NL, Leebens-Mack J, Lewis SE, Lord P, Mallon A, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889–896, DOI 10.1038/nbt.1411, URL http://dx.doi.org/10.1038/nbt.1411
70. Thompson A, Schfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75(8):1895–1904, DOI 10.1021/ac0262560, URL http://dx.doi.org/10.1021/ac0262560
71. Vignali DA (2000) Multiplexed particle-based flow cytometric assays. J Immunol Meth 243(1–2):243–255, URL http://www.ncbi.nlm.nih.gov/pubmed/10986418, PMID: 10986418
72. Vizcano JA, Ct R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L (2009) A guide to the proteomics identifications database proteomics data repository. Proteomics 9(18):4276–4283, DOI 10.1002/pmic.200900402, URL http://www.ncbi.nlm.nih.gov/pubmed/19662629, PMID: 19662629
73. Vogel C, de Sousa Abreu R, Ko D, Le S, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol 6, DOI 10.1038/msb.2010.59, URL http://dx.doi.org/10.1038/msb.2010.59
74. Watterson S, Marshall S, Ghazal P (2008) Logic models of pathway biology. Drug Discov Today 13(9–10):447–456, DOI 10.1016/j.drudis.2008.03.019, URL http://www.ncbi.nlm.nih.gov/pubmed/18468563, PMID: 18468563
75. Wolf-Yadlin A, Sevecka M, MacBeath G (2009) Dissecting protein function and signaling using protein microarrays. Curr Opin Chem Biol 13(4):398–405, DOI 10.1016/j.cbpa.2009.06.027, URL http://www.ncbi.nlm.nih.gov/pubmed/19660979, PMID: 19660979
76. Wu F, Wang P, Zhang J, Young LC, Lai R, Li L (2010) Studies of phosphoproteomic changes induced by nucleophosmin-anaplastic lymphoma kinase (ALK) highlight deregulation of tumor necrosis factor (TNF)/Fas/TNF-related apoptosis-induced ligand signaling pathway in ALK-positive anaplastic large cell lymphoma. Mol Cell Proteom MCP 9(7):1616–1632, DOI 10.1074/mcp.M000153-MCP201, URL http://www.ncbi.nlm.nih.gov/pubmed/20393185, PMID: 20393185

Chapter 3
An Integrated Bayesian Framework for Identifying Phosphorylation Networks in Stimulated Cells

Tapesh Santra, Boris Kholodenko, and Walter Kolch

Abstract One of the primary mechanisms of signal transduction in cells is protein phosphorylation. Upon ligand stimulation a series of phosphorylation events take place which eventually lead to transcription. Different sets of phosphorylation events take place due to different stimulating ligands in different types of cells. Knowledge of these phosphorylation events is essential to understand the underlying signaling mechanisms. We have developed a Bayesian framework to infer phosphorylation networks from time series measurements of phosphosite concentrations upon ligand stimulation. To increase the prediction accuracy we integrated different types of data, e.g., amino acid sequence data, genomic context data (gene fusion, gene neighborhood, and phylogentic profiles), primary experimental evidence (physical protein interactions and gene coexpression), manually curated pathway databases, and automatic literature mining with time series data in our inference framework. We compared our results with data available from public databases and report a high level of prediction accuracy.

1 Introduction

There have been several attempts to reverse engineer the phosphorylation networks of cellular signaling pathways in recent years. These studies can be categorized in two main classes. Some of these studies used high throughput quantitative data such as mass spectrometry data, flow-cytometry data, RNAi screening data, etc., to establish causal relationship among kinase-substrate pairs [1–6] and some used non quantitative data such as protein sequence data and cellular context data to

T. Santra (✉) • B. Kholodenko • W. Kolch
Systems Biology Ireland, Conway Institute, University College Dublin (UCD),
Belfield, Dublin 4, Ireland
e-mail: tapesh.santra@ucd.ie; boris.kholodenko@ucd.ie; walter.kolch@ucd.ie

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_3, © Springer Science+Business Media, LLC 2012

determine probable phosphorylation patterns [7–10]. The studies which predict phosphorylation networks from quantitative data use a range of statistical methods such as least square regression, Bayesian network, dynamic Bayesian network, etc. Least square based methods such as [1] are suitable for models of small number of proteins, usually 20–50. On the other hand, Bayesian models such as [2, 3, 5, 6] require many repetitions of high throughput experiments under different experimental conditions. Usually high throughput experiments are repeated 2–5 times which may not be sufficient for a reliable inference when using the Bayes net methods. Hence, efficient implementations of these powerful tools require rigorous and often expensive biological experiments. Some other techniques such as probabilistic boolean networks [11] and clustering methods [12] have been used for network inference from quantitative data with different levels of success.

The network inference methods that use non quantitative data have some fundamental differences from the other methods. For example, these methods alone are not sufficient to determine whether a particular interaction takes place in a cell under certain experimental condition. Another limitation of these methods is that they do not differentiate between different phosphorylated states of the kinases themselves when predicting phosphorylation networks. Such information may be important for understanding the signaling mechanism under investigation. However, given the limitations these methods also reported reasonable prediction accuracy [7].

In this study we developed a Bayesian framework which integrates both quantitative and non quantitative data to infer phosphorylation network. We integrated protein sequence data and cellular context data with mass spectrometry data in a naive Bayes model to infer phosphorylation interactions that take place under certain experimental conditions. We used our method on the Olsen et al. [13] data in an effort to understand the phosphorylation interactions which occur when a Hela cell is stimulated by Epidermal Growth Factor (EGF). We analyzed the performance of our algorithm by benchmarking our results against curated data.

2 Algorithm

Let us assume that we have a set of kinase phosphopeptides and a set of substrate phosphopeptides which are denoted by $K = \{K_l : l = 1 \ldots \kappa\}$ and $S = \{S_i : i = 1 \ldots s\}$, respectively. Phosphopeptides are small phosphorylated fragments of proteins. Each phosphopeptide represents a phosphorylation state of the corresponding protein. In this study our objective is to identify which phosphorylation state of what kinase phosphorylates which substrates at what sites under certain experimental condition. From here on, we shall call K and S as kinases and substrates instead of phosphopeptides of kinases and substrates for convenience. Since, some kinases can phosphorylate each other $K \cap S \neq \emptyset$. We denote a phosphorylation event by $\rightarrow$, e.g., the phosphorylation of substrate S_i by kinase K_l is denoted by $K_l \rightarrow S_i$. Our objective is to find a

partition $\Sigma = \Sigma_l : l = 1 \ldots \kappa, \Sigma_j \subset S$ on the set of substrates S such that $K_l \rightarrow S_i \forall S_i \in \Sigma_l$. We used three different types of data to achieve the above objective, i.e., protein sequence data, cellular context data, and temporal measurement data which represents the concentrations of the phospho-peptides at different instants of time.

Sequence data is used in the form of motifs corresponding to the phosphorylation sites. Each motif consists of 15 amino acids, the phosphorylation site itself and seven amino acids to its left and right. The set of motifs is denoted by $M = \{M_i : i = 1 \ldots s\}$, where $M_i = \langle \alpha_{ij} : j = 1 \ldots 15 \rangle$. Here, α_{ij} represents the jth amino acid of the ith motif.

Contextual data is represented in the form of context network. In a context network two proteins are connected by an edge if a range of different types of data such as genomic context data (gene fusion, gene neighborhood, and phylogenetic profiles), primary experimental evidence (physical protein interactions and gene coexpression), manual and automatic literature, and database curation data indicate a probable interaction between them. It should be noted that an edge in a context based protein interaction network does not mean a physical interaction between two proteins. Instead, it suggests that the proteins are contextually close to each other. A probabilistic measure is associated with each edge which reflects confidence of interaction. This type of network data is available from STRING database [14] which we shall discuss in detail in the implementation section. The contextual proximity of a kinase K_l and a substrate S_i is denoted by ε_{li}.

The temporal measurements of kinase and substrate phospho-peptides are given by $K_l(t)$ and $S_i(t)$. We implemented a time lag correlation model in our method. For each kinase substrate pair K_l and S_i a cross correlation $C_{li}(\tau)$ between $K_l(t)$ and $S_i(t)$ is calculated using the following formula:

$$C_{li}(\tau) = \frac{\sum_{t=0}^{T} \hat{K}_l(t-\tau)\hat{S}_i(t)}{\sqrt{\sum_{t=0}^{T} \hat{K}_l(t-\tau)\hat{K}_l(t-\tau) \sum_{t=0}^{T} \hat{S}_i(t)\hat{S}_i(t)}}, \tag{3.1}$$

τ is the time lag at which the cross correlation $C_{li}(\tau)$ is calculated, $\hat{K}_l(t) = \frac{K_l(t)-\bar{K}_l(t)}{\sigma_{K_l}}$ and $\hat{S}_i(t) = \frac{S_i(t)-\bar{S}_i(t)}{\sigma_{S_i}}$. Here, $\bar{K}_l(t)$ and $\bar{S}_i(t)$ are sample means and σ_{K_l} and σ_{S_i} are the standard deviation of $K_l(t)$ and $S_i(t)$, respectively. The time lag at which maximum correlation occurs between $K_l(t)$ and $S_i(t)$ is denoted by $\tau_{li}^{\max}$, i.e., $\tau_{li}^{\max} = \text{argmax}(C_{li}(\tau))$. The maximum value of $C_{li}(\tau)$ is denoted by $C_{li}^{\max} = C_{li}(\tau_{li}^{\max})$. We transform $C_{li}(\tau)$ using Fisher transform [15] as shown below:

$$Z_{li}(\tau) = \left(\frac{1}{2}\right) \frac{1 + \ln(C_{li}(\tau))}{1 - \ln(C_{li}(\tau))}. \tag{3.2}$$

Hence, $Z_{li}^{\max} = Z_{li}(\tau_{li}^{\max})$. We used $Z_{li}^{\max}$ values instead of $C_{li}^{\max}$ in our model because under certain assumptions $Z_{li}(\tau)$ is shown to be normally distributed [15]. This particular feature of Fisher transform is convenient for further calculations.

The properties of Fisher transform and the assumptions under which the transformed values are normally distributed will be discussed in the Subsect. 2.3. Based on the above notations the posterior probability of $K_l \rightarrow S_i$ given motif M_i, contextual proximity ε_{li} and time lag correlation data $\tau_{li}^{\max}$ and $Z_{li}^{\max}$ can be given as follows:

$$P\left(K_l \rightarrow S_i | M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max}\right) = \frac{P(M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max} | K_l \rightarrow S_i) P(K_l \rightarrow S_i)}{\sum_{K_l \in K, S_i \in S} P(M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max} | K_l \rightarrow S_i) P(K_l \rightarrow S_i)}. \tag{3.3}$$

If the prior $P(K_l \rightarrow S_i)$ is equal for all $K_l \in K$ and $S_i \in S$ then (3.3) can be rewritten as below.

$$P\left(K_l \rightarrow S_i | M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max}\right) \propto P(M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max} | K_l \rightarrow S_i). \tag{3.4}$$

The right hand side of (3.4) can be easily calculated under certain assumptions. We assume that the phosphorylation motifs, the contextual proximity of the proteins and the temporal measurements of phospho-peptide concentrations are independent of each other. Due to the strong independence assumption our model can be classified as a Naive Bayes model. The strong independence assumption may be over simplification of biological reality. However, earlier studies suggested that Naive Bayes models are quite robust against such oversimplifying and often erroneous assumptions and oftentimes outperform more sophisticated models [16]. Under the above independence assumption (3.4) can be rewritten as follows:

$$P(K_l \rightarrow S_i | M_i, \varepsilon_{li}, \tau_{li}^{\max}, Z_{li}^{\max}) \propto P(M_i | K_l \rightarrow S_i) \times P(\varepsilon_{li} | K_l \rightarrow S_i) \times P(Z_{li}^{\max}, \tau_{li}^{\max} | K_l \rightarrow S_i). \tag{3.5}$$

The calculations of the three probability measures on the right hand side of (3.5) are described in details in the following subsections.

2.1 Calculating $P(M_i | K_l \rightarrow S_i)$

As stated before, a motif consists of 15 amino acids, $M_i = \langle \alpha_{ij} : j = 1 \ldots 15 \rangle$, which represent a small protein fragment surrounding a phosphorylation site. There are 20 possible amino acids. Let us denote the set of amino acids by $\mathscr{A} = \{A_k : k = 1 \ldots 20\}$. We assume that each α_{ij} is an independent random variable the values of which are drawn from the set of amino acids $\mathscr{A}$ with multinomial distributions, i.e., $\alpha_{ij} \sim$ Multinomial$(\overrightarrow{\pi_j})$. Here, $\overrightarrow{\pi_j} = \{\pi_{jk} : k = 1 \ldots 20\}$ is the set of

multinomial parameters. Let us consider that the multinomial parameters π_{jk} have Dirichlet priors, i.e., $\pi_{jk} \sim D(\lambda_k : k = 0 \ldots 20)$, where $\lambda_k : k = 1 \ldots 20$ are hyper parameters and $\lambda_0 = \sum_{k=1}^{20} \lambda_k$. Let us denote the multinomial parameters of the amino acids found in a set of motifs $\Delta_l = \{\acute{M}_i : i = 1 \ldots N_{\acute{M}_i}\}$ which are targets of kinase K_l by π_{jl}. The posterior of $\overrightarrow{\pi_{jl}}$ can be given by:

$$P\left(\overrightarrow{\pi_{jl}}|K_l \rightarrow S_i\right) = P\left(\overrightarrow{\pi_{jl}}|\Delta_l\right) = D(\lambda_1 + N_{j1l}, \lambda_2 + N_{j2l} \ldots \lambda_{20} + N_{j20l}, \lambda_0 + N_l). \tag{3.6}$$

In (3.6), N_{jkl} is the number of times the kth amino acid is observed at the jth position in the motifs of Δ_l. Given the above posterior, the probability $P(M_i|K_l \rightarrow S_i)$ can be calculated as follows:

$$\begin{aligned} P(M_i|K_l \rightarrow S_i) &= P(\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{i15}|K_l \rightarrow S_i) \\ &= \prod_{j=1}^{15} P(\alpha_{ij}|K_l \rightarrow S_i) \\ &= \prod_{j=1}^{15} \int P\left(\alpha_{ij}|\overrightarrow{\pi_{jl}}\right) P\left(\overrightarrow{\pi_{jl}}|K_l \rightarrow S_i\right) d\overrightarrow{\pi_{jl}} \\ &= \prod_{j=1}^{15} E(\pi_{jkl}|K_l \rightarrow S_i) \\ &= \prod_{j=1}^{15} \frac{\lambda_k + N_{jkl}}{\lambda_0 + N_l}. \end{aligned} \tag{3.7}$$

Initially Δ_ls are the training sets for motif data which can be found in publicly available databases such as HPRD [17], Phosida [18], Phosphosite [19], PhosphoELM [20], etc. We shall discuss about these databases in the implementation section. As the algorithm progresses the newly classified motifs are also included in Δ_l. The above model can be summarized as follows:

$$\begin{aligned} M_i &= \{\alpha_{ij} : j = 1 \ldots 15\} \\ \alpha_{ij} &\sim \text{Mult}(\overrightarrow{\pi_{jl}}) \\ \overrightarrow{\pi_j} &= \{\pi_{jkl} : k = 1 \ldots 20\} \\ \overrightarrow{\pi_{jl}} &\sim D(\lambda_k : k = 0 \ldots 20) \end{aligned} \tag{3.8}$$

The values of λ_k are fixed at $\lambda_k = 1; \forall k = 1 \ldots 20$.

Some of the assumptions of the above model might be over simplistic. For example, the assumption that α_{ij}'s are independent is an oversimplification. More sophisticated models, such as Hidden Markov Models [21–23] take the interdependencies among consecutive amino acids into account when modeling consensus phosphorylation motifs. However, we find that our simplistic approach performs at least as well as the more sophisticated models especially when applied in conjunction with other features.

2.2 Calculating $P(\varepsilon_{li}|K_l \rightarrow S_i)$

ε_{li} represents the proximity between K_l and S_i in a context based protein interaction network. A context based protein interaction network is a probabilistic network where nodes represent proteins and edges represent protein–protein interactions (here an interaction does not necessarily mean physical interaction). Each edge has a probability score which represents the confidence of interaction between corresponding proteins. The probability scores are calculated from a range of data such as gene neighborhood, gene fusion, co-occurrence, homology, co-expression, experimental evidence, knowledge base, and text mining [14]. Let us denote a context based protein interaction network by $\mathscr{G} = \langle V, E,\ P(E)\rangle$ where V is the set of vertices, E is the set of undirected edges, $P(E)$ is a probability measure on the set of edges, $P : E \rightarrow \{0, 1\}$. We opt to use this probability measure to determine the contextual proximity between a kinase and a substrate. Such information have been used before by Linding et al. [7] in order to infer kinase substrate specificity. In cases where a phosphorylation site contains consensus motifs which can be recognized by multiple kinases Linding et al. [7] chose those kinases which are directly connected to the substrate in the context network. In most cases this approach may suffice to identify kinase substrate specificity efficiently. However, in general, the method of preferring directly connected kinase-substrate pairs over indirectly connected ones have some conceptual ambiguity. The ambiguity of the above method is demonstrated in the following example.

Consider a probabilistic context network which consists of two kinases $\acute{K}_1$, $\acute{K}_2$, a substrate $\acute{S}_1$ and a protein $\acute{p}$ which connects $\acute{K}_2$ with $\acute{S}_1$ as shown in Fig. 3.1. The edges and their probability scores are defined as $\acute{e}_1 = (\acute{K}_1, \acute{S}_1), \acute{e}_2 = (\acute{K}_2, \acute{p}), \acute{e}_3 = (\acute{p}, \acute{S}_1)$, $P(\acute{e}_1) = 0.75, P(\acute{e}_2) = 0.9, P(\acute{e}_2) = 0.9$. In this network, $\acute{K}_1$ is directly connected to $\acute{S}_1$ and $\acute{K}_2$ is indirectly connected to $\acute{S}_1$. But, the probability of interaction between $\acute{K}_2$ and $\acute{S}_1$, $P(\varepsilon_{21}) = P(\acute{e}_2) \times P(\acute{e}_3) = 0.81$ is higher than the probability of interaction between $\acute{K}_1$ and $\acute{S}_1$, $P(\varepsilon_{11}) = 0.75$. Hence, in this particular case, choosing $\acute{K}_1$ over $\acute{K}_2$ as a potential kinase of $\acute{S}_2$ may not be correct.

Due to the above difficulty we define ε_{li} using a concept called "two-point reliability" which is well studied in communication theory [24]. Reliability of communication between two nodes in a probabilistic network is defined as the probability of existence of a connecting path between the nodes [24]. Following this notion, in a probabilistic context network a measure of proximity between

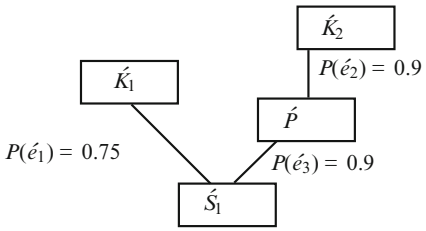


Fig. 3.1 Toy context network consisting of two kinses $\acute{K}_1$, $\acute{K}_2$, one substrate $\acute{S}_1$ and a protein $\acute{P}$. $\acute{K}_1$ interacts with $\acute{S}_1$ with probability $P(\acute{e}_1) = 0.75$, $\acute{K}_2$ interacts with $\acute{P}$ with probability $P(\acute{e}_2) = 0.9$ and $\acute{P}$ interacts with $\acute{S}_1$ with probability $P(\acute{e}_3) = 0.9$. Here, $P(\acute{e}_2) \times P(\acute{e}_3) > P(\acute{e}_1)$

two proteins can be defined as the probability of existence of a connecting path between them. For example, the proximity ε_{li} between the kinase K_l and substrate S_i can be measured as the probability that a path exists between K_l and S_i in the context network $\mathscr{G}$. Calculating the proximity measure is a NP-hard problem and there are several potential solutions proposed over the years [24–26]. We shall use some of the concept developed in these earlier studies. Given a context network $\mathscr{G} = \langle V, E, P(E)\rangle$ and two of its vertices $v_i, v_j \in V$, let us denote the set of paths that connects v_i, v_j by $\mathscr{P}_{ij} = \{\mathscr{P}_{ij}^k : k = 1 \ldots N_p\}$ where $N_p = |\mathscr{P}_{ij}|$. Using inclusion–exclusion theory the probability that a path exists between v_i and v_j is given in the following equation [24].

$$P\left(\cup_{k=1}^{N_p}\mathscr{P}_{ij}^k\right) = \sum_{k=1}^{N_p} P\left(\mathscr{P}_{ij}^k\right) - \sum_{k\neq l} P\left(\mathscr{P}_{ij}^k \mathscr{P}_{ij}^l\right) + \cdots + (-1)^{N_p-1} P\left(\mathscr{P}_{ij}^1 \mathscr{P}_{ij}^2 \ldots \mathscr{P}_{ij}^{N_p}\right). \tag{3.9}$$

In (3.9), $P\left(\cup_{k=1}^{N_p}\mathscr{P}_{ij}^k\right)$ is the probability that a path exists between v_i and v_j in the context network $\mathscr{G}$, i.e., $P(\varepsilon_{ij}) = P\left(\cup_{k=1}^{N_p}\mathscr{P}_{ij}^k\right)$. The terms of the right hand side of (3.9) can be calculated as follows:

$$\begin{aligned}
P\left(\mathscr{P}_{ij}^k\right) &= \prod_{e_k \in \mathscr{P}_{ij}^k} P(e_k) \\
P\left(\mathscr{P}_{ij}^k \mathscr{P}_{ij}^l\right) &= \prod_{e_k \in \mathscr{P}_{ij}^k} P(e_k) \prod_{e_l \in \mathscr{P}_{ij}^l} P(e_l) \\
&\cdots = \cdots \\
P\left(\mathscr{P}_{ij}^1 \mathscr{P}_{ij}^2 \cdots \mathscr{P}_{ij}^{N_p}\right) &= \prod_{e_1 \in \mathscr{P}_{ij}^1} P(e_1) \prod_{e_2 \in \mathscr{P}_{ij}^2} P(e_2) \cdots \prod_{e_{N_p} \in \mathscr{P}_{ij}^{N_p}} P(e_{N_p})
\end{aligned} \tag{3.10}$$

In (3.10), e_k is an edge, i.e., $e_k \in E$ and $P(e_k)$ is its probability score. However, the proximity measure shown in (3.9) can also be calculated using a simple recursive formula. Given a set of paths $\mathscr{P}_{ij} = \{\mathscr{P}_{ij}^k : k = 1 \ldots N_p\}$, between v_i and v_j the probability measure shown in (3.9) can be calculated in the following manner. The probability that a path exists between v_i and v_j is the probability of traversing at least one path in $\mathscr{P}_{ij}$. Hence,

$$\begin{aligned} P(\varepsilon_{ij}) =& P\left(\mathscr{P}_{ij}^1\right) + \left(1 - P\left(\mathscr{P}_{ij}^1\right)\right) P\left(\mathscr{P}_{ij}^2\right) \\ &+ \left(1 - P\left(\mathscr{P}_{ij}^1\right) + \left(1 - P\left(\mathscr{P}_{ij}^1\right)\right) P\left(\mathscr{P}_{ij}^2\right)\right) P(\mathscr{P}_{ij}^3) + \cdots \end{aligned} \tag{3.11}$$

The right hand side of (3.11) can be interpreted as follows: $P(\mathscr{P}_{ij}^1)$ is the probability of traversing $\mathscr{P}_{ij}^1$, $(1 - P(\mathscr{P}_{ij}^1))P(\mathscr{P}_{ij}^2)$ is the probability of traversing $P(\mathscr{P}_{ij}^2)$ and not $P(\mathscr{P}_{ij}^1)$, $(1 - P(\mathscr{P}_{ij}^1) + (1 - P(\mathscr{P}_{ij}^1))P(\mathscr{P}_{ij}^2))P(\mathscr{P}_{ij}^3)$ is the probability of traversing $P(\mathscr{P}_{ij}^3)$ but not $P(\mathscr{P}_{ij}^1)$ and $P(\mathscr{P}_{ij}^2)$, and so on. A recursive representation of (3.11) is given below.

$$\acute{P}(\varepsilon_{ij})^{k+1} = \acute{P}(\varepsilon_{ij})^k + (1 - \acute{P}(\varepsilon_{ij})^k) * P(\mathscr{P}_{ij}^{k+1}) \text{ where } \acute{P}(\varepsilon_{ij}^0) = 0. \tag{3.12}$$

$P(\varepsilon_{ij})$ is calculated from (3.12) by replacing $k + 1$ by N_p, i.e., $P(\varepsilon_{ij}) = \acute{P}(\varepsilon_{ij})^{N_p}$.

The proximity measure $P(\varepsilon_{li})$ between a kinase K_l and substrate S_i can be calculated using (3.12) given a context network which contains both. In this case, the context network is an undirected graph whose edge probabilities are calculated from contextual data and are independent of the assumption of direct post translational modification events, i.e., $P(\varepsilon_{li}|K_l \rightarrow S_i) = P(\varepsilon_{li})$. Hence, $P(\varepsilon_{li}|K_l \rightarrow S_i)$ can directly be calculated from databases such as STRING [14].

2.3 Calculating $P(Z_{li}^{\max}, \tau_{li}^{\max}|K_l \rightarrow S_i)$

As defined before, $Z_{li}(\tau)$ is the Fisher transformation of the correlation coefficient $C_{li}(\tau)$. Fisher [15] showed that Z_{li}^{τ} is normally distributed if $K_l(t)$ and $S_i(t)$ are normal independent variables. Using extreme value theory, the cumulative distribution function of the extremum (in this case maximum) of a set of independent normal variables can be given as follows [27]:

$$P(Z \leq z) = \exp\left(-\exp-\left(\frac{z-\mu}{\sigma}\right)\right). \tag{3.13}$$

In (3.13), μ is the location parameter and σ is the scale parameter. $Z_{li}(\tau)$ are normally distributed but not independent. However, Leadbetter et al. [28] (Chapters 4–6) showed that (3.13) holds fairly generally even under many dependency conditions. Hence, the probability density function of $Z_{li}^{\max}$ can be given as follows:

$$P\left(Z_{li}(\tau_{li}^{\max})|\mu(\tau_{li}^{\max}),\sigma(\tau_{li}^{\max})\right) = \frac{1}{\sigma(\tau_{li}^{\max})}\exp\left(-\frac{z-\mu\left(\tau_{li}^{\max}\right)}{\sigma\left(\tau_{li}^{\max}\right)}\right) \times \exp\left(-\exp\left(-\frac{z-\mu\left(\tau_{li}^{\max}\right)}{\sigma\left(\tau_{li}^{\max}\right)}\right)\right). \quad (3.14)$$

In (3.14), $\mu(\tau_{li}^{\max})$ and $\sigma(\tau_{li}^{\max})$ are location and scale parameters, and are dependent on $\tau_{li}^{\max}$. The dependence of the distribution parameters on $\tau_{li}^{\max}$ arises from the biological mechanism of phosphorylation. If a substrate is phosphorylated by a kinase then the maximum phosphorylation of the substrate occurs almost instantaneously (within ≈5 min) after the maximum phosphorylation of the kinase (e.g., see [29]). Hence, it can be assumed that the maximum correlation between $K_l(t)$ and $S_i(t)$ occurs at a time lag no greater than ≈5 min if $K_l \rightarrow S_i$ holds true, i.e., $\tau_{li}^{\max} \leq 5$ if $K_l \rightarrow S_i$. Hence, given $\tau_{li}^{\max} \leq 5$, the location parameter corresponds to high values of $C_{li}(\tau)$. However we have no prior information about the distribution of $Z_{li}^{\max}$ and parameter $\mu(\tau_{li}^{\max})$ when $\tau_{li}^{\max} > 5$. In this case, we assume that the location parameter has a flat distribution. To reflect the above assumptions we define the conditional density of $\mu(\tau_{li}^{\max})$ as follows:

$$\begin{aligned} P(\mu(\tau_{li}^{\max})|\tau_{li}^{\max}, K_l \rightarrow S_i) &= N(m_\mu, \sigma_\mu) \\ \text{where } m_\mu &= 1.5, \sigma_\mu = 1 \text{ if } \tau_{li}^{\max} \leq 5 \\ m_\mu &= 0, \sigma_\mu = 3 \text{ if } \tau_{li}^{\max} > 5 \end{aligned} \quad (3.15)$$

In (3.15), the $m_\mu = 1.5|\tau_{li}^{\max} \leq 5$ since it corresponds to $C_{li}^{\max} \approx 0.9$. $m_\mu = 0$, $\sigma_\mu = 3|\tau_{li}^{\max}$ represents a flat distribution over all possible values. We fix the scale parameter at $\sigma(\tau_{li}^{\max}) = 1$. Under the above model the conditional density function of $Z_{li}^{\max}$ can be calculated and given $\tau_{li}^{\max}$ as shown in the following equation:

$$\begin{aligned} &P(Z_{li}^{\max}|\tau_{li}^{\max}, K_l \rightarrow S_i) \\ &= \int P\left(Z_{li}^{\max}|\mu(\tau_{li}^{\max})\right) P(\mu(\tau_{li}^{\max})|\tau_{li}^{\max}, K_l \rightarrow S_i) d(\mu(\tau_{li}^{\max})) \\ &= \int \left(\frac{1}{\sqrt{(2\pi\sigma_{mu}^2)}}\right) \frac{1}{\sigma(\tau_{li}^{\max})}\exp\left(-\frac{z-\mu\left(\tau_{li}^{\max}\right)}{\sigma\left(\tau_{li}^{\max}\right)}\right) \\ &\quad \times \exp\left(-\exp\left(-\frac{z-\mu\left(\tau_{li}^{\max}\right)}{\sigma\left(\tau_{li}^{\max}\right)}\right)\right)\exp-\left(\frac{\left(\mu(\tau_{li}^{\max})-m_\mu\right)^2}{2\sigma_\mu^2}\right)\mathrm{d}\mu(\tau_{li}^{\max}). \end{aligned} \quad (3.16)$$

We have demonstrated the conditional density function $P(Z_{li}^{\max}|\tau_{li}^{\max}, K_l \rightarrow S_i)$ in Fig. 3.2.

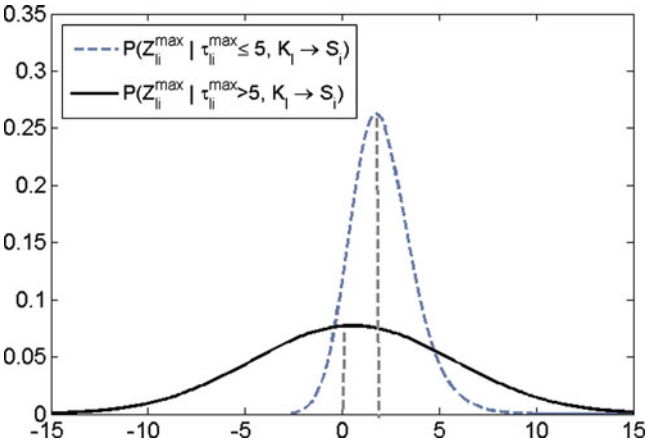


Fig. 3.2 The conditional probability density $P(Z_{li}^{\max}|\tau_{li}^{\max}, K_l \rightarrow S_i)$ is shown for $\tau_{li}^{\max} \leq 5$ and $\tau_{li}^{\max} > 5$ in this figure. $P(Z_{li}^{\max}|\tau_{li}^{\max} \leq 5, K_l \rightarrow S_i)$ has a sharp peak at $Z_{li}^{\max} = 1.5$, whereas $P(Z_{li}^{\max}|\tau_{li}^{\max} > 5, K_l \rightarrow S_i)$ has a flat distribution over a large range of values

$\tau_{li}^{\max}$ can only have discrete values since in biological experiments time series data is measured at discrete time interval. Hence, the conditional density of $\tau_{li}^{\max}$ can be given as follows:

$$\begin{aligned} \tau_{li}^{\max}|K_l \rightarrow S_i &\sim Mult(\overrightarrow{\phi_l}) \\ \overrightarrow{\phi_l} &= \{\phi_{tl} : t = 0 \ldots T\} \\ \overrightarrow{\phi_l} &\sim D(\nu_0, \nu_1, \ldots \nu_T, \acute{\nu}) \end{aligned} \tag{3.17}$$

In (3.17), ϕ_t are multinomial parameters, and $\nu_t : t = 0, \ldots, T$ are Dirichlet parameters where $\acute{\nu} = \sum_{t=0}^{T} \nu_t$. We fix the values of $\nu_t = 1 \forall t = 1 \ldots T$. Though it is known that given $K_l \rightarrow S_i$, $\tau_{li}^{\max}$ is usually ≤ 5 min the little information is available a priori about the parameter values of the multinomial. Hence we adopt an online update technique to calculate these parameter values. Given a set of substrates $\acute{S} = \{\acute{S}_i : K_l \rightarrow \acute{S}_i\}$, the posterior of $\overrightarrow{\phi_l}$ can be given by $P(\overrightarrow{\phi_l}|\acute{S}) = D(\nu_t + N_{\tau_t}^l : t = 0 \ldots T, \acute{\nu} + N_{\tau}^l)$. Here $N_{\tau_t}^l$ is the number of substrates whose temporal concentration correlates maximally with that of kinase K_l at $\tau = t$, and N_{τ}^l is the total number of substrates which are phosphorylated by K_l. The posterior of $\tau_{li}^{\max}$ can now easily be calculated using the following formula.

$$\begin{aligned} P\left(\tau_{li}^{\max} = t|\acute{S}\right) &= P\left(\tau_{li}^{\max} = t|K_l \rightarrow S_i\right) \\ &= \int P(\tau_{li}^{\max} = t|\phi_{tl}) P(\phi_{tl}|K_l \rightarrow S_i) d\phi_{tl} \\ &= \int \phi_{tl} P(\phi_{tl}|K_l \rightarrow S_i) d\phi_{tl} = \frac{\nu_t + N_{\tau_t}^l}{\acute{\nu} + N_{\tau}^l} \end{aligned} \tag{3.18}$$

Since, in this case, we do not have a reference database, we start with $N^l_{\tau_t} = 0$ and $N^l_{\tau} = 0$, and update the parameters using (3.18) as we sample from the overall distribution. Finally, given (3.18) and (3.16) the conditional density $P(Z^{\max}_{li}, \tau^{\max}_{li} | K_l \rightarrow S_i)$ can be calculated using the following equation:

$$P(Z^{\max}_{li}, \tau^{\max}_{li} | K_l \rightarrow S_i) = P(Z^{\max}_{li} | \tau^{\max}_{li}, K_l \rightarrow S_i) P(\tau^{\max}_{li} | K_l \rightarrow S_i). \quad (3.19)$$

2.4 The Log Likelihood Function

Given the posterior of $K_l \rightarrow S_i$ the log likelihood can be defined by the following equation:

$$\begin{aligned} \mathscr{L} &= \log \left(\prod_{K_l \in K, S_i \in S} P(K_l \rightarrow S_i | M_i, \varepsilon_{li}, Z^{\max}_{li}, \tau^{\max}_{li}) \right) \\ &= \sum_{K_l \in K, S_i \in S} \log \left(P(K_l \rightarrow S_i | M_i, \varepsilon_{li}, Z^{\max}_{li}, \tau^{\max}_{li}) \right) \end{aligned} \quad (3.20)$$

Finally an optimization algorithm can be used to maximize the log likelihood shown in (3.20). For our implementation we used a MCMC based sampling scheme to find optimal assignment of K_l, S_i pairs.

2.5 Pseudo Code for the Above Algorithm

The mathematical equations described above can be put together in a clustering algorithm to identify kinase substrate interactions from quantitative data. The pseudocode of the algorithm we have implemented is shown below.

$\lambda_k \leftarrow 1; \forall k = 0 \ldots 20$
$\nu_t \leftarrow 1; \forall t = 1 \ldots T$
$\sigma \left(\tau^{\max}_{ij} \right) \leftarrow 1$
Calculate N_{jkl} from motif databases
$N_{\tau_t} \leftarrow 0$ for $t = 1 \ldots T$
Calculate $P(\varepsilon_{li})$ from STRING database using (3.12)
$\pi_{jkl} \leftarrow \frac{\lambda_k + N_{jkl}}{\lambda_0 + N_l}$
$\mathscr{L} \leftarrow 0$
$\mathscr{L}_1 \leftarrow 0$

$th \leftarrow 10^{-6}$
$P[l] \leftarrow 0; \forall l = 1 \ldots \kappa$
$Labels[i] \leftarrow -1; \forall 1 \ldots s$ {*Stores the previous cluster labels*}
$Label \leftarrow -1$
$\tau^{\max}[i] \leftarrow -1; \forall i = 1 \ldots s$ {*Stores the previous values of* $\tau_{li}^{\max}$}
while $d(\mathscr{L}) > th$ **do**
  $\mathscr{L}_1 \leftarrow \mathscr{L}$
  $\mathscr{L} \leftarrow 0$
  for $i = 1 \rightarrow s$ **do**
    for $l = 1 \rightarrow \kappa$ **do**
      $P[l] \leftarrow P(M_i|K_l \rightarrow S_i) \times P(\varepsilon_{li}) \times P(Z_{li}^{\max}, \tau_{li}^{\max}|K_l \rightarrow S_i)$
    end for
    $Label \leftarrow Sample(P)$
    $\mathscr{L} \leftarrow \mathscr{L} + \log(P(Label))$
    if $Labels[i] == -1$ **then**
      $N_{jkl} \leftarrow N_{jkl} + 1$ for $l = Label$
      $N_l \leftarrow N_l + 1$ for $l = Label$
      $\pi_{jkl} \leftarrow \frac{\lambda_k + N_{jkl}}{\lambda_0 + N_l}$ for $l = Label$
      $N_{\tau_t}^l \leftarrow N_{\tau_t}^l + 1$ at $t = \tau_{li}^{\max}, l = Label$
      $\phi_{tl} \leftarrow \frac{\nu_t + N_{\tau_t}^l}{\acute{\nu} + N_\tau^l}$ at $t = \tau_{li}^{\max}, l = Label$
      $Labels[i] = Label$
      $\tau^{\max}[i] \leftarrow \tau_{li}^{\max}$ for $l = Label$
    else
      if $Labels[i] \neq Label$ **then**
        {*Update the parameters of the new cluster*}
        $N_{jkl} \leftarrow N_{jkl} + 1$ for $l = Label$
        $N_l \leftarrow N_l + 1$ for $l = Label$
        $\pi_{jkl} \leftarrow \frac{\lambda_k + N_{jkl}}{\lambda_0 + N_l}$ for $l = Label$
        $N_{\tau_t}^l \leftarrow N_{\tau_t}^l + 1$ at $t = \tau_{li}^{\max}, l = Label$
        $N_\tau^l \leftarrow N_\tau^l + 1$ for $l = Label$
        $\phi_{tl} \leftarrow \frac{\nu_t + N_{\tau_t}^l}{\acute{\nu} + N_\tau^l}$ at $t = \tau_{li}^{\max}, l = Label$
        {*Update the parameters of the old cluster*}
        $N_{jkl} \leftarrow N_{jkl} - 1$ for $l = Labels[i]$
        $N_l \leftarrow N_l - 1$ for $l = Labels[i]$
        $\pi_{jkl} \leftarrow \frac{\lambda_k + N_{jkl}}{\lambda_0 + N_l}$ for $l = Labels[i]$
        $N_{\tau_t}^l \leftarrow N_{\tau_t}^l - 1$ at $t = \tau_{li}^{\max}, l = Labels[i]$
        $N_\tau^l \leftarrow N_\tau^l - 1$ for $l = Labels[i]$
        $\phi_{tl} \leftarrow \frac{\nu_t + N_{\tau_t}^l}{\acute{\nu} + N_\tau^l}, t = \tau_{li}^{\max}, l = Labels[i]$
        $\tau^{\max}[i] \leftarrow \tau_{li}^{\max}$, for $l = Label$
        $Labels[i] \leftarrow Label$

```
        end if
      end if
    end for
    d(ℒ) ← |ℒ − ℒ1|
end while
```

3 Implementation of Our Algorithm to Analyze Phosphoproteomic Data

We implemented our algorithm on phosphoproteomic data to infer phosphorylation networks. We used time resolved mass spectrometry data from [13], phosphorylation site and sequence motif data from HPRD [17], Phosida [18], Phosphosite [19], PhosphoELM [20] databases, and probabilistic interaction network data from STRING database [14]. The main problem in using multitude of data from different databases is ID conflict among them. For example, [13] uses IPI numbers as protein IDs, STRING database uses protein names and Swissprot identifiers as protein IDs, PHOSIDA uses gene names, and IPI numbers as protein IDs, etc. Although there are some ID mapping services available (e.g., IPI database mapping tools [30]) these softwares usually do not produce a one to one mapping for any two ID types. Hence, we used five different types of data to uniquely identify the proteins among different databases. We used IPI numbers, Swissprot IDs, gene names, protein aliases, and protein sequences to establish an unique identity for each protein among different databases. For any two types of IDs we first established a map by two way validation which was then refined by using protein sequence similarity. This was carried out for every pair of IDs that we came across in these databases. Once protein identities were established and all necessary data were collected, we implemented our algorithm to establish pairwise kinase substrate specifications from the data. The time resolved mass spectrometry data [13] consists of relative concentrations of phosphopeptides measured over a period of 20 min, after EGF stimulation of Hela cells. The phosphopeptides are fragments of both kinases and substrate proteins. For the set of kinases we used only those for which mass spectrometry data is available in [13] dataset. Based on motif and STRING data we filtered out all the non-kinase phosphopeptides which are likely to be phosphorylated by kinases not observed in [13]'s experiment. This is done because our algorithm can not deal with hidden variable (unobserved kinases) at its current stage of development. We implemented both the MCMC method as shown in Sect. 2.5 and simulated annealing method on the filtered datasets for sampling plausible networks. For simulated annealing we sampled from $\left(P(M_i||K_l \to S_i) \times P(\varepsilon_{li}||K_l \to S_i) \times P(\tau_{li}^{\max}, Z_{li}^{\max}|K_l \to S_i)\right)^{1/T}$ where T is the temperature parameter. The temperature is reduced after every 200 iterations

using a geometric annealing function $T = T \times 0.95$. We find that the best results can be obtained by collecting a large number of samples (we took 50 samples for our study) after convergence of the MCMC algorithm and taking those kinase substrate pairs which occur in at least, e.g., 10 of the samples.

3.1 Results

In the dataset of [13] the phosphopeptide concentrations were measured after stimulating Hela cells with EGF. Hence, application of our algorithm on this data enables us to detect the phosphorylation events that take place immediately after EGF stimulation to Hela cells. The best way to verify our results is to validate each phosphorylation event individually in vivo. But such experiment is out of the context of this study. There are, however, limited data on in-vivo kinase substrate specification, e.g., Phosphosite [19]. But these datasets cover a small fraction of human proteome and there are very little overlap between them and [13]'s dataset. The database that covers the largest set of phosphorylation interactions is NetworKIN [7]. Only $\approx$62% of filtered phosphorylation sites from [13]'s dataset is documented in [7]. Approximately 72% of the phosphosites in the filtered datasets are documented in three databases put together, i.e., NetworKIN [7], RegPhos [31], and Phosphosite [19]. Some of these databases provide online prediction of putative kinases for new phosphorylation sites, e.g., NetworKIN [7] and NetPhosK [32]. The phosphosites which are not included in the above databases are fed into the online prediction services and results are collected. Among these phosphosites we selected only those for which at least one putative kinase is predicted by the online prediction services. This covers a total of $\approx$88% of the filtered phosphosites.

Due to unavailability of gold standard data sets and low overlap between [13]'s data and any individual database, we benchmarked our results with those of the databases and online prediction services mentioned above. For benchmarking we chose five well studied kinases, e.g. EGFR, MAPK3, MAPK14, GSK3β, and Rock2. The result of benchmarking is shown in Fig. 3.3.

The results shown in Fig. 3.3 suggest that there is a high level of similarity among the predictions made by our algorithm and other publicly available prediction services and databases. This only shows that the performance of our algorithm is comparable to those of the others. However, how accurate is our algorithm in detecting in-vivo phosphorylation events can not be determined this way. This is because there are no experimental proof for a large number of predictions made by our algorithm or the other publicly available prediction services.

We also tried to construct a signaling pathway from our predicted interactions. The main difficulty in constructing a pathway using our algorithm on quantitative data is that most data sets do not contain all phosphosites of a particular pathway. Additionally, a pathway is made of multiple different types of interactions such as phosphorylation, complex formation, ubiquitination, etc. Our algorithm can infer only phosphorylation interactions and this leads to inconsistency in different parts

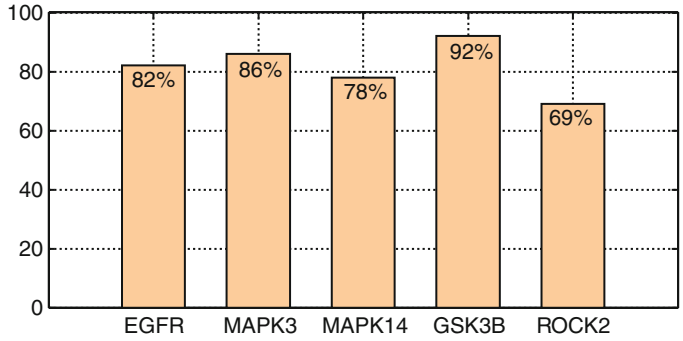


Fig. 3.3 Bechmarking our results. The percentages are calculated using the following formula: $P = \frac{|S_p \cap S_d|}{|S_p|} \times 100$, where S_p is the set of interactions predicted by our algorithm, S_d is the set of interactions either documented in other databases or predicted by online prediction services as mentioned in the main text

of a pathway. For example, in EGFR pathway, EGF receptors form complexes with Shc and Grb2 which then forms complex with Sos1 before phosphorylating Ras-GDP. This entire series of events can not be inferred using our methodology. Different types of quantitative datasets are needed to infer all different types of interactions in a pathway. Hence, in this paper we constructed a partial picture of the EGFR/MEK/ERK pathway using only those phosphosites present in the dataset of [13]. A cartoon of the inferred pathway is shown in Fig. 3.4. In Fig. 3.4, the black arrows indicate interactions which are predicted using our algorithm and are detected in other studies such as [7, 18, 19, 32]. The red arrows are phosphorylation interactions which are predicted in only our study. The gray arrows indicate interactions which are not detected in our study mainly due to absence of the related phosphosites in the [13]'s database. For example, Raf1 phosphorylates MEK in Ser218 and Ser222 site but these phosphosites are not measured in [13]'s dataset. We included these phosphorylations into our diagram in order to maintain consistency of the pathway. The green arrows in the diagram represent translocation of proteins. The interactions which are indicated to take place in the cytoplasm are inferred from the concentration measurements of the cytoplasmic fraction of the corresponding phosphosites and the interactions which are indicated to take place in the nucleus are inferred from the concentration measurements of the nuclear fractions of the corresponding phosphosites. It is not, however, clear whether the phosphorylation events which are inferred from the nuclear fractions of the phosphosites take place in the nucleus or they take place in the cytoplasm and then the phosphorylated proteins are translocated to the nucleus. The same is true for the interactions which are inferred from the cytoplasmic fractions of the phosphosite concentrations.

Despite the partial nature of the inferred pathway our algorithm has several advantages over currently available pathway inference techniques. The advantages and disadvantages of our algorithm over other pathway inference techniques are discussed in the following section.

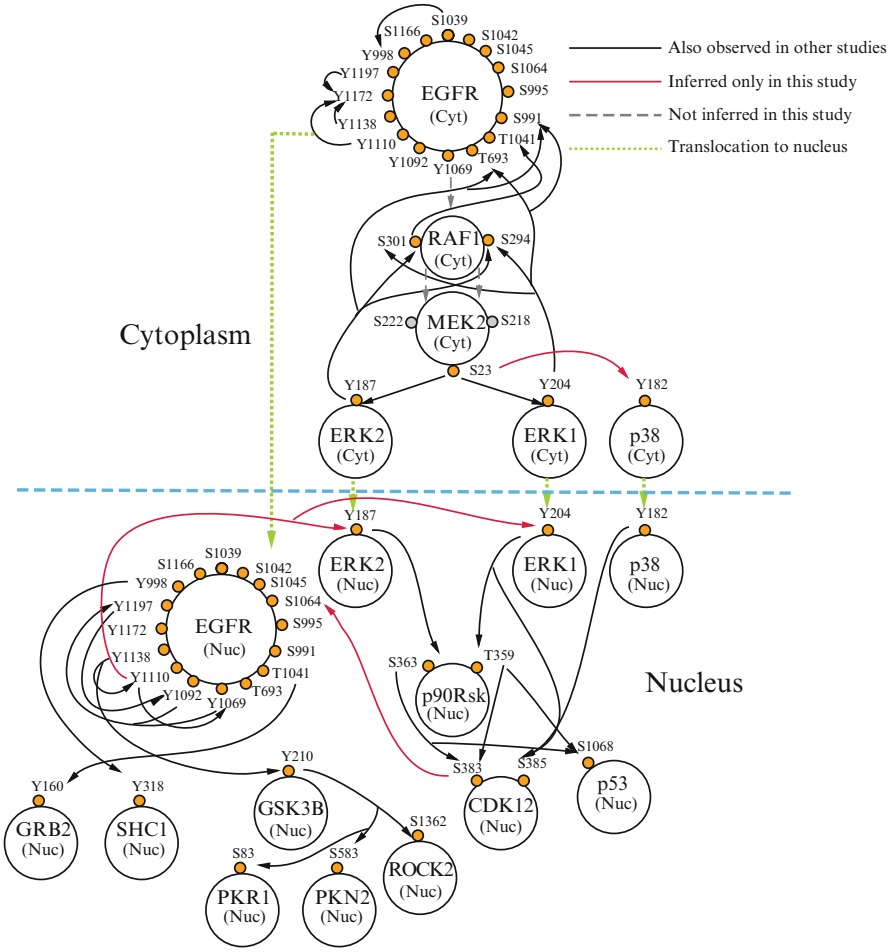


Fig. 3.4 Partial EGFR/MEK/ERK pathway inferred by our algorithm. In this pathway we have shown interactions between some of the kinases which are present in [13]'s data. We have also included some well knows proteins such as Shc1, Grb2, and p53. The partial nature of the pathway arises from inclompleteness of the data

4 Advantages and Disadvantages of Our Algorithm

Many of the currently available algorithms for inferring phosphorylation networks use non quantitative data, e.g., [7] and [32]. Given a substrate, these algorithms efficiently predict the probable kinases which phosphorylate its phosphorylation sites. But if a particular phosphosite has more than one probable kinases, the above algorithms fall short of determining which of the probable kinases phosphorylate it under certain experimental condition. Though [7] developed an efficient

methodology to discriminate among probable kinases based on contextual data, the story might be very different in some cases when quantitative data is taken into account. Another important problem of most network inference techniques is that they do not differentiate among the different phosphorylation states of the kinases themselves when inferring probable substrates. Our algorithm can deal with these problems efficiently. On the other hand some of the algorithms which explicitly use quantitative data such as static Bayesian network inference [5] rely on directed acyclic network architecture and are inefficient in detecting feedback loops. Our algorithm is based on clustering technique and do not have this limitation. Some of the advantages of our algorithm is discussed below with demonstrating examples.

4.1 Advantages

4.1.1 Inter Kinase Specificity

When a phosphosite has more than one probable kinases, it is often difficult to detect which kinase phosphorylates it under certain experimental condition. We call this phenomena inter kinase specificity. An example of inter kinase specificity we encountered in [13]'s data is as follows. When Hela cells are stimulated by EGF, Dcp1 protein is found phosphorylated at Ser 315 site. The amino acid sequence surrounding this phosphosite is PTYTIPLS(p)PVLSPTL which contains the consensus motifs for both GSK3B (S-X-X-X-S) and ERK (P-X-(S/T)-P). But which of these kinases phosphorylates Dcp1 at Ser 215 is not clear. Based on contextual data the NetworKIN algorithm [7] predicts that GSK3B has a higher probability of phosphorylating Dcp1 at Ser 215 compared to ERK. But our algorithm finds that ERK has a higher probability of phosphorylating Dcp1 compared to GSK3B mainly due to relatively high correlation between their concentrations (see Fig. 3.5) under the experimental setup of [13]. It should also be noted that the high correlation between the concentrations of Dcp1(S315) phosphosite occurs with that of ERK2(T201) phosphosite and not any other phosphorylated form of ERK. This reveals another important feature of our algorithm, i.e., intra kinase specificity detection. A more well known example of intra kinase specificity detected by our algorithm is as follows.

4.1.2 Intra Kinase Specificity

It is well known that EGFR phosphorylates itself upon EGF stimulation [33]. But the detailed mechanism of this autophosphorylation in not clear, i.e., which phosphorylated state of EGFR phosphorylates itself at what site is unclear. From [13]'s data our algorithm predicts that upon EGF stimulation EGFR, when phosphorylated at Y1110, Y1138, and Y1197 phosphorylates itself at Y1172. However, when EGFR is phosphorylated at S695 or T693 it can not

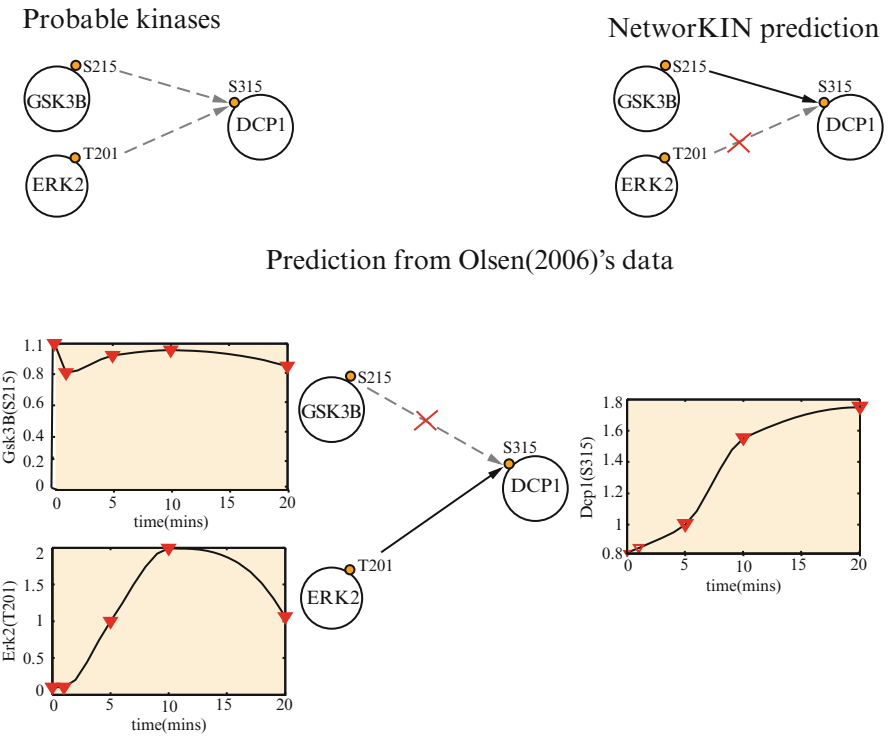


Fig. 3.5 Interkinase specificity. The gray arrows indicate probable interactions. The black arrows indicate predicted interaction

phosphorylate itself at Y1172. We have demonstrated the temporal profiles of the phosphorylation sites mentioned above and the predicted interactions in Fig. 3.6. The intra kinase specificity of EGFR may be important since it is found that EGFR autophosphorylates itself at different sites when stimulated by different EGF like growth factors such as EGF and betacellulin [34].

4.1.3 Feedback Interactions

Feedback interactions have important roles in the dynamic behavior of signaling pathways. Our algorithm is based on a clustering framework and not specifically designed to detect feedback regulations. However, it detects some of the feedback interactions as byproducts. For example, the auto phosphorylation of EGFR and few feedback regulations from ERK to Raf and ERK to EGFR are detected (see Fig. 3.4).

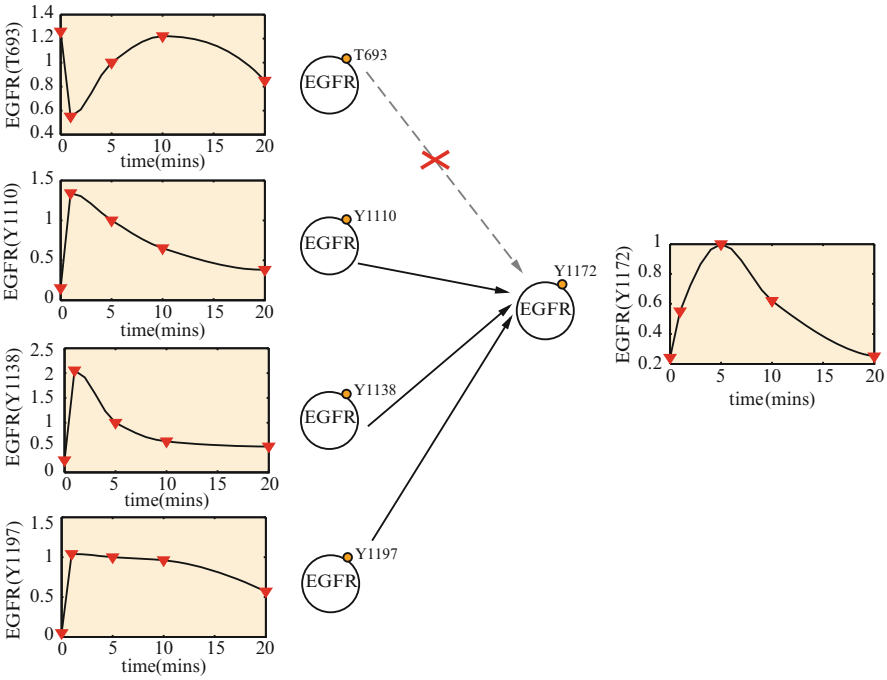


Fig. 3.6 Intrakinase specificity. The gray arrows indicate possible interaction and the black arrows indicate predicted interactions

4.2 Disadvantages

There are, however, several disadvantages of our algorithm. The most important disadvantage is its inability to deal with hidden variables, in this case, kinases which are not measured in a particular experimental set up. For example many of the phosphosites of MEK are not measured in [13]'s dataset which leads to both incomplete and false predictions. Accommodating such unobserved variables in our algorithm will provide a more complete picture and enhance the accuracy of the predictions.

Another problem of our algorithm is its inability to deal with inhibitory phosphorylations. For example, the inhibitory phosphorylation site Y527 of SRC is measure in [13]'s experiment but our algorithm is unable to predict the SRC substrates due to low correlation between the kinase and the substrate phosphosite concentrations. Though this problem may appear to be easier to deal with a systematic formulation for such cases is yet to be incorporated in our algorithm.

Finally, our algorithm is based on a clustering framework and not on any network inference framework. Many important features of a signaling pathways such as

feedback interactions, competitive phosphorylation, etc., can only be thought of as byproducts of our algorithm since it is not specifically designed to detect these properties and hence might introduces errors in prediction.

5 Conclusion

Recent high throughput proteomic experiments generated a wealth of data. How to make biologically interpretable inference from these data using mathematical and statistical techniques is a major challenge. We undertook the problem of network inference from high throughput proteomic data. Such data usually come in limited number of repetitions which limits the applicability of most well known statistical network inference methods such as Bayesian Network inference, etc. Hence, ours is an effort to make new statistical tools to make useful prediction from such data. We used a Naive Bayes based clustering framework to determine pairwise relationship between the kinases and their substrates. Our results suggest that our method is at the very least comparable to other methodologies developed for the same purpose. Additionally, it offers some new capabilities such as detection of inter kinase specificity and intra kinase specificity. These type of information may reveal more detailed picture of signaling mechanisms than what is already known. However, the Naive Bayes architecture of our algorithm may be too simplistic and can be improved by using more sophisticated models which takes into account the interrelationship between different types of features. Additionally, our algorithm can deal only with time resolved mass spectrometry data. Recent high throughput proteomic data comes in many flavors. For example, inhibitory mass spectrometry data, imaging data, etc. We are currently engaged in developing more accurate methods for different types of high throughput data and planning to use these algorithms on data being generated in our lab.

References

1. Janes K, Kelly J, Gaudet S, Albeck J, Sorger P, Lauffenburger D (2004) Cue-signal-response analysis of tnf-induced apoptosis by partial least squares regression of dynamic multi-variate signaling network measurements. J Comp Biol (11):544–561
2. Woolf P, Prudhomme W, Daheron L, Daley G, Lauffenburger D (2005) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics (21):741–753
3. Sachs K, Perez O, Peter D, Lauffenburger D, Nolan G (2005) Causal protein signaling networks derived from multiparameter single-cell data. Science (308):523–529
4. Locasale J, Yadlin A (2009) Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PLoS One (4):e6522
5. Wagner J, Lauffenburger D (2009) Bayesian network inference of phosphoproteomic signaling networks. In: Seventh Annual Workshop on Bayes Applications, Montreal, Canada
6. Sachs K, Itani S, Carlisle J, Nolan G, Peer D, Lauffenburge D (2009) Learning signaling network structures with sparsely distributed data. J Comput Biol (16):1–12

7. Linding R, Jensen LJ, Ostheimer G, Vugt M, Jorgensen C, Miron I, Diella F, Colwill K, Taylor L, Elder K, Metalnikov P, Nguyen V, Pasculescu A, Jin J, Park J, Samson L, Woodgett J, Russell RB, Bork P, Yaffe M, Pawson T (2007) Systematic discovery of in vivo phosphorylation networks. Cell (129):1415–1426
8. Hjerrild M, Stensballe A, Rasmussen T, Kofoed C, Blom N, Sicheritz-Pontén T, Larsen M, Brunak S, Jensen O, Gammeltoft S (2004) Gammeltoft, identification of phosphorylation sites in protein kinase a substrates using artificial neural networks and mass spectrometry. J Proteome Res (3):426–433
9. Obenauer J, Cantley L, Yaffe M (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res (31):3635–3641
10. Puntervoll P, Linding R, Gemnd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin D, Ausiello G, Brannetti B, Costantini A, et al. (2003) Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res (31):3625–3630
11. Kaderali L, Dazert E, Zeuge U, Frese M, Bartenschlager R (2009) Reconstructing signaling pathways from rnai data using probabilistic boolean threshold network. Bioinformatics (25):2229–2235
12. Froehlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T (2007) Large scale statistical inference of signaling pathways from rnai and microarray data. BMC Bioinformatics (8):1–15
13. Olsen J, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen M, Mann P (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell (127):635–648.
14. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen L, von Mering, C (2011) The string database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res (39):D561–D568
15. Fisher RA (1921) On the probable error of a coefficient of correlation deduced from a small sample. Metron (1):03–32
16. Hand DJ, Yu K (2001) Idiot's bayes: not so stupid after all? Int Stat Rev (69):385–398
17. Prasad T, et al. (2009) Human protein reference database – 2009 update. Nucleic Acids Res (37):D767–772
18. Gnad F, Ren S, Cox J, Olsen J, Macek B, Oroshi M, Mann M (2007) Phosida (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol (8):R250
19. Hornbeck P, Chabra I, Kornhauser J, Skrzypek E, Zhang B (2004) Phosphosite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics (4):1551–1561
20. Dinkel H, Chica C, Via A, Gould C, Jensen L, Gibson T, Diella F (2010) Phospho.elm: a database of phosphorylation sites – update 2011. Nucleic Acids Res (39):D261–D267
21. Huang H, Lee T, Tzeng S, Horng J (2005) Kinasephos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res (33):W226–W229
22. Huang H, Lee T, Tzeng S, Wu L, Horng J et al. (2005) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. J Comput Chem (26):1032–1041
23. Senawongse P, Dalby A, Yang Z (2005) Predicting the phosphorylation sites using hidden markov models and machine learning methods. J Chem Inf Model (45):1147–1152
24. Satyanarayana A (1982) A unified formula for analysis of some network reliability problems. IEEE Trans Reliab (R31):23–31
25. Satyanarayana A, Prabhakar A (1978) New topological formula and rapid algorithm for reliability analysis of complex networks. IEEE Trans Reliability (R-27):82–100
26. Satyanarayana A, Chan M (1983) Network reliability and the factoring theorem, Networks (13):107–120
27. Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat (3):119–131
28. Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences and processes. Springer-Verlag, New York
29. Zhang Y, Wolf-Yadlin A, Ross PL, Pappin D, Rush J, Lauffenburger D, White F (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. Mol Cell Proteom (4):1240–1250

30. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The international protein index: an integrated database for proteomics experiments. Proteomics (4):1985–1988
31. Lee TY, Hsu J, Chang W, Huang H (2010) Regphos: a system to explore the protein kinase-substrate phosphorylation network in humans. Nucleic Acids Res (39):D777–D787
32. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics (4):1633–1649
33. Guoa L, Kozloskya C, Ericssona L, Daniela TO, Cerrettia DP, Johnson R (2003) Studies of ligand-induced site-specific phosphorylation of epidermal growth factor receptor. J Am Soc Mass Spectrom (14):1022–1031
34. Saito T, Okada S, Ohshima K, Yamada E, Sato M, Uehara Y, Shimizu H, Pessin J, Mori, M (2004) Differential activation of epidermal growth factor (egf) receptor downstream signaling pathways by betacellulin and egf. Endocrinology (145): 4232–4243

Chapter 4
Signaling Cascades: Consequences of Varying Substrate and Phosphatase Levels

Elisenda Feliu, Michael Knudsen, and Carsten Wiuf

Abstract We study signaling cascades with an arbitrary number of layers of one-site phosphorylation cycles. Such cascades are abundant in nature and integrated parts of many pathways. Based on the Michaelis–Menten model of enzyme kinetics and the law of mass-action, we derive explicit analytic expressions for how the steady state concentrations and the total amounts of substrates, kinase, and phosphatates depend on each other. In particular, we use these to study how the responses (the activated substrates) vary as a function of the available amounts of substrates, kinase, and phosphatases. Our results provide insight into how the cascade response is affected by crosstalk and external regulation.

1 Introduction

Reverse phosphorylation of proteins is one of the principal mechanisms by which signals are transmitted in living cells. Signaling pathways typically contain a cascade of phosphorylation cycles involving kinases and phosphatases, where the activated (in general, the phosphorylated) protein in one layer acts as the kinase in the next layer. The levels of substrate, phosphatase or stimulus in these cycles might be regulated externally by other proteins. Many disease-related proteins are

E. Feliu (✉)
Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
e-mail: efeliu@birc.au.dk

M. Knudsen • C. Wiuf
Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

Centre for Membrane Pumps in Cells and Disease (PUMPKIN), Aarhus University, Aarhus, Denmark
e-mail: michaelk@birc.au.dk; wiuf@birc.au.dk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_4, © Springer Science+Business Media, LLC 2012

part of signaling pathways, for example, the tumor suppressor proteins BRCA1 and p53 exist in many phosphoforms and the PTEN protein is a phosphatase. Cascade malfunctioning might therefore be a cause of disease (e.g., [1,2]). It is a goal of this work to understand how a signaling cascade adjusts to changes in the amount of initial stimulus or the amount of phosphatase and substrate in specific layers.

The biological relevance of this signaling mechanism is well-established theoretically [3–7]. Properties of general signaling cascades, such as ultrasensitivity and signal amplification, might be elucidated from the study of signaling cascades with an arbitrary number of layers n, where each layer is a one-site phosphorylation cycle. Such cascades are part of many pathways [8, page 342], [9, 10] and have been investigated mathematically: $n = 1$, e.g., [11,12], $n = 1, 2$, e.g., [5,13,14], and arbitrary n, e.g., [3, 7, 15, 16]. In much previous work, a cascade is modeled as a system of independent layers, thereby ignoring the effect of kinase sequestration. This was pointed out in [7]. This simplification further implies that one cannot study how activation of one layer effects the concentration levels in the layers of the upstream. To study this it is crucial to consider connected layers.

Here, we give an analysis of a cascade with n connected layers. We provide analytic expressions for how species concentrations and total amounts of substrates and phosphatases are related. Specifically, we use Michaelis–Menten's classical model of an enzyme reaction which includes the formation of intermediate complexes (thus accounting for sequestration). Based on mass-action kinetics we derive a system of differential equations and compute the steady states using an iterative procedure to eliminate variables. This approach makes it possible to derive exact relationships between concentrations and total amounts at steady state and to study aspects of the system in detail without relying on simulation or numerical evaluations. Our work is an extension of the work in [17], where we gave a detailed mathematical analysis of this cascade at steady state.

The outline of the paper is provided in the following manner. In Sect. 2 we describe the system. In Sect. 3 we give the main mathematical results that we derive about the system. Non-mathematically inclined readers might skip this section. In Sect. 4 we study how species concentrations at steady state vary as a function of the overall substrate and phosphatase levels. We take this further in Sect. 5, where we study stimulus–response and signal amplification. Finally, in Sect. 6, the question of how the maximal response relates to the number of layers as well as the levels of phosphatase or substrate is addressed.

2 One-Site Linear Signaling Cascades

We consider signaling cascades with n layers and a one-site phosphorylation cycle at each layer (Fig. 4.1). The species in each cycle are the *unmodified* substrate S_i^0, the *modified* substrate S_i^1, the *phosphatase* F_i, the *kinase* S_{i-1}^1, and the *intermediate (enzyme–substrate) complexes* Y_i^0 and Y_i^1 for $i = 1, \ldots, n$. That is, in each layer the kinase is the phosphorylated substrate of the previous layer. The kinase of the first

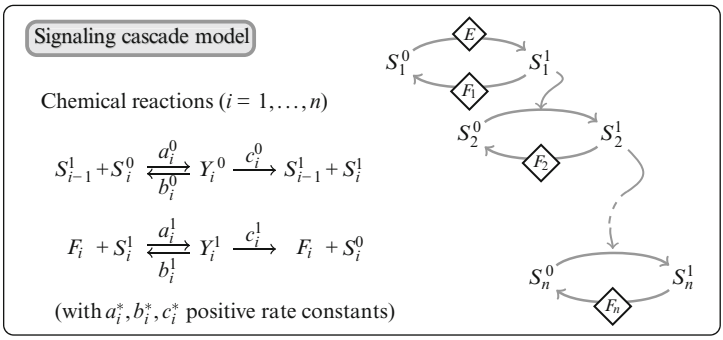


Fig. 4.1 One-site cascade of length n. The enzyme mechanism follows the classical Michaelis–Menten model. In the first layer, the substrate S^1_0 is the kinase E is the first cycle

layer is not a substrate in any other layer and we denote it by $E = S^1_0$ (corresponding to a 0th layer). The modified substrate S^1_i of the ith layer is called the *response* of the ith layer; in particular the response of the nth layer is called the *final* response or simply the response of the cascade.

The system is specified by the set of chemical reactions (Fig. 4.1). The enzyme mechanism follows the classical model of Michaelis and Menten, in which an enzyme–substrate complex is formed reversibly, while its dissociation into product and enzyme is irreversible. Further, the phosphate donor, typically ATP, is assumed in abundance and embedded into the rate constants. This reaction set-up has frequently been used to study signaling cascades, see e.g., [5, 7, 14, 18–20].

2.1 Steady States

Assuming mass-action kinetics, the differential equations describing the dynamical system over time t are given by:

$$\dot{S}^1_i = (b^0_{i+1} + c^0_{i+1})Y^0_{i+1} + c^0_i Y^0_i + b^1_i Y^1_i - (a^0_{i+1}S^0_{i+1} + a^1_i F_i)S^1_i \tag{4.1}$$

$$\dot{S}^0_i = b^0_i Y^0_i + c^1_i Y^1_i - a^0_i S^0_i S^1_{i-1} \tag{4.2}$$

$$\dot{Y}^0_i = -(b^0_i + c^0_i)Y^0_i + a^0_i S^0_i S^1_{i-1} \tag{4.3}$$

$$\dot{E} = (b^0_1 + c^0_1)Y^0_1 - a^0_1 S^0_1 E \tag{4.4}$$

$$\dot{F}_i = (b^1_i + c^1_i)Y^1_i - a^1_i F_i S^1_i \tag{4.5}$$

$$\dot{Y}^1_i = a^1_i F_i S^1_i - (b^1_i + c^1_i)Y^1_i \tag{4.6}$$

for $i = 1\ldots,n$ and where we put $Y^0_{n+1} = S^0_{n+1} = 0$. It follows from Equations (4.5) and (4.6) that $\dot{F}_i + \dot{Y}^1_i = 0$. Similarly, from (4.4) and (4.3) for $i = 1$ we have

that $\dot{E} + \dot{Y}_1^0 = 0$. This implies that the values $F_i + Y_i^1$ and $E + Y_1^0$ are independent of time. Similarly, $S_i^0 + S_i^1 + Y_i^0 + Y_i^1 + Y_{i+1}^0$ is also constant. Hence, the system has the following *conservation laws*:

$$\overline{F}_i = F_i + Y_i^1, \quad \overline{E} = E + Y_1^0, \quad \overline{S}_i = S_i^0 + S_i^1 + Y_i^0 + Y_i^1 + Y_{i+1}^0, \tag{4.7}$$

for $i = 1, \ldots, n$, and $Y_{n+1}^0 = 0$. The quantities $\overline{E}$, $\overline{F}_i$, and $\overline{S}_i$ are called the total amounts of enzymes and substrates, or just the total amounts.

The steady states of the cascade are found by setting the right hand side of (4.1)–(4.6) to zero. The conservation laws imply that the equations corresponding to $\dot{S}_i^0, \dot{E}, \dot{F}_i = 0$ are redundant. Therefore, given total amounts $\overline{E}, \overline{F}_i, \overline{S}_i$, the steady states of the system are the concentrations that fulfill the conservation laws (4.7) (linear equations) together with $\dot{S}_i^1, \dot{Y}_i^0, \dot{Y}_i^1 = 0$ (quadratic equations).

These equations provide a system of polynomial equations with $5n + 1$ equations and variables which, because of the quadratic equations, may have many solutions. However, we are only interested in biologically relevant solutions for which all concentrations at steady state are positive or zero. This suggests the following definition: A *Biologically Meaningful Steady State* (BMSS) is a steady state for which all total amounts are positive and all species concentrations are positive or zero.

In [17], we prove that the cascade has precisely one BMSS for any choice of kinetic rate constants. Further, we show that the BMSS concentrations are in fact positive (i.e., non-zero) and hence each concentration at steady state is strictly smaller than a corresponding total amount, e.g., $E < \overline{E}$. By abuse of language, we often say "the steady state", while meaning the BMSS. Likewise, we say, e.g., "the kinase E fulfills..." when in fact we mean "the concentration of the kinase E fulfills...".

Having set the notation, we can formalize the scope of this work: we seek to study how the BMSS (in particular the response S_i^1) changes when the total amounts $\overline{E}$, $\overline{S}_i$ or $\overline{F}_i$ change, and how a change in one layer effects the responses in other layers.

2.2 Concentrations at Steady State

Using (4.1)–(4.6) together with the conservation laws the following relations apply at steady state,

$$F_i = \frac{\overline{F}_i}{1 + \delta_i S_i^1}, \quad Y_i^1 = \frac{\delta_i \overline{F}_i S_i^1}{1 + \delta_i S_i^1}, \quad Y_i^0 = \frac{\gamma_i \overline{F}_i S_i^1}{1 + \delta_i S_i^1}, \quad S_i^0 = \frac{\lambda_i \overline{F}_i S_i^1}{(1 + \delta_i S_i^1) S_{i-1}^1} \tag{4.8}$$

for $i = 1, \ldots, n$, with constants $\delta_i = a_i^1/(b_i^1 + c_i^1)$, $\gamma_i = (c_i^1/c_i^0)\delta_i$, and $\lambda_i = \gamma_i(b_i^0 + c_i^0)/a_i^0$. The constant δ_i is the inverse of the Michaelis–Menten constant for F_i, γ_i is the catalytic efficiency $c_i^1 \delta_i$ of F_i divided by the dissociation constant c_i^0 of S_{i-1}^1, and λ_i is the relative catalytic efficiency in layer i, that is, the quotient of the catalytic efficiency of F_i by that of S_{i-1}^1.

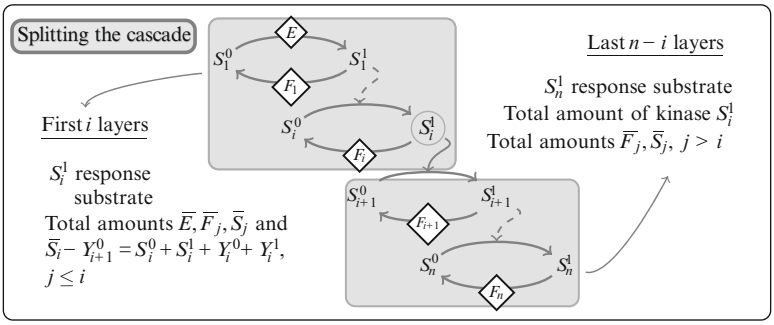


Fig. 4.2 Splitting the cascade at the ith layer

Equation (4.8) is essential: It provides simple relationships between different concentrations at steady state. The values Y_i^0, Y_i^1, and F_i depend only on the rate constants in the ith layer and are increasing in $\overline{F}_i$ and S_i^1. The value S_i^0, however, depends on the steady state values of the modified substrates in the ith and $(i-1)$th layers, providing a link between the two layers.

2.3 Splitting the Cascade

Consider the cascade obtained from the first i layers. Its connection to the last $n-i$ layers is through the intermediate complex Y_{i+1}^0 accounting for the conversion of S_{i+1}^0 to S_{i+1}^1 via the kinase S_i^1. If Y_{i+1}^0 is known, then the steady state concentrations in the first i layers satisfy the steady state equations of a cascade of length i with total amounts $\overline{E}, \overline{F}_1, \ldots, \overline{F}_i, \overline{S}_1, \ldots, \overline{S}_{i-1}$, and $\overline{S}_i - Y_{i+1}^0 = S_i^0 + S_i^1 + Y_i^0 + Y_i^1$. Thus, the intermediate complex Y_{i+1}^0 influences the layers upstream of layer $i+1$ by reducing the total amount of substrate available at layer i. This effect is known as sequestration.

Similarly for the cascade consisting of the last $n-i$ layers. If S_i^1 is known (fixed), then the steady state concentrations in the layers $i+1, \ldots, n$ satisfy the steady state equations of a cascade of length $n-i$ with total amounts $\overline{F}_{i+1}, \ldots, \overline{F}_n$, $\overline{S}_{i+1}, \ldots, \overline{S}_n$ and total amount of kinase S_i^1.

This split is illustrated in Fig. 4.2. The results presented in the following sections rely on splitting the cascade in this way.

3 Relationships Between Response Concentrations

In this section we provide an iterative expression for the ith response S_i^1 in terms of the final response S_n^1. Some consequences of this result are discussed in the forthcoming sections.

3.1 The Last Layer

If the expressions in (4.8) are substituted for Y_n^0, Y_n^1, S_n^0 in the conservation law $\overline{S}_n = S_n^0 + S_n^1 + Y_n^0 + Y_n^1$ we obtain S_{n-1}^1 as an (increasing) function of S_n^1,

$$S_{n-1}^1 = f_{n-1}(S_n^1) = \frac{\lambda_n \overline{F}_n S_n^1}{d_n(S_n^1, 0)},$$

with $d_i(x, y) = (\overline{S}_i - y) - x - \overline{F}_i(\delta_i + \gamma_i)x + \delta_i(\overline{S}_i - y)x - \delta_i x^2$, $1 \leq i \leq n$. If S_n^1 is positive, then S_{n-1}^1 is positive provided $d_n(S_n^1, 0)$ is positive. This is the case only if $S_n^1 \in [0, \alpha_n)$, where α_n is the only positive root of $d_n(x, 0)$. Therefore, $\overline{F}_n$, $\overline{S}_n$, and the rate constants of layer n restrict S_n^1 at steady state, $S_n^1 < \alpha_n$, independently of the parameters in the other layers.

If S_n^1 is close to α_n, the denominator of f_{n-1} is close to zero and hence S_{n-1}^1 is large. Since the amount of substrate in layer $n-1$ is bounded by $\overline{S}_{n-1}$, S_{n-1}^1 cannot be arbitrarily large. Thus, upstream layers limit the possible values of S_n^1 further.

3.2 Intermediate Layers Response

In (4.8), S_{i+1}^1 gives Y_{i+1}^0. This observation allows us iteratively to calculate all responses S_i^1 as functions of S_n^1. Specifically, consider the ith layer of the cascade. For every $Y_{i+1}^0 < \overline{S}_i$, the steady state values of the species in the first i layers are found by solving the steady state equations for the cascade consisting of the layers from 1 to i with the total amount of substrate in layer i being $\overline{S}_i - Y_{i+1}^0$. Therefore, we obtain

$$S_{i-1}^1 = g_i(S_i^1, Y_{i+1}^0) = \frac{\lambda_i \overline{F}_i S_i^1}{d_i(S_i^1, Y_{i+1}^0)}, \tag{4.9}$$

with $g_n(S_n^1, Y_{n+1}^0) = f_{n-1}(S_n^1)$, since $Y_{n+1}^0 = 0$. The response S_{i-1}^1 can be found in terms of S_n^1 by repeated application of (4.9). Positivity of S_{i-1}^1 imposes an upper bound β_{i-1} to S_n^1, which is smaller than the upper bound β_i imposed by S_i^1. Indeed, when S_n^1 is close to β_i, S_i^1 is large and then d_i becomes negative.

We have outlined the following result, which is proven in [17, Prop. 2.31].

Result 1 (Response relationships) *For $i = 0, \ldots, n-1$, the BMSS value of S_i^1 satisfies $S_i^1 = f_i(S_n^1)$, where f_i is an increasing function of S_n^1 defined on an interval $[0, \beta_i)$. Furthermore,*

- *$\beta_i < \beta_{i+1}$ and $\beta_i < \beta_{n-1} = \alpha_n$ for $i < n-1$. β_i depends on $\overline{F}_j, \overline{S}_j$, $j \geq i+1$ only.*
- *Let α_i be the positive root of $d_i(x, 0)$ which depends on $\overline{F}_i$ and $\overline{S}_i$. Then $S_i^1 < \alpha_i$ for any BMSS.*

This result is important and shows how each additional layer further constrains the maximal value of S_n^1. Also, the response S_i^1 is bounded by α_i, which depends exclusively on the rate constants and total amounts of layer i. This upper bound is obtained by ignoring sequestration, i.e., assuming $Y_{i+1}^0 = 0$.

Result 1 provides an iterative procedure for calculating response relationships. All terms that appear in the function f_i are mathematically simple (polynomials) and hence f_i is a rational function. Such functions are easy to manipulate, for example, using programs like Mathematica™.

3.3 Total Amount of Kinase $\overline{E}$

Using Result 1 we obtain the increasing relations $E = S_0^1 = f_0(S_n^1)$ and $Y_1^0 = \frac{\gamma_1 \overline{E}_1 f_1(S_n^1)}{1+\delta_1 f_1(S_n^1)}$. The latter we denote $Y_1^0 = f_1^Y(S_n^1)$. These functions do not dependent on the stimulus $\overline{E}$ and hence

$$\overline{E} = r\left(S_n^1\right) = f_0\left(S_n^1\right) + f_1^Y\left(S_n^1\right)$$

gives $\overline{E}$ as an increasing function of S_n^1. The function f_0 is defined for $S_n^1 < \beta_0$ and tends to infinity as S_n^1 tends to β_0. The function f_1^Y is defined for $S_n^1 < \beta_1$. Since $\beta_0 < \beta_1$, the function r is increasing and defined for $S_n^1 \in [0, \beta_0)$. It tends to infinity when S_n^1 tends to β_0.

As a consequence, for positive $\overline{E}$, there is a unique value of S_n^1 satisfying the relation $\overline{E} = r(S_n^1)$. This is the BMSS value of S_n^1. All other concentrations can be derived from this using (4.8) and the functions g_i. Further, the upper bound β_0 of S_n^1 is only obtained if $\overline{E}$ is very large. We introduce a distinctive symbol for this upper bound, or the *maximal response* of the cascade: $\sigma_n := \beta_0$. Writing r as a quotient of polynomials, σ_n is simply the first positive root of the denominator.

4 Regulation Through Substrate and Phosphatase Variation

In the previous section we found S_i^1 in terms of S_n^1. This relation provides means to explore how noise and regulation at intermediate layers (e.g., *crosstalk* [14]) propagate upstream and downstream in the cascade and effects the responses.

The following result is from [17, Th. 2.32, Th. 2.33] and illustrated in Fig. 4.3a.

Result 2 (Variation in the total amount of substrate) *Consider a cascade with n layers and fix all total amounts but $\overline{S}_i$ for some layer i. Then an increase of $\overline{S}_i$ causes*:

- *The BMSS values of the response S_j^1 and the intermediate complexes Y_j^0 and Y_j^1 increase downstream of layer i, that is, for layers $j = i, \ldots, n$.*

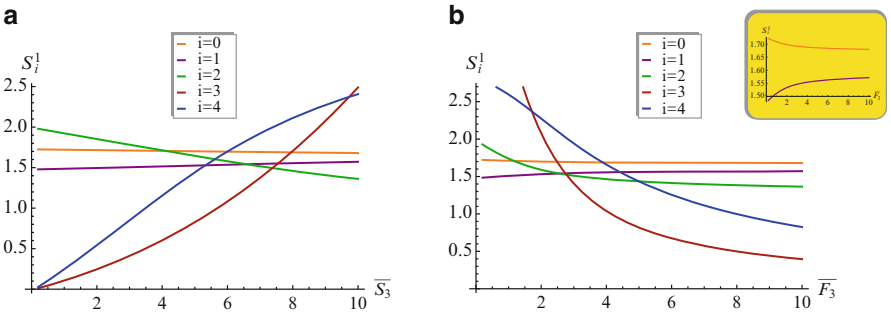


Fig. 4.3 Variation of S_i^1 when the total amounts of phosphatase or substrate are varied. Fixed parameters: $a_*^* = b_*^* = c_*^* = 1, \overline{F}_* = 3, \overline{E} = 3, \overline{S}_* = 7$

- *The BMSS value of the response S_j^1 increases upstream of layer i if $j = 1, \ldots, i-1$ has the same parity as i and decreases otherwise.*

Thus an increase in the total amount of substrate in one intermediate layer propagates downstream as an increase in the concentrations of the modified substrates. This corresponds to increasing the initial kinase or stimulus S_i^1 in the smaller cascade consisting of the layers below the one undergoing variation. However, these layers have fixed total amounts and the modified substrates downstream are therefore bounded by their respective α_i (Result 1).

Also, an increase in the total amount of substrate in an intermediate layer propagates upstream in an alternating fashion. If S_i^1 is increased, so is the sequestered substrate Y_i^0 and hence the total amount at layer $i-1$, $\overline{S}_{i-1} - Y_i^0$ decreases. In turn, this causes S_{i-1}^1 to decrease. In turn, this causes Y_{i-1}^0 to decrease and hence $\overline{S}_{i-2} - Y_{i-1}^0$ to increase and so S_{i-2}^1 increases. This effect is strongly dependent on the intermediate complexes and cannot be demonstrated in a model without these.

Similarly, Result 1 provides insight into how the response varies when the total amount of phosphatase is changed (see Appendix A for a proof).

Result 3 (Variation in total amount of phosphatase) *Consider a cascade with n layers and fix all total amounts but $\overline{F}_i$ for some layer i. If the total amount of phosphatase at layer i, $\overline{F}_i$, is increased then*:

- *The BMSS value of the response S_j^1 decreases downstream of layer i, that is, for $j = i, \ldots, n$.*
- *The BMSS value of the response S_j^1 increases upstream of layer i if $j = 1, \ldots, i-1$ has the same parity as i and decreases otherwise.*

Result 3 is illustrated in Fig. 4.3b. As expected, an increase of phosphatase at layer i causes the amount of phosphorylated substrate at layer i to decrease and likewise all downstream responses to decrease too. In particular, the final

response decreases. Thus, controlling the level of phosphatase at any layer serves as a regulator of the response level. Upstream of layer $i-1$ the response increase/decrease in an alternating way, using the same argument as above.

5 Stimulus–Response Curves

The relationship between stimulus and response has been studied extensively, e.g., [7,11,14,16,21]. Much attention has been devoted to whether a system exhibits ultrasensitivity, that is, whether it reacts to input in a switch-like mode [18,22].

The plot of S_n^1 (the final response) against $\overline{E}$ (the stimulus) is usually called the *stimulus–response* curve. We showed in Sect. 3 that the stimulus and the response are related by an increasing function

$$\overline{E} = r(S_n^1)$$

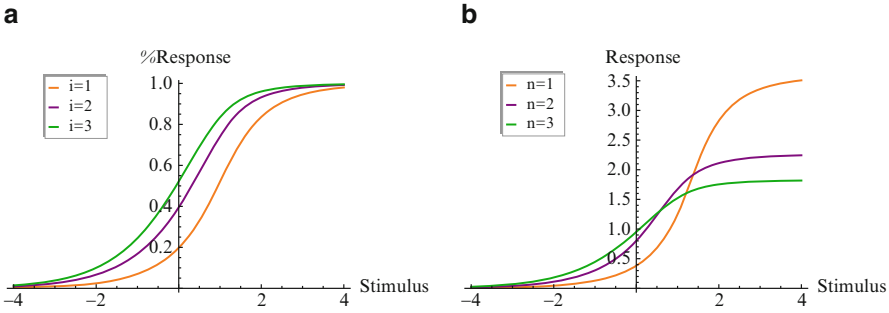
defined on an interval $[0, \sigma_n)$. Thus, the inverse of r is the stimulus–response curve.

When the stimulus $\overline{E}$ is arbitrarily large the final response S_n^1 saturates at its maximal value σ_n. The stimulus required to achieve a certain percentage of the maximal response can be determined from the explicit expression of the inverse stimulus–response curve. Let $\overline{E}_M$ be the value of $\overline{E}$ required to obtain $M\%$ of the maximal response, that is, $\overline{E}_M = r(M\sigma_n/100)$. For instance, 90% of the maximal response is obtained with $\overline{E}_{90} = r(0.9\sigma_n)$. This provides means to compute measures of sensitivity and switch behavior of biological systems: the *response coefficient* (also called cooperativity index) $R = \overline{E}_{90}/\overline{E}_{10}$ [5], the *switch value* $\overline{E}_{90} - \overline{E}_{10}$ [18], and the *Hill coefficient* $n_H = \log(81)/\log(\overline{E}_{90}/\overline{E}_{10})$ [23].

The maximal response σ_i of S_i^1 is easily derived from the maximal response σ_n using $\sigma_i = f_i(\sigma_n)$. Now consider the response in any layer normalized with its maximal response, that is, the normalized, or relative, response is between 0 and 1. We provide conditions for which the normalized response increases when moving down the layers in a cascade for a fixed stimulus $\overline{E}$, Fig. 4.4a. In other words, the normalized stimulus–response curves are shifted to the left as we move down the layers. This is known as signal amplification.

Result 4 (Signal amplification) *If* $1 > \delta_i \overline{S}_i - (\delta_i + \gamma_i)\overline{F}_i$, *then the level of kinase* $\overline{E}$ *required to achieve* $M\%$ *of the maximal response at layer* i *is always smaller than the amount of kinase required to achieve* $M\%$ *of maximal response at layer* $i-1$.

Thus, if the level of phosphatase is in excess relatively to the substrate in all layers, then for any given amount of stimulus, the last layer will always have a higher relative response than the intermediate responses, and the response in the first layer will always have the lowest relative response.



Shift of the normalized stimulus-response curves for the modified substrates S_i^1, i = 1, 2, 3, for $n = 3$.

Reduction of the maximal value of response for different cascade lengths $n = 1, 2, 3$ with parameters in the last layer fixed.

Fig. 4.4 Stimulus–response curves in semi-log scale, $\log(\overline{E})$ versus S_n^1. Fixed parameters: $a_*^* = b_*^* = c_*^* = 1, \overline{F}_1 = 5, \overline{F}_2 = 4, \overline{F}_3 = 5; \overline{S}_1 = 8, \overline{S}_2 = 9, \overline{S}_3 = 10$

6 Maximal Response

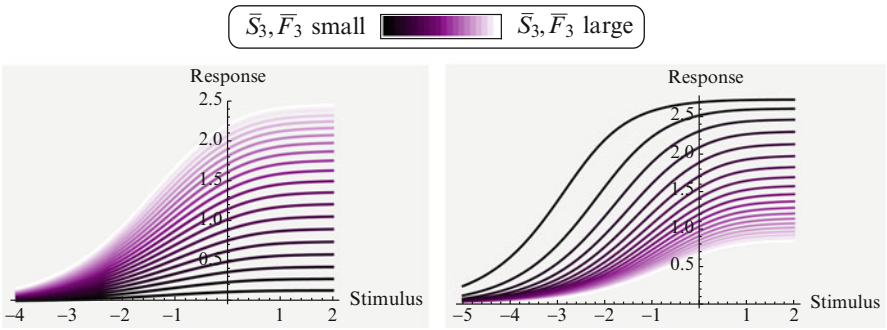
The maximal response is restricted by the total amount $\overline{S}_n$ in the last layer and further by any additional layer, as described in Result 1. The reduction of the maximal response is exemplified in Fig. 4.4b for three cascades with one, two, and three layers. The maximal response of the single-layer cascade is 3.58, but after adding one (respectively two) additional layer(s) on top of it, the maximal response drops to 2.23 (respectively 1.83), which is much lower than the upper bound set by the total amount (fixed to 10). The decline of the maximal response is caused by substrate sequestration: In layers above the last layer, substrates are trapped in intermediate complexes and therefore not able to participate as kinases driving the cascade of modifications that ultimately results in phosphorylation of S_n^0.

How the maximal response changes with changing total amounts of phosphatase and substrate can be quantified. It is stated below and illustrated in Fig. 4.5 (a proof can be found in Appendix A).

Result 5 (Maximal response) *Consider a cascade of length n.*

- *If the total amount of phosphatase at layer i, $\overline{F}_i$, increases then the maximal response σ_n decreases.*
- *If the total amount of substrate at layer i, $\overline{S}_i$, increases then the maximal response σ_n increases.*

Interestingly, an increase in the level of phosphatase at any layer cannot be compensated fully by an increase in the stimulus. Only locally, for low responses, such a loss could be overcome.



Increase of $\overline{S}_3$ causes the maximal response to increase.

Increase of $\overline{F}_3$ causes the maximal response to decrease.

Fig. 4.5 Variation of the maximal response when $\overline{S}_3$ and $\overline{F}_3$ are increased. $\overline{S}_3, \overline{F}_3$ are increased from 0.5 to 10. Fixed parameters: $a_*^* = b_*^* = c_*^* = 1; \overline{F}_* = 3; \overline{E} = 3; \overline{S}_* = 7$

7 Discussion

We have provided a theoretical discussion of linear cascades with arbitrary number of layers of one-site phosphorylation cycles. In particular, we have focused on *intrinsic* properties of the cascade, that is, properties that do not rely on specific reaction rate constants. Such studies may be useful for testing new hypotheses, since experimental data is difficult to obtain and rate constants are hard to estimate [24, 25].

Many cascades are regulated externally, but the effect of such regulation is generally unclear. Our study sheds light on how the steady state changes as a consequence of changing levels of phosphatases and substrates. If a level increases at some layer in the cascade, then all responses downstream decrease (phosphatase) or increase (substrate). Thus, regulation at each layer of the final response is possible. Further, the maximal response (obtained when stimulus is very large) follows the same pattern. An increase in the phosphatase level at some layer causes the maximal response to decrease. This loss cannot be compensated for by an increase in the stimulus.

Upstream of the modified layer variations in the responses follow an alternating behavior: If the response at some layer above the modified layer increases, the response in the next layer decreases, and so forth. This behavior is counterintuitive: The response of one layer is the kinase of the next layer, and we might expect the same qualitative change in each layer. However, the result relies strongly on the formation of intermediate complexes and thus relates to (hidden) sequestration. It is therefore important, that intermediate complexes are modeled explicitly as in our approach.

Under some conditions, signal amplification also occurs in the cascade in the sense that the relative response increases down through the cascade. Thus, in a long

cascade the final response can come up faster than in a short cascade. However, this gain in signal amplification has to be contrasted to a reduction of the maximal response with increasing cascade length. Consequently, the cascade length is a compromise between when and how high the final response should be.

A deeper study is required to understand to what extend this balance between gain and lost is beneficial for the cell. It may depend on the specific levels of phosphatase and substrate as well as on the reaction rate constants. Although the results presented here are qualitatively independent of the rate constants, their effect is crucial in determining the magnitude of a change.

Acknowledgements EF has received support from a postdoctoral grant of the "Ministerio de Educación" of Spain and the project MTM2009-14163-C02-01 from the "Ministerio de Ciencia e Innovación". CW is supported by the Lundbeck Foundation, Denmark.

Appendix

Proofs

The proofs follow very closely the proofs for Results 1 and 2 which can be found in our previous paper [17].

Proof of Result 3. In the sequel, we assume that all total amounts but $\overline{F}_i$ are fixed. Consider a cascade of length n and fix a value of S_n^1. Define

$$Y_i^0 = g_i^Y(S_i^1) = \frac{\gamma_i \overline{F}_i S_i^1}{1 + \delta_i S_i^1}. \tag{4.10}$$

Using (4.10) and (4.9), we see that for $j \geq i$, S_j^1, Y_{j+1}^0 are independent of $\overline{F}_i$. Then, by (4.10), Y_i^0 is an increasing function of $\overline{F}_i$, and so is S_{i-1}^1 by (4.9) (there might be singularities). For $j \geq i-2$, S_j^1, Y_j^0 are increasing in S_{j+1}^1, Y_{j+2}^0 (with expressions not involving $\overline{F}_i$). We conclude that they are increasing in $\overline{F}_i$ for fixed S_n^1. It follows that the steady state value of S_n^1 must decrease if $\overline{F}_i$ is increased. Indeed, we have $\overline{E} = E + Y_1^0$ with E, Y_1^0 increasing both in S_n^1 and $\overline{F}_i$. Since the functions $f_i, \ldots, f_{n-1}$ are independent of $\overline{F}_i$ and increasing in S_n^1, the concentrations S_j^1 for $j = i, \ldots, n$ decrease in $\overline{F}_i$.

As shown in [17], the BMSS of a cascade of length n satisfies $S_n^1 = \psi(\overline{S}_n)$, with ψ an increasing continuous function defined over the non-negative real numbers. Hence, if we consider now the split of the cascade at layer $i-1$, the steady state value of S_{i-1}^1 is given by a decreasing continuous function of Y_i^0, $S_{i-1}^1 = f(Y_i^0) := \psi(\overline{S}_{i-1} - Y_i^0)$, obtained by considering the first $i-1$ layers of the cascade with total amounts $\overline{E}, \overline{F}_1, \ldots, \overline{F}_{i-1}, \overline{S}_1, \ldots, \overline{S}_{i-2}$, and $\overline{S}_{i-1} - Y_i^0$. The function f is independent of $\overline{F}_i$.

Let now $h(\overline{F}_i)$ denote the value of Y^0_{i+1} at steady state, corresponding to the total amount $\overline{F}_i$. By (4.9), and writing S^1_i as a function of Y^0_i using (4.10), we have that $S^1_{i-1} = \widetilde{g}_i(Y^0_i, Y^0_{i+1}) = \widetilde{g}_i(Y^0_i, h(\overline{F}_i))$. If we write

$$\widetilde{g}_i(Y^0_i, Y^0_{i+1}) = \frac{p_1(Y^0_i, Y^0_{i+1})}{p_2(Y^0_i, Y^0_{i+1})},$$

then $p_1(y,z) = \lambda_i y(\xi - y)$, and $p_2(y,z) = (\delta_i + \gamma_i)y^2 - \gamma_i(1/\delta_i + \overline{F}_i + \xi + (\overline{S}_i - z))y + \gamma_i\xi(\overline{S}_i - z)$ with $\xi = \gamma_i\overline{F}_i/\delta_i$. Computing the partial derivative of this function with respect to Y^0_{i+1} and $\overline{F}_i$, we see that $\widetilde{g}_i$ is decreasing in $\overline{F}_i$ and increasing in Y^0_{i+1}. Since h is decreasing in $\overline{F}_i$, it follows that $\widetilde{g}_i(Y^0_i, h(\overline{F}_i))$ is decreasing in $\overline{F}_i$ for any fixed Y^0_i.

The steady state value of the pair (Y^0_i, S^1_{i-1}) for a fixed $\overline{F}_i$, must satisfy both equalities $S^1_{i-1} = f(Y^0_i) = \widetilde{g}_i(Y^0_i, h(\overline{F}_i))$. Since f is independent of $\overline{F}_i$ and $\widetilde{g}_i$ decreases in $\overline{F}_i$, we have that Y^0_i increases in $\overline{F}_i$ while S^1_{i-1} decreases.

It follows that Y^0_{i-1} decreases too. If for $j \le i-1$, Y^0_j increases, then the total amount of layer $j-1$, $\overline{S}_{j-1} - Y^0_j$ decreases and thus S^1_{j-1} decreases. On the contrary, if Y^0_j decreases, then the same arguments shows that S^1_{j-1} increases completing the proof. □

Proof of Result 5. Let $\sigma_n(\overline{F}_i)$ denote the maximal response of S^1_n corresponding to the total amount of phosphatase $\overline{F}_i$. Similarly, denote by $\beta_j(\overline{F}_i)$ the upper bounds of Result 1. Note that d_j is decreasing in $\overline{F}_j$. Since S^1_n decreases in $\overline{F}_i$, $\sigma_n(\overline{F}_{i,1}) \le \sigma_n(\overline{F}_{i,2})$ if $\overline{F}_{i,1} > \overline{F}_{i,2}$. The question is whether they can be equal or not.

Fix $S^1_n = \sigma_n := \sigma_n(\overline{F}_{i,2})$. Let $\overline{\rho}_1(S^1_n) = \rho_1 \circ f^Y_2(S^1_n)$ be defined as the positive root of the polynomial $d_1(x, f^Y_2(S^1_n))$. The maximal response $\sigma_n = \beta_0$ is given by the positive value of S^1_n for which $f_1(S^1_n) = \overline{\rho}_1(S^1_n)$. Thus, we have

$$f_1(\sigma_n, \overline{F}_{i,2}) = \overline{\rho}_1(\sigma_n, \overline{F}_{i,2}), \tag{4.11}$$

where we add the reference to the total amount of phosphatase. As noted in the preceding proof, if σ_n is fixed and $\overline{F}_i$ is increased, then f_1 is an increasing function. Similarly, $f^Y_2(\sigma_n, \overline{F}_i)$ is increasing too, and since $\rho_1(Y^0_2)$ is decreasing in Y^0_2, the function $\overline{\rho}_1$ is decreasing in $\overline{F}_i$. Note that since d_1 decreases in $\overline{F}_1$, the argument applies even if $i = 1$.

It follows that if $\overline{F}_{i,2}$ satisfies equality (4.11), then the equality cannot be satisfied by $\overline{F}_{i,1} \neq \overline{F}_{i,2}$ and the first part of the result follows.

The same reasoning applies to the maximal response following variation on the total amount $\overline{S}_i$ at some layer i. By Result 2, S^1_n increases if $\overline{S}_i$ increases, and thus, using the corresponding notation, we have $\sigma_n(\overline{S}_{i,1}) \le \sigma_n(\overline{S}_{i,2})$ if $\overline{S}_{i,1} < \overline{S}_{i,2}$. It is easy to see that we can proceed as above to rule out equality. One just have to observe that if $\overline{S}_i$ increases, then, for fixed $S^1_n = \sigma_n$, both $f_1, \overline{\rho}^1$ are decreasing functions. □

References

1. Blume-Jensen P, Hunter T (2001) Oncogenic kinase signalling. Nature 411:355–365
2. Sesti G, Federici M, Hribal ML, Lauro D, Sbraccia P, Lauro R (2001) Defects of the insulin receptor substrate (IRS) system in human metabolic disorders. FASEB J 15:2099–2111
3. Chaves, M, Sontag ED, Dinerstein RJ (2004) Optimal length and signal amplification in weakly activated signal transduction cascades. J Phys Chem B 108(39):15311–15320
4. Ferrell JE, Xiong W (2001) Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible. Chaos 11:227–236
5. Goldbeter A, Koshland DE (1981) An amplified sensitivity arising from covalent modification in biological systems. Proc Natl Acad Sci USA 78:6840–6844
6. Qiao L, Nachbar RB, Kevrekidis IG, Shvartsman SY (2007) Bistability and oscillations in the Huang–Ferrell model of MAPK signaling. PLoS Comput Biol 3:1819–1826
7. Ventura AC, Sepulchre JA, Merajver SD (2008) A hidden feedback in signaling cascades is revealed. PLoS Comput Biol 4:e1000041
8. Cooper GM, Hausman RE (2009) The cell. 5th edn. ASM Press, Washington
9. MacFarlane RG (1964) An enzyme cascade in the blood clotting mechanism, and its function as a biochemical amplifier. Nature 202:498–499
10. Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. Ann Rev Cell Dev Biol 21:319–346
11. Bluthgen N, Bruggeman FJ, Legewie S, Herzel H, Westerhoff HV, Kholodenko BN (2006) Effects of sequestration on signal transduction cascades. FEBS J 273:895–906
12. Salazar C, Höfer T (2006) Kinetic models of phosphorylation cycles: a systematic approach using the rapid-equilibrium approximation for protein–protein interactions. Biosystems 83:195–206
13. Goldbeter A, Koshland DE (1984) Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. J Biol Chem 259:14441–14447
14. Legewie S, Bluthgen N, Schafer R, Herzel H (2005) Ultrasensitization: switch-like regulation of cellular signaling by transcriptional induction. PLoS Comput Biol 1:e54
15. Kholodenko BN, Hoek JB, Westerhoff HV, Brown GC (1997) Quantification of information transfer via cellular signal transduction pathways. FEBS Lett 414:430–434
16. Qu Z, Vondriska TM (2009) The effects of cascade length, kinetics and feedback loops on biological signal transduction dynamics in a simplified cascade model. Phys Biol 6:016007
17. Feliu E, Knudsen M, Andersen LN, Wiuf C (2011) An algebraic approach to signaling cascade with n layers. Bull Math Biol, DOI 10.1007/s11538-011-9658-0, http://www.springerlink.com/content/9718g720118r9666
18. Gunawardena J (2005) Multisite protein phosphorylation makes a good threshold but can be a poor switch. Proc Natl Acad Sci USA 102:14617–14622
19. Salazar C, Höfer T (2009) Multisite protein phosphorylation – From molecular mechanisms to kinetic models. FEBS J 276:3177–3198
20. Thomson M, Gunawardena J (2009) The rational parameterization theorem for multisite post-translational modification systems. J Theor Biol 261:626–636
21. Heinrich R, Neel BG, Rapoport TA (2002) Mathematical models of protein kinase signal transduction. Mol Cell 9:957–970
22. Markevich NI, Hoek JB, Kholodenko BN (2004) Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J Cell Biol 164:353–359
23. Huang CY, Ferrell JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci USA 93:10078–10083
24. Gunawardena J (2010) Biological systems theory. Science 328:581–582
25. Shinar G, Feinberg M (2010) Structural sources of robustness in biochemical reaction networks. Science 327:1389–1391

Chapter 5
Heterogeneous Biological Network Visualization System: Case Study in Context of Medical Image Data

Erno Lindfors, Jussi Mattila, Peddinti V. Gopalacharyulu, Antti Pesonen, Jyrki Lötjönen, and Matej Orešič

Abstract We have developed a system called megNet for integrating and visualizing heterogeneous biological data in order to enable modeling biological phenomena using a systems approach. Herein we describe megNet, including a recently developed user interface for visualizing biological networks in three dimensions and a web user interface for taking input parameters from the user, and an in-house text mining system that utilizes an existing knowledge base. We demonstrate the software with a case study in which we integrate lipidomics data acquired in-house with interaction data from external databases, and then find novel interactions that could possibly explain our previous associations between biological data and medical images. The flexibility of megNet assures that the tool can be applied in diverse applications, from target discovery in medical applications to metabolic engineering in industrial biotechnology.

Abbreviations

API	Application programming interface
BIND	Biomolecular interaction network database
BioGrid	Biological general repository for interaction datasets
CCA	Curvilinear component analysis
CDA	Curvilinear distance analysis
Cer	Ceramide

E. Lindfors (✉) • P.V. Gopalacharyulu • A. Pesonen • M. Orešič
VTT Technical Research Centre of Finland, Tietotie 2, Espoo, Finland
e-mail: Erno.Lindfors@vtt.fi; gopal.peddinti@vtt.fi; antti.pesonen@vtt.fi; matej.oresic@vtt.fi

J. Mattila • J. Lötjönen
VTT Technical Research Centre of Finland, Sinitaival 6, Tampere, Finland
e-mail: jussi.mattila@vtt.fi; jyrki.lotjonen@vtt.fi

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_5, © Springer Science+Business Media, LLC 2012

DAG	1,2-Diacyl-*sn*-glycerol
DIP	Database of interacting proteins
EMBL	European molecular biology laboratory
EMPath	Enriched molecular path detection
GEO	Gene expression omnibus
GO	Gene ontology
JDBC	Java data base connectivity
JVM	Java virtual machine
LysoPC	Lysophosphocholine
LysoPE	Lysophosphatidylethanolamine
MINT	Molecular interaction database
MR	Magnetic resonance
NML	Sammon's non-linear mapping
OAT	Ontology aided text mining
PC	Phosphatidylcholine
PE	Phosphatidylethanolamine
SIF	Simple identifier format
SM	Sphingomyelin
SOAP	Simple object access protocol
TAG	Triacylglycerol
TEAFS	Topological enrichment analysis for functional subnetworks
TransFac	Database of transcription factors
TransPath	Database of signal transduction pathways
UMLS	Unified medical language system
XML	eXtensible markup language

1 Introduction

We have earlier introduced a software system megNet for integrating and visualizing heterogeneous biological data, with the aim to address the needs of systems biology, integrate data from many sources into a single platform, and model it as holistic biological networks [1, 2]. At the methodological level, this system has addressed the need of evolving ontologies in biology by allowing the user to define a desired biological context by assigning weights to the edges, and map the internal distances of nodes into two dimensions. The prototype of the software is currently installed in our facility at VTT Technical Research Centre of Finland and it has being used by VTT's researchers.

We have recently made several improvements to megNet. Specifically, we have developed two new interfaces in order to improve the usability: a desktop application for visualizing networks in three dimensions and a web application for taking input parameters from the user. This enables several new use cases, for example, text mining from our databases that can be further included in network construction. This application is integrated with Cytoscape [3], a popular biological network

visualization tool. Also, we have developed a text mining system called ontology aided text (OAT) mining system [4] which creates ontologies for biological entities by utilizing an existing knowledge base. The content of this system is represented as an ontological database, and it is integrated as part of our database repository, and its ontological relationships can be visualized in megNet's networks. In parallel, we have recently developed advanced medical image techniques [5] and computational methods to integrate this data with biological data [6].

Herein we describe a conceptual framework and technical architecture that reflects the current status of megNet (Sect. 2). Then we show illustrative examples demonstrating how the biological data from our previous case study [6] can be visualized in megNet, and how they help us find novel associations with the medical image data (Sect. 3). In the end, we discuss the significance of megNet and its future challenges (Sect. 4).

2 Materials and Methods

2.1 Conceptual Framework of megNet

The conceptual framework of megNet is shown in Fig. 5.1. The primary aim of megNet development has been data integration; there is a huge amount of heterogeneous biological data available across diverse databases. This data comprises mainly publicly available data as well as commercial data repositories. We model this data as biological networks in which nodes are either low level molecular entities (e.g., proteins and metabolites) or more complex biological entities and concepts (e.g., diseases and biological processes), and edges are relationships between them (e.g., protein–protein interactions, metabolic reactions, signal transductions, and ontological relationships). We can also extract relationships from OAT [4] for biological entities based on their occurrences as subjects, predicates, or objects in sentences extracted from biological articles.

We can enrich contextual information in the network model, for example, by incorporating gene expression or metabolic profiles that will be manifested as bars inside nodes or as co-expression edges. In the practical examples of this chapter, the context is medical images but more broadly the methods are applicable to any other biological context.

We can access the integrated database repository to construct a complex network of several interaction types following the network presentation model. We can restrict the network to a biological context, for example, including only entities that are involved in a specific biological process. Once the network is constructed, we can browse it to find novel interactions, for example, between different biological processes via multiple interaction types. It is becoming increasingly evident that this kind of cross-talk can lead to new testable hypotheses [7–9].

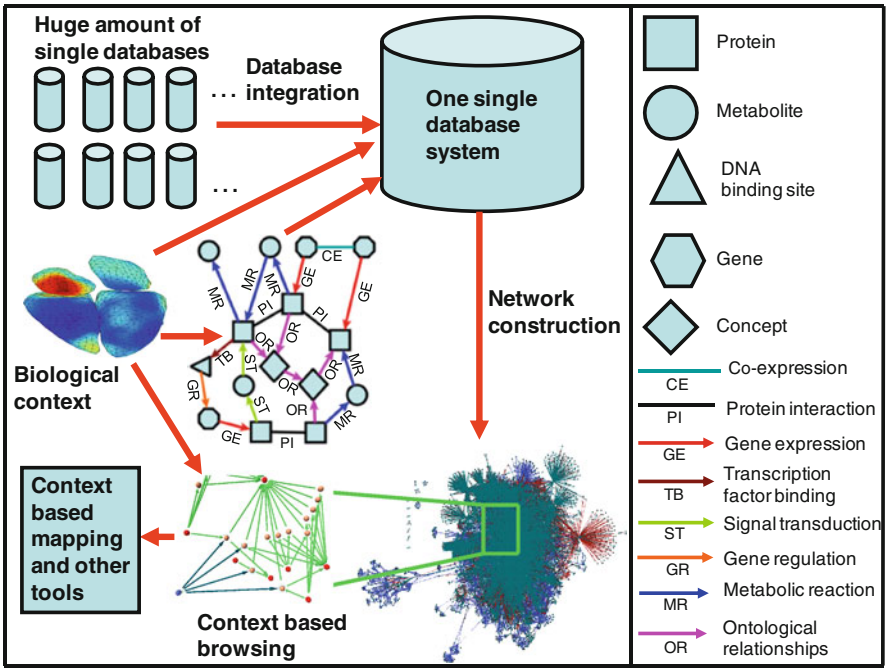


Fig. 5.1 Conceptual framework of megNet

In the end, we can map the network into two dimensions to easily visualize the proximities or similarities among the biological entities in a specific biological context. Also, we can export the network to other advanced computational tools for contextual analysis.

2.2 Technical Architecture

megNet is technically implemented as an architecture as shown in Fig. 5.2. Its main components are database end, middle tier, input client, and network client.

The middle tier includes business logic processing, for example, network construction and text mining. It is implemented in Java programming level by using Java virtual machine (JVM) v.1.6.16 (Oracle, Inc.) and we have been running it on a JBoss application server (JBoss, Inc.), but in general it can be run on any J2EE application server. It uses Tamino Java application programming interface (API) and Oracle Java data base connectivity (JDBC) thin drivers to communicate with the databases, and simple object access protocol (SOAP) messages to communicate with the network and input clients.

megNet has two main user interfaces: input client and network client. The input client is implemented as a web application. It takes all input parameters from the

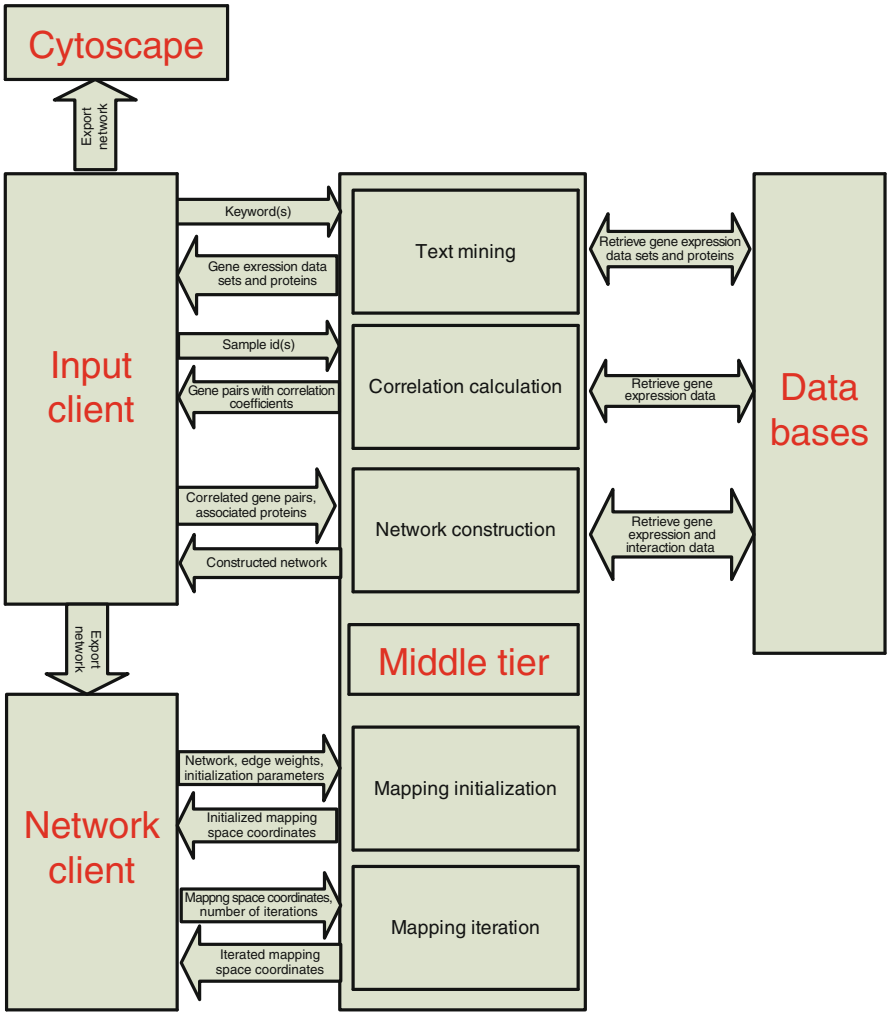


Fig. 5.2 megNet's architecture and features

user for business logic. It is implemented by using Google Web Toolkit (http://code.google.com/intl/fi/webtoolkit/). We have tested that it works in most common web browsers (e.g., Internet Explorer 8.0.6, Mozilla Firefox 3.5.15).

The network client is a desktop application. Its main task is to visualize networks and mapping results. It is a stand-alone Windows (Microsoft) application developed in C# 2.0 by using Microsoft.NET Framework Version 2.0. The three-dimensional visualization is implemented by using Microsoft's DirectX 9.0c platform, which allows hardware acceleration of the three-dimensional scenes. In addition, Cytoscape [3] can be used as an alternative visualization for megNet's network client.

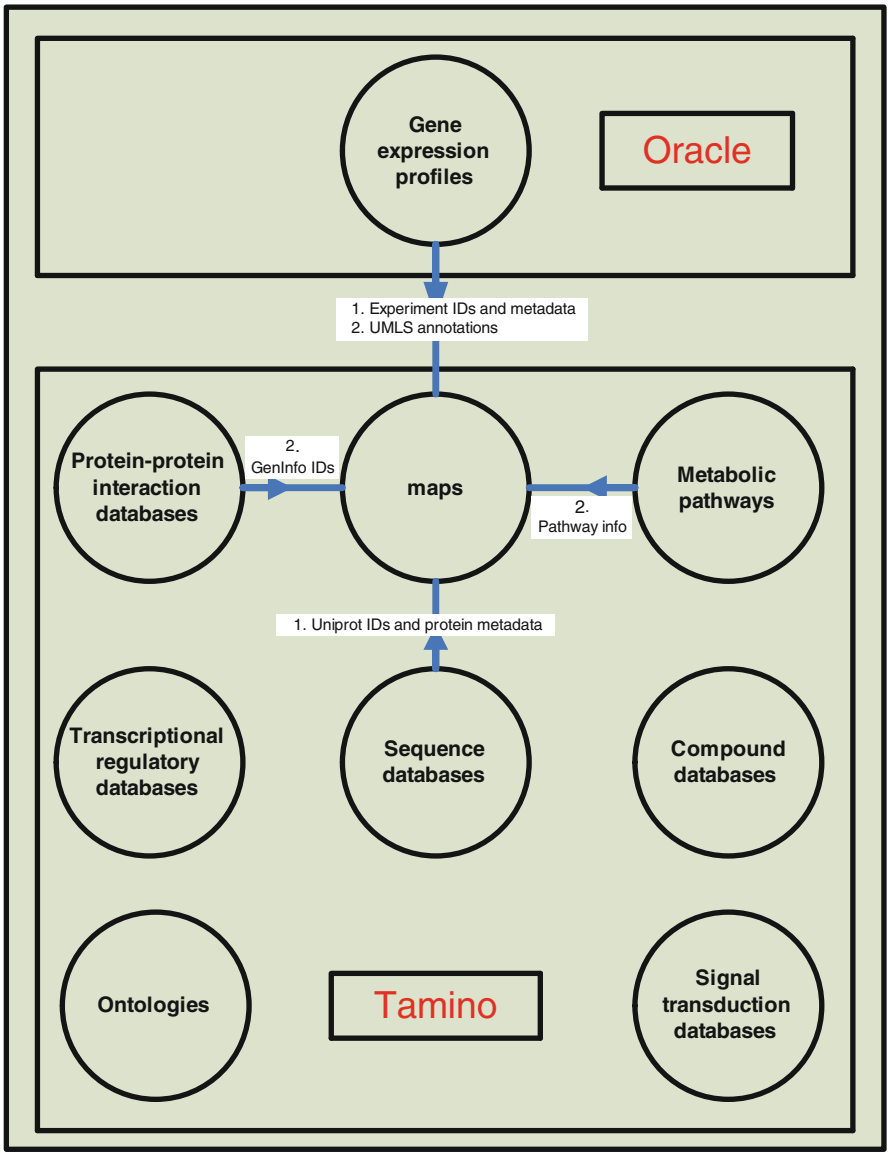


Fig. 5.3 The content of back end

2.3 Databases

megNet's database end comprises all databases to be integrated. The content of these databases is presented in Fig. 5.3. They are physically stored in two databases: Oracle and Tamino.

The Oracle database runs on an Oracle 10g database server (Oracle, Inc.) in which data are stored as relational tables. It comprises the following database:

- Gene expression profiles that are obtained from a public gene expression repository called gene expression omnibus (GEO) [10].

The Tamino database runs on an eXtensible markup language (XML) data management system Tamino XML server (Software AG) in which data are stored in XML format. It comprises the following databases:

- Metabolic pathway databases: Kyoto encyclopedia of genes and genomes (KEGG) [11] and genome-scale yeast metabolic models [12, 13].
- Protein–protein interaction databases: Biological general repository for interaction datasets (BioGrid) [14], database of interacting proteins (DIP) [15], MINT [16], biomolecular interaction network database (BIND) [17].
- Transcriptional regulatory database: database of transcription factors (TransFac) [18].
- Signal transduction database: database of signal transduction pathways (TransPath) [19].
- Compound databases: PubChem [20] and KEGG compounds [11].
- Ontological databases: gene ontology (GO) [21] and OAT [4].
- Sequence databases: universal protein resource (UniProt) [22] and European molecular biology laboratory (EMBL) [23].

In addition, we have developed a database called "maps" in Tamino. This database is based on a premise that in each interaction and pathway database proteins are identified by a unique protein identifier called UniProt identifier [22] and each experiment in the gene expression database is identified by a unique experiment identifier. This database comprises XML documents in such a way that in every document, an experiment or protein identifier is mapped to its metadata (e.g., experiment description and protein name). This enables retrieving effectively data across multiple databases; first we retrieve experiment or protein identifiers for given metadata, and then we use experiment identifiers to retrieve more specific data on the experiments (e.g., experiment description, samples taken in the experiment), or UniProt identifiers [22] to retrieve interactions and reactions in which the proteins are involved. This database can easily be extended to include similar mappings also for other types of entity (e.g., for genes and compounds).

Figure 5.3 describes how the "maps" database is populated. The experiment "maps" are populated by retrieving experiment identifiers and metadata from the gene expression database. Also, these documents comprise unified medical language system (UMLS) annotations [24] that are incorporated by using GENOTEXT [25], which finds annotations for a part of experiment. We can manually incorporate UMLS annotations for the rest of experiments. The protein "maps" are populated by retrieving first UniProt identifiers and most of the metadata from UniProt [22]. Then, the metadata is augmented by retrieving pathway information from the pathway databases and GenInfo identifiers from the BIND database [17].

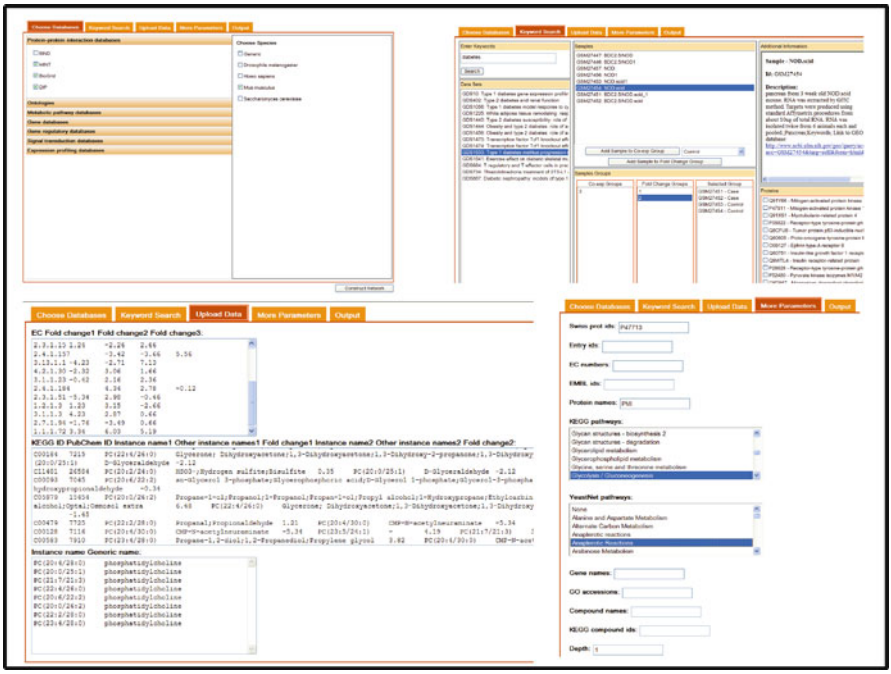


Fig. 5.4 Overall user interface of input client. It comprises four panels in which the user can give input parameters: database and species selection panel (*top-left*), keyword search panel (*top-right*), panel for uploading user's data, and panel for finalizing network construction (*downright*)

2.4 Features

In this section, we describe the basic features of megNet. Figure 5.2 describes how megNet's components interact with each other and with the user when implementing these features.

2.4.1 Text Mining

The purpose of text mining is to help the user find most relevant data from the massive amount of data that we have in megNet's databases; this works like Google in a biological jungle. In the beginning, the user has some biological concept(s) in mind (e.g., diabetes). She types this concept in the "keyword search" tab of megNet's input client (top-right corner of Fig. 5.4). Then the middle tier accesses the GEO database [10] to retrieve all gene expression datasets of which description contains the given keyword, and the "maps" database to retrieve all proteins that are annotated with the given keyword. After that, the input client displays the retrieved gene expression datasets and proteins, as illustrated in Fig. 5.4. Then the user can browse the results in order to assess their relevance.

The gene expression datasets can be of two types: single channel [26] or dual channel [27] microarrays. The single channel datasets contain log2 ratios between case and control intensities (e.g., healthy and disease). The dual channel datasets contain separate values for case and control intensities, so for these datasets we calculate the log2 ratios by normalizing the case and control intensities. More precisely, for each case sample, we calculate the log2 ratio of intensity versus the average intensity of control samples from the same dataset. This enables the user to use both single and dual channel datasets in identical fashion. From these datasets the user can create sample groups for the correlation calculation and network construction. In case of single channel dataset, she has to separately select case and control samples.

2.4.2 Correlation Calculation

The purpose of correlation calculation is to find strongly associated genes or other biological entities in a specific context. As described in the "text mining" section, the user can create sample groups from text mining results. She can select some of these groups for correlation calculation. Also, she sets a cut-off for correlation co-efficient meaning that all correlations of which absolute value is less than this cut-off value will be ignored. Then the middle tier accesses the GEO database [10] to retrieve gene expression data for the selected samples. Based on this data, the middle tier first filters out genes that do not have enough variation between case and control samples by using student's t-test [28]. Then it calculates correlations between the case samples for remaining genes. In the end, a list of the remaining gene pairs along with their correlations is displayed in megNet's input client.

2.4.3 Network Construction

The user can choose from which databases she wants to retrieve data in the "choose databases" tab of megNet's input client (top-left corner of Fig. 5.4). In this tab, all megNet's databases are listed. Also, in this tab the user can choose in which species the network will be constructed. After that there are four basic use cases that the user can use to construct networks. Next we will briefly describe each of them.

- The user can construct a network by giving directly a name or identifier of biological entities (e.g., endothelial lipase) in the "more parameters" tab of megNet's input client (down-right corner of Fig. 5.4). In this case, the middle tier retrieves all interactions in which the given biological entities participate from the selected databases. Also, in this tab she can select metabolic pathways. In this case, the middle tier constructs a network that contains the given metabolic pathways combined with interactions from other selected databases. Also, in this tab, the user can give a depth for the network construction. This means how many nearest neighbors will be retrieved for the given biological entities or pathways. The default depth is one.

- The user can construct a network from the text mining results. In the same way as in the correlation calculation, she can select sample groups for network construction. Then the genes that are in the selected samples will be enriched in the network based on their expression in the sample groups. In megNet's network client, they are visualized as bars inside a gene node so that one bar corresponds to one sample group. Also the user can restrict the network construction with proteins from the text mining results. She can select proteins from the text mining results. After that, the selected proteins are added in the "more parameters" tab of megNet's input client (down-right corner of Fig. 5.4).
- The user can construct a network after performing correlation calculation. After the correlation calculation, the user has a list of most correlated gene pairs. From this list, she can optionally select which pairs she want to include in the network construction. Then the middle tier constructs a network creating co-expression edges for the selected gene pairs and retrieving interactions from other selected databases.
- The user can upload his or her data to the network construction in the "upload data" tab of megNet's input client (down-left corner of Fig. 5.4). This supports two different types of data: gene expression data for enzymes describing how strongly their encoding genes are expressed in given conditions and concentration data for lipid molecular species that are mapped to their generic lipid names on a specific metabolic pathway by using the biochemical knowledge of the side chain length and saturation, as described in [29]. Enzymes are enriched with uploaded gene expression data; in megNet's network client, one bar inside a protein node corresponds to expression of its encoding genes in a specific condition. Compounds are enriched with uploaded lipidomics data; in megNet's network client, one bar inside a compound node corresponds to a concentration of one lipid molecular species.

The user can use these use cases in overlapping manner, which is actually quite common case. For example, she can first make a text mining search and select sample groups for network construction. Then she can calculate correlations between some samples and select some of gene pairs for network construction. Then she can upload his or her data for network construction. Also, she can give more input parameters, for example, she may want to restrict the network construction to a specific metabolic pathway.

2.4.4 Network Visualization

The output of network construction is presented in such formats that the network can be exported to megNet's network client or to Cytoscape [3] for visualization. The megNet network is presented in an XML document for which we have defined an XML schema. Briefly, this schema comprises a node element for each node containing its unique identifier and metadata, and an edge element for each edge containing its identifier, metadata, and the identifiers of connected nodes. The

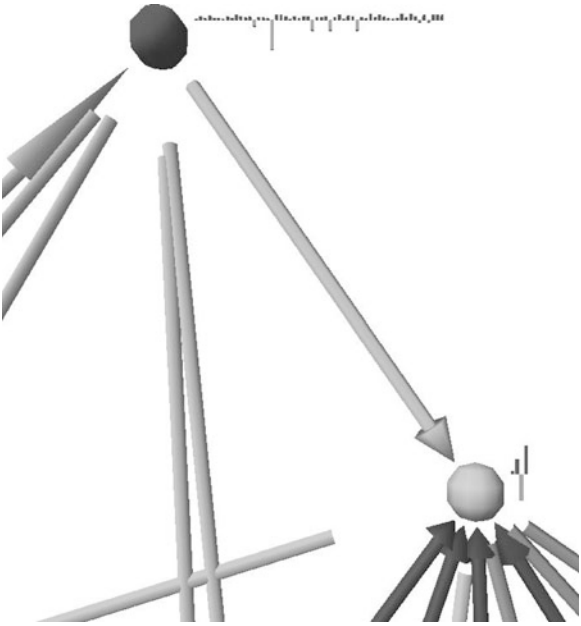


Fig. 5.5 Bar visualization in network client. The upper node is a compound and the lower node is a protein. One bar inside the compound node corresponds to a concentration of a lipid molecular species mapped to the compound. One bar inside a protein node corresponds to expression of the encoding genes in a specific condition. The up-pointing bar means that concentration or gene expression in case is higher than in control group, and the down-pointing bar means the opposite case

Cytoscape network is presented in simple identifier format (SIF). Briefly, this format is a flat file format in which one row corresponds to one edge; it comprises identifiers of connected nodes and interaction types. Also, there are separate flat files for edge and node attributes (e.g., colors and shapes). In megNet's input client, these outputs are presented in text boxes, so that the XML document for megNet's network client is in one text box, the SIF format for Cytoscape is in another text box, and each attribute type for Cytoscape is in a separate text box. In order to visualize the network in Cytoscape or in megNet's network client, the user should copy-paste these outputs to text files. And then in megNet's network client she should import the XML document. Or in Cytoscape first import the SIF format, and then import each edge and node type separately.

The idea is that the user can visualize networks in Cytoscape and megNet's network client in a complementary manner; in some aspects, megNet's network client outweighs Cytoscape and vice versa. Most obviously, in megNet's network client, we use third dimension for some features which enables elegant visualization, whereas Cytoscape is a large open source community effort which enables continuously growing amount of new features augmented with many useful plugins. One very useful feature that the third dimension brings to the network client is that it is possible to visualize bars inside nodes. As described in the "Network construction" section, the user can enrich gene expression data to genes and enzymes, and lipidomics data to compounds. This data is manifested as bars inside gene, protein, and compound nodes as illustrated in Fig. 5.5.

2.4.5 Context-Based Mapping

The purpose of the context based mapping is to enable investigating how the biological entities are related to each other in a specific biological context. We have implemented three different mapping methods: Sammon's non-linear mapping (NLM) [30], curvilinear component analysis (CCA) [31], and curvilinear distance analysis (CDA) [32]. We have described these methods in detail in our previous publication [2]. Briefly, the idea is that we non-linearly map the internal distances of nodes into two dimensions. All of these methods try to minimize iteratively discrepancy between the original high-dimensional distance space and two-dimensional mapping space. In NLM, the mapping is calculated based on a steepest gradient descent, whereas in CCA and CDA it is based on a stochastic gradient descent. In the network client, the user can assign weights to the edges in an appropriate way. For example, if she is interested in a specific biological process, she can assign low weights to edges that are close to the corresponding GO concept node [21]. Then the middle tier initializes the mapping by calculating the internal distances of nodes based on the assigned weights, and returns an initialized mapping as two-dimensional coordinates along with the mapping discrepancy to the network client. The network client visualizes the initialized mapping in the user interface. After that the user can send an iteration request to the middle tier, and then the middle tier iterates the mapping, and sends new mapping coordinates along with the discrepancy to the network client. The user can keep iterating the mapping as long as she feels that the mapping discrepancy is small enough.

3 Results

In this section, we show how megNet can be used to make novel findings by studying biological networks in context of medical image data from Lamin A/C mutation patients. In Sect. 3.1, we describe the biological and medical image data that we use in this case study. In Sect. 3.2, we show two examples that demonstrate how we can find associations between biological network and medical images via multilevel cross-talk. In Sect. 3.3, we perform a context based mapping to show how biological entities are related to each other in context of medical images.

3.1 Biological and Medical Image Data for Lamin A/C Case Study

We have previously derived magnetic resonance (MR) image parameters from Lamin A/C mutation patients [5]. In a follow-up study, we performed lipidomics analysis in the same patient, and developed a statistical model to find associations between the lipidomics profiles and medical image parameters [6]. In order to study

how these associations are manifested in biological networks, in this chapter we use megNet to construct biological networks in context of the lipidomics profiles.

More specifically, we first mapped lipid molecular species to their generic lipid names on *glycerophospho-, glycero-*, and *sphingolipid* metabolic pathways from KEGG [11]. Exact mappings are presented in Tables 5.1–5.3. Then we uploaded this data into megNet's input client, and chose all other databases in which these pathways are involved (OAT [4], BioGrid [14], MINT [16], DIP [15], GO [21], and EMBL [23]) for network construction in human. As a result, we thus obtained a network in which these pathways are integrated with interactions from these databases (Fig. 5.6). In this network, there are bars inside compound nodes, so that one bar represents fold change between concentrations of Lamin A/C mutation carriers and their non-mutated controls in one lipid molecular species.

3.2 Multilevel Cross-talk Examples

We can see from Fig. 5.6 that interestingly between many metabolic reactions there is quite dense cross-talk via many interaction levels. An interesting cross-talk example is visualized in Fig. 5.7. In this figure, *arachidonate 12-lipoxygenase* interacts with two isoforms of *phospholipase A2* [33]. One of these isoforms catalyzes a metabolic reaction in which *1-acyl-sn-glycero-3-phosphocholine* is a product, and the other isoform catalyzes a reaction in which *phosphatidylethanolamine* (PE) is a substrate. Many *lysophosphocholine* lipid molecular species (LysoPCs) are mapped to the former lipid, and many *phosphatidylethanolamine* lipid molecular species (PEs) to the latter one. In our previous case study [6], PEs were correlated quite strongly with image parameters, whereas there was a LysoPC in which the correlation was not so obvious. Perhaps the *arachidonate 12-lipoxygenase* has some role in these correlations, for example, it may via signaling regulate activities of the phospholipases. And interestingly there is some evidence that lipoxygenases have important roles in cardiovascular diseases [34].

Another interesting cross-talk example is visualized in Fig. 5.8. This figure comprises glycerolipid metabolism in which two isoforms of *endothelial lipase* break down *1,2-diacyl-sn-glycerol* (DAG) and *triacylglycerol* (TAG) into free fatty acids. Both of these lipases are involved in *cholesterol transport and homeostasis* biological processes. This is interesting since in our previous case study [6] triglyceride lipid molecular species (TGs) were associated with increased end-diastolic wall thickness. This may be a sign that cholesterol metabolism is associated with the increased end-diastolic wall thickness via the TG. Also, interestingly according to the OAT text mining system [4], the endothelial lipases are associated with diabetes prevention [35] and maintenance of cell homoeostasis [36] in type 1 diabetes mouse models. This may be a sign that the end-diastolic wall thickness prevents type 1 diabetes and it may have an important role in the maintenance of cell homeostasis in diabetes development.

Table 5.1 Mapping lipid molecular species (the second column) to generic lipids (the first column) on glycerophospholipid metabolic pathway

C00157 Phosphatidylcholine	PC(34:5), PC(36:6), PC(38:7), PC(36:5), PC(28:0), PC(40:8), PC(30:1), PC(36:7), PC(36:7), PC(42:9), PC(32:2) (sodiated), PC(32:2), PC(38:6) (sodiated), PC(38:6), PC(36:5), PC(38:8), PC(40:7), PC(36:4), PC(36:4)(sodiated), PC(34:3), PC(sodiated), PC(38:6)(sodiated), PC(38:6), PC(40:7), PC(36:6), PC(40:6), PC(36:4), PC(38:7), PC(38:5), PC(40:8), PC(34:3), PC(32:1), PC(32:1) (sodiated), PC(36:3), PC(40:5), PC(34:2), PC(34:3), PC(36:5), PC(38:5), PC(40:8), PC(36:3), PC(36:3) (sodiated), PC(38:7), PC(40:6), PC(42:9), PC(40:5), PC(38:4) (sodiated), PC(38:4), PC(40:8), PC(30:1), PC(42:8), PC(32:3), PC(38:3) (sodiated), PC(38:3), PC(38:4) (sodiated), PC(38:4), PC(34:3), PC(36:4), GPCho(32:0), PC(34:1), PC(36:2), PC(36:2) (sodiated), PC(38:6), PC(40:5), PC(40:7), PC(34:3), PC(38:3) (sodiated), PC(38:3), PC(38:6) (sodiated), PC(38:6), PC(34:2) (sodiated), PC(34:2), PC(40:7), PC(32:0), PC(36:1), PC(34:2), PC(38:2) (sodiated), PC(38:2), PC(40:4), PC(32:0), PC(34:2), PC(34:2)(sodiated), PC(36:0), PC(38:0), PC(34:0)
C00350 Phosphatidylethanolamine	PE(34:1), PE(34:0), PE(34:0), PE(36:0), PE(36:1), PE(36:2), PE(36:3), PE(36:3), PE(36:3), PE(36:4), PE(36:4), PE(38:0), PE(38:0), PE(38:1), PE(38:2), PE(38:3), PE(38:3), PE(38:4), PE(40:1), PE(40:2), PE(40:3), PE(40:4), PE(40:5), PE(40:6), PE(42:1), PE(42:8), PE(42:9), PE(44:10), PE(46:5), PE(44:11), PE(44:5), PE(44:7), PE(48:10), PE(48:8), PE(48:9), PE(48:9)

(continued)

Table 5.1 (continued)

C00641 1,2-Diacylglycerol	DAG(34:6), DAG(36:2), DAG(36:7), DAG(40:8)
C04230 1-Acyl-*sn*-glycero-3-phosphocholine	LysoPC(20:5), LysoPC(16:1), LysoPC(22:6), LysoPC(16:1), LysoPC(20:4), LysoPC(18:2), LysoPC(18:2) (sodiated), LysoPC(22:6), LysoPC(20:4), LysoPC(18:2), LysoPC(18:2) (sodiated), LysoPC(18:3), LysoPC(16:0), LysoPC(20:3), LysoPC(16:0), LysoPC(16:0) (sodiated), LysoPC(18:1), LysoPC(18:1) (sodiated), LysoPC(20:3), LysoPC(18:1), LysoPC(20:4) LysoPC(18:1), LysoPC(18:0), LysoPC(18:0) (sodiated), LysoPC(18:0) (sodiated), LysoPC(18:0), LysoPC(20:1), LysoPC(18:0)
C04233 2-Acyl-*sn*-glycero-3-phosphocholine	LysoPC(20:5), LysoPC(16:1), LysoPC(22:6), LysoPC(16:1), LysoPC(20:4), LysoPC(18:2), LysoPC(18:2) (sodiated), LysoPC(22:6), LysoPC(20:4), LysoPC(18:2), LysoPC(18:2) (sodiated), LysoPC(18:3), LysoPC(16:0), LysoPC(20:3), LysoPC(16:0), LysoPC(16:0) (sodiated), LysoPC(18:1), LysoPC(18:1) (sodiated), LysoPC(20:3), LysoPC(18:1), LysoPC(20:4) LysoPC(18:1), LysoPC(18:0), LysoPC(18:0) (sodiated), LysoPC(18:0) (sodiated), LysoPC(18:0), LysoPC(20:1), LysoPC(18:0)
C04438 1-Acyl-*sn*-glycero-3-phosphoethanolamine	LysoPE(18:0), LysoPE(18:2), LysoPE(20:1), LysoPE(22:0), LysoPE(22:3), LysoPE(22:3)
C05973 2-Acyl-*sn*-glycero-3-phosphoethanolamine	LysoPE(18:0), LysoPE(18:2), LysoPE(20:1), LysoPE(22:0), LysoPE(22:3), LysoPE(22:3)

Table 5.2 Mapping lipid molecular species (the second column) to generic lipids (the first column) on glycerolipid metabolic pathway

C00641 1,2-Diacylglycerol	DAG(34:6), DAG(36:2), DAG(36:7), DAG(40:8)
C00422 Triacylglycerol	TAG(33:0), TAG(33:0), TAG(36:0), TAG(38:0), TAG(40:0), TAG(44:0), TAG(44:1), TAG(44:2), TAG(46:0), TAG(46:1), TAG(46:2), TAG(47:0), TAG(47:1), TAG(47:2), TAG(48:0), TAG(48:1), TAG(48:1), TAG(48:2), TAG(48:3), TAG(48:4), TAG(48:4), TAG(48:4), TAG(48:5), TAG(48:5), TAG(49:0), TAG(49:1), TAG(49:2), TAG(49:3), TAG(49:4), TAG(49:7), TAG(49:9), TAG(50:0), TAG(50:1), TAG(50:2) TAG(50:3), TAG(50:3), TAG(50:4), TAG(50:5), TAG(50:9), TAG(51:1), TAG(51:2), TAG(51:2), TAG(51:3) TAG(51:4), TAG(52:0), TAG(52:1), TAG(52:2), TAG(52:3), TAG(52:4), TAG(52:5), TAG(52:5), TAG(52:6), TAG(52:6), TAG(52:7), TAG(53:10), TAG(53:2), TAG(53:3), TAG(53:4), TAG(53:4), TAG(53:5), TAG(53:5), TAG(53:6), TAG(53:6), TAG(53:7), TAG(53:8), TAG(54:1), TAG(54:2), TAG(54:3), TAG(54:3), TAG(54:4), TAG(54:4), TAG(54:5), TAG(54:5), TAG(54:5), TAG(54:6), TAG(54:6), TAG(54:7), TAG(54:7), TAG(54:8), TAG(54:8), TAG(54:8), TAG(55:3), TAG(55:4), TAG(55:5), TAG(55:6), TAG(56:10),

(continued)

Table 5.2 (continued)

	TAG(56:2), TAG(56:2), TAG(56:3), TAG(56:3), TAG(56:4), TAG(56:4), TAG(56:4), TAG(56:5), TAG(56:5), TAG(56:5), TAG(56:6), TAG(56:6), TAG(56:6), TAG(56:6), TAG(56:7), TAG(56:7), TAG(56:8), TAG(56:8), TAG(56:9), TAG(57:10), TAG(57:11), TAG(57:8), TAG(58:10) TAG(58:13), TAG(58:3), TAG(58:5), TAG(58:5), TAG(58:5), TAG(58:6), TAG(58:6), TAG(58:8), TAG(58:8), TAG(58:9), TAG(59:12), TAG(60:10), TAG(60:11)

3.3 Context Based Mapping Example

In the previous section, we made a tentative observation that cholesterol metabolism may explain why TGs are associated with increased end-diastolic wall thickness. In order to gain our understanding of the role of cholesterol metabolism in this context, we performed a mapping in context of cholesterol metabolism. More specifically, we assigned low weights (=0.01) to the incident edges of the biological processes in which the phospholipases were involved in the previous section; *cholesterol homeostasis* (GO:0042632), *reserve cholesterol transport* (GO:0043691), *positive regulation of cholesterol transport* (GO:0032376), and we assigned one as weight to the other edges. Then we performed the CDA mapping [32] using 100 iterations. In the mapping results, we took a zoom from the neighborhood of TG (Fig. 5.9). We can see that for example an *ethanolamine kinase 1* and a *receptor signaling* biological process are in this figure. Maybe this is a sign that there are some receptor signaling cascades that stimulate the TG to participate in cholesterol metabolism and in turn associate it with the increased end-diastolic wall thickness. Also, interestingly these entities are close to a *neuron differentiation* biological process, so the stimulating signaling could be neurological. Another interesting observation is that the regulation of macrophage activation is quite close to the TG. Interestingly there has been discussion that macrophages may play critical role in the pathogenesis of type 1 diabetes [37]. Maybe this is related to the observation in the previous section stating that the end-diastolic wall thickness might prevent type 1 diabetes.

Table 5.3 Mapping lipid molecular species (the second column) to generic lipids (the first column) on sphingolipid metabolic pathway

C00195 *N*-Acylsphingosine	Cer(d18:1/22:0), Cer(d18:1/22:1), Cer(d18:1/23:0), Cer(d18:1/24:1)
C00550 Sphingomyelin	SM(d18:1/14:0), SM(d18:1/16:1) (sodiated), SM(d18:1/16:1), SM(d18:1/16:1), SM(d18:1/15:0), SM(d18:1/17:1), SM(d18:1/16:0), SM(d18:1/16:0) (sodiated), SM(d18:1/18:1) (sodiated), SM(d18:1/18:1), SM(d18:0/16:0), SM(d18:1/18:2), SM(d18:1/16:0), SM(d18:1/18:0) (sodiated), SM(d18:1/18:0), SM(d18:0/22:5), SM(d18:0/18:0), SM(d18:1/18:4), SM(d18:1/18:0), SM(d18:1/24:4), SM(d18:1/18:0), SM(d18:1/22:5). SM(d18:1/23:1). SM(d18:1/21:0), SM(d18:1/21:0) (sodiated), SM(d18:1/11:0), SM(d18:1/22:1), SM(d18:1/23:3), SM(d18:1/24:1), SM(d18:1/24:2), SM(d18:1/22:0), SM(d18:1/22:0) (sodiated), SM(d18:1/21:0), SM(d18:1/23:1), SM(d18:1/24:0) SM(d18:1/25:1), SM(d18:1/23:0), SM(d18:1/23:1)

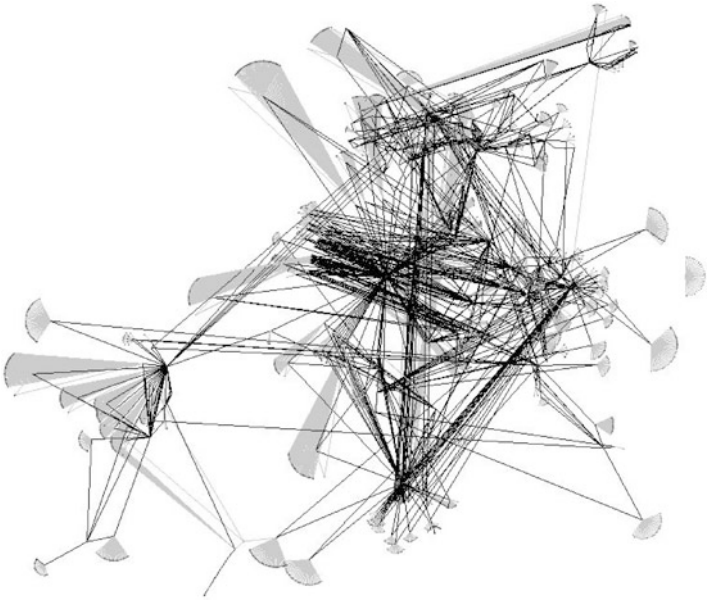


Fig. 5.6 Lipid molecular species metabolic pathway network integrated with other types of interactions. The *dark edges* represent metabolic pathways. The *light edges* represent other types of interactions that make cross-talk between metabolic reactions

4 Discussion

In this chapter, we described our system for integrating and visualizing heterogeneous biological data reflecting to its current status. We showed its practical utility in the context of medical images. We first integrated lipidomics data from our laboratory into a biological network constructed from many data sources, including relationships from our own text mining system. From this network, we showed two interesting examples in which cross-talk via multi-level interaction types could explain associations between lipidomics and medical image data. Also, we showed a context based mapping example in which we studied how biological entities are related to each other in medical context leading to interesting observations. We believe these examples show that our system has potentiality for making novel medical findings in biological network level, though it is good to keep in mind that the findings made in this chapter are very preliminary and they naturally require more validation.

In parallel, we have developed a fingerprint analysis tool [5] that finds statistical differences between two patients groups (e.g., disease versus healthy) in MR image and biological measurements (i.e., gene expression or metabolic profiles). Currently this tool is implemented as a separate web application, so it is not directly integrated with megNet. However, the user can handle data in integrative manner, since some

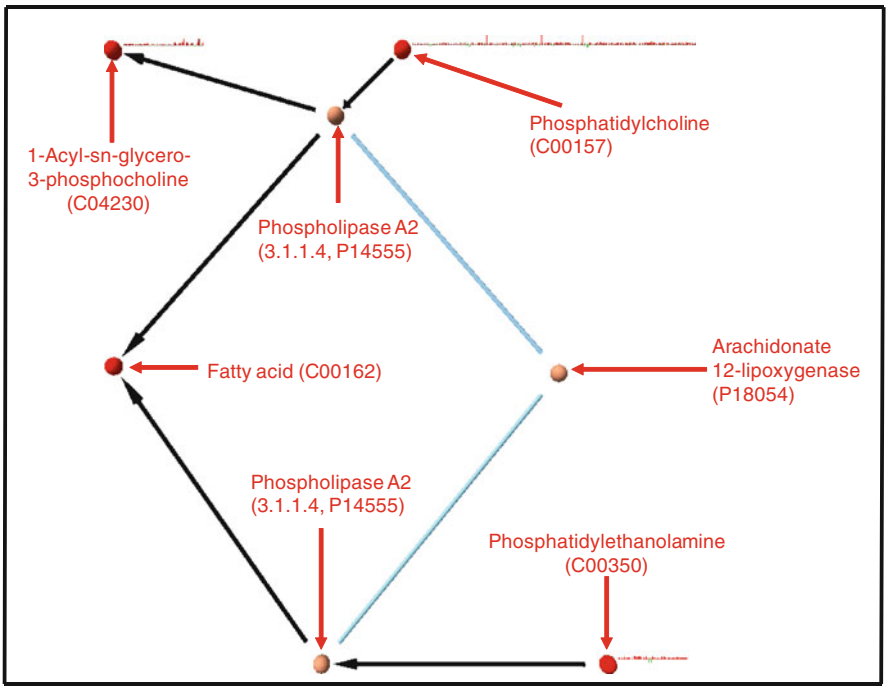


Fig. 5.7 Cross-talk between metabolic reactions. The *dark lines* represent metabolic connection between metabolites and enzymes. The *light lines* represent protein–protein interactions. In *brackets* there are unique identifiers of biological entities; KEGG compound identifiers [11] for metabolites, and UniProt identifiers [22] and EC (Enzyme Commission) numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) for proteins

gene expression profiles and images are annotated with common identifiers called unified medical language system (UMLS) annotations [24]. Also, she can integrate non-annotated data in a heuristic way by using the keyword search in megNet's input client. Also, in the future we may enhance the integration, for example, by creating hyperlinks between megNet's input client and the fingerprint analysis tool.

In addition, megNet is scalable and may incorporate new databases. For example, we have incorporated metabolic profiles from our laboratory into databases, and we have large clinical phenotype data repositories, for example, from cardiovascular diseases, diabetes, and nutritional intervention studies. Our plan is to create a model that enables using this data as part of megNet.

At the moment, megNet is not publicly available, since Tamino's and Oracle's database licenses that we have do not allow unlimited number of users. However, we are considering making parts of megNet publicly available. For example, network construction and mapping methods from megNet's middle tier could be implemented as an open source Cytoscape plugin, so they would be freely available for the systems biology community.

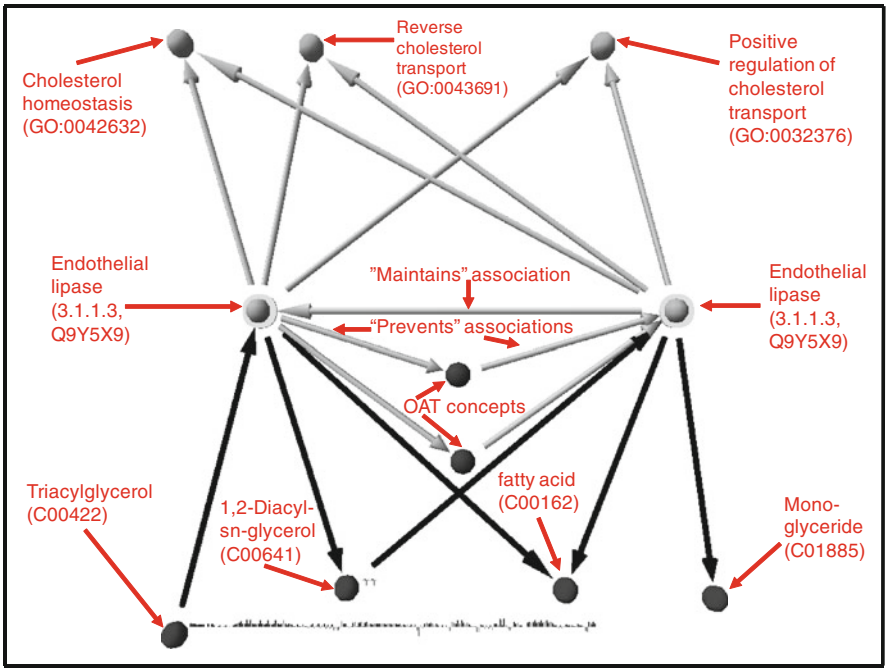


Fig. 5.8 Cross-talk between endothelial lipases. The *light edges* are metabolic reactions that the endothelial lipases catalyze. The *dark edges* are OAT text mining associations and GO biological process relationships. The *light edges* are GO biological relationships [21] and OAT text mining associations [4]. In *brackets* there are unique identifiers of biological entities; GO terms for GO concepts [21], KEGG compound identifiers [11] for metabolites, and UniProt identifiers [22] and EC numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) for proteins

The researchers have used megNet as part of practical biological applications. As a drug target discovery example, megNet was used to construct an integrated metabolic, protein–protein interaction and signal transduction network in non-obese diabetic mouse [38]. Then, the enriched molecular path detection method (EMPath) was used to detect type 1 diabetes specific paths in this network. The results were very interesting in terms of medical biology; ether phospholipid biosynthesis was down-regulated in pre-state of type 1 diabetes, which was consistent with recent findings in clinical level. As an industrial biotechnology example, megNet was applied in a case study in which dynamical topology of modules was studied in an integrated yeast network [39]. MegNet was first applied to construct an integrated metabolic, protein–protein interaction and transcriptional regulatory network in yeast. Then, the topological enrichment analysis of function subnetworks (TEAFS) method was used to rank modules of the integrated network based on their topological measures under time course of a gene expression dataset from oxidative stress. The modules related to the biosynthesis of toxic lipids were found to be modulated during this time course. These results were further validated by

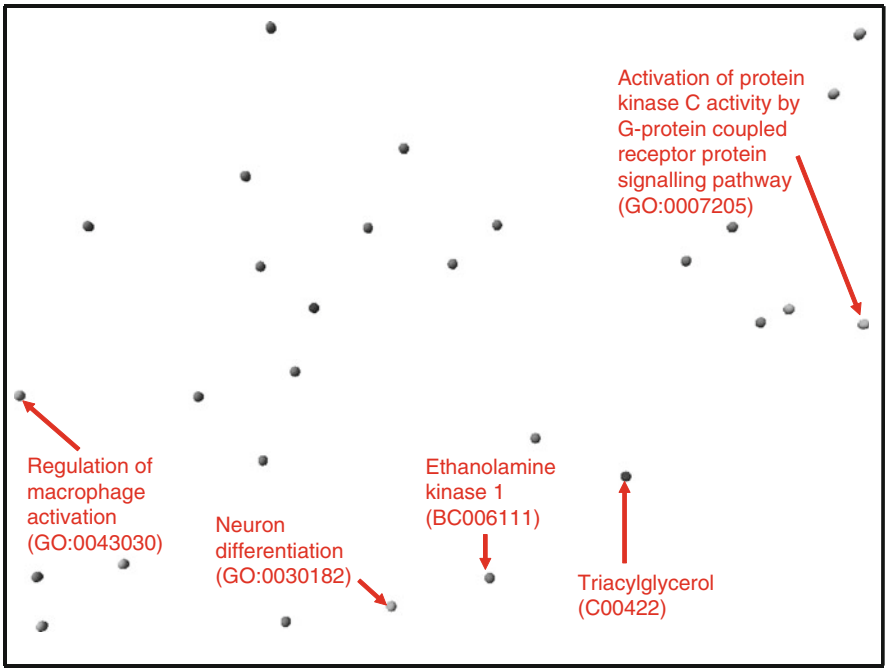


Fig. 5.9 Mapping results in context of cholesterol metabolism. A zoom from the neighborhood of triacylglycerol

metabolomic analysis by showing that the toxic lipids were accumulated during the time course. These examples indicate that megNet has potential to be used in diverse types of biological applications.

Acknowledgments We thank Dr. Laxman Yetukuri for technical assistance in mapping lipidomics data to metabolic pathways. The project was supported by the research program "White Biotechnology – Green Chemistry" (Academy of Finland; Finnish Centre of Excellence programme, 2008–2013, Decision number 118573), by the EU project MITIN (HEALTH-F4–2008–223450), by the National Graduate School in Informational and Structural Biology (ISB), and by the TRANSCENDO project of the Tekes MASI Program.

References

1. Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, Hollmén J, Oresic M (2005) Data integration and visualization system for enabling conceptual biology. Bioinformatics 21:i177–i185
2. Gopalacharyulu PV, Lindfors E, Miettinen J, Bounsaythip CK, Oresic M (2008) An integrative approach for biological data mining and visualisation. Int J Data Min Bioinform 2(1):54–77
3. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B et al (2007) Integration of biological networks and gene expression data using cytoscape. Nat Protocols 2(10):2366–2382

4. Timonen M, Pesonen A (2008) Combining context and existing knowledge when recognizing biological entities – early results. Adv Knowl Discov Data Min 5012:1028–1034
5. Koikkalainen JR, Antila M, Lotjonen JMP, Helio T, Lauerma K, Kivisto SM, Sipola P, Kaartinen MA, Karkkainen STJ, Reissell E et al (2008) Early familial dilated cardiomyopathy: identification with determination of disease state parameter from cine MR image data 10.1148/radiol.2491071584. Radiology 249(1):88–96
6. Sysi-Aho M, Koikkalainen J, Seppänen-Laakso T, Kaartinen M, Kuusisto J, Peuhkurinen K, Kärkkäinen S, Antila M, Lauerma K, Reissell E et al (2011) Serum lipidomics meets cardiac magnetic resonance imaging: profiling of subjects at risk of dilated cardiomyopathy. PLoS ONE 6(1):e15744
7. Papin JA, Palsson BO (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. J Theor Biol 227(2):283–297
8. Min Lee J, Gianchandani EP, Eddy JA, Papin JA (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Comput Biol 4(5):e1000086
9. Li X, Gianoulis TA, Yip KY, Gerstein M, Snyder M (2010) Extensive in vivo metabolite–protein interactions revealed by large-scale systematic analyses. Cell 143(4):639–650
10. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al (2009) NCBI GEO: archive for high-throughput functional genomic data 10.1093/nar/gkn764. Nucl Acids Res 37(suppl_1):D885–890
11. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T et al (2008) KEGG for linking genomes to life and the environment. Nucl Acids Res 36(suppl_1):D480–484
12. Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Buthgen N, Borger S, Costenoble R, Heinemann M et al (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26(10):1155–1160
13. Dobson P, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn W, Fisher P et al (2010) Further developments towards a genome-scale metabolic model of yeast. BMC Syst Biol 4(1):145
14. Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hon G, Myers C, Parsons A, Friesen H, Oughtred R, Tong A et al (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol 5(4):11
15. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucl Acids Res 32(suppl_1):D449–451
16. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the molecular interaction database. Nucl Acids Res 35(suppl_1):D572–574
17. Bader GD, Betel D, Hogue CWV (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31:248–250
18. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV et al (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucl Acids Res 31(1):374–378. doi:10.1093/nar/gkg108
19. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C et al (2006) TRANSPATH(R): an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucl Acids Res 34(suppl_1):D546–551
20. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) Pubchem: a public information system for analyzing bioactivities of small molecules. Nucl Acids Res 37(suppl_2):W623–633. doi:10.1093/nar/gkp456
21. The Gene Ontology Consortium (2008) The gene ontology project in 2008. Nucl Acids Res 36(suppl_1):D440–444. doi:10.1093/nar/gkm883
22. Consortium TU (2010) The universal protein resource (uniprot) in 2010. Nucl Acids Res 38(suppl_1):D142–148
23. Cochrane GR, Galperin MY (2010) The 2010 nucleic acids research database issue and online database collection: a community of data resources. Nucl Acids Res 38(suppl_1):D1–4. doi:10.1093/nar/gkp1077

24. Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucl Acids Res 32:D267–D270
25. Butte AJ, Kohane IS (2006) Creation and implications of a phenome–genome network. Nat Biotechnol 24(1):55–62
26. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Norton H et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14(13):1675–1680
27. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 270:(5235):467–470. doi:10.1126/science.270.5235.467
28. Box JF (1987) Guinness, Gosset, Fisher, and small samples. Stat Sci 2(1):45–52
29. Yetukuri L, Katajamaa M, Medina-Gomez G, Seppanen-Laakso T, Vidal-Puig A, Oresic M (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. BMC Syst Biol 1(1):12
30. Sammon JWJ (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comp C-18(5):401–409
31. Demartines P, Hérault J (1997) Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. IEEE Trans Neur Netw 8:148–154
32. Lee JA, Lendasse A, Verleysen M (2004) Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis. Neurocomputing 57:49–76
33. Coffey MJ, Coles B, Locke M, Bermudez-Fajardo A, Williams PC, Jarvis GE, O'Donnell VB (2004) Interactions of 12-lipoxygenase with phospholipase A2 isoforms following platelet activation through the glycoprotein VI collagen receptor. FEBS Lett 576(1):165–168
34. Zhao L, Funk CD (2004) Lipoxygenase pathways in atherogenesis. Trends Cardiovasc Med 14(5):191–195
35. Mizuno M, Masumura M, Tomi C, Chiba A, Oki S, Yamamura T, Miyake S (2004) Synthetic glycolipid OCH prevents insulitis and diabetes in NOD mice. J Autoimmun 23(4):293–300
36. Mi Q-S, Ly D, Zucker P, McGarry M, Delovitch TL (2004) Interleukin-4 but not interleukin-10 protects against spontaneous and recurrent Type 1 diabetes by activated CD1D-restricted invariant natural killer T-cells. Diabetes 53(5):1303–1310. doi:10.2337/diabetes.53.5.1303
37. Yang L-J (2008) Big mac attack: does it play a direct role for monocytes/macrophages in Type 1 diabetes? Diabetes 57(11):2922–2923. doi:10.2337/db08–1007
38. Lindfors E, Gopalacharyulu PV, Halperin E, Orešic M (2009) Detection of molecular paths associated with insulitis and Type 1 diabetes in non-obese diabetic mouse. PLoS ONE 4(10):e7323
39. Gopalacharyulu P, Velagapudi V, Lindfors E, Halperin E, Orešic M (2009) Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast. Mol BioSyst 5:276–287

Chapter 6
Evolution of the Cognitive Proteome: From Static to Dynamic Network Models

J. Douglas Armstrong and Oksana Sorokina

Abstract Integrative analysis of the neuronal synapse proteome has uncovered an evolutionarily conserved signalling complex that underpins the cognitive capabilities of the brain. Highly dynamic, cell type specific and intricately regulated, the synaptic proteome presents many challenges to systems biology approaches, yet this is likely to be the best route to unlock a new generation of neuroscience research and CNS drug development that society so urgently demands. Most systems biology approaches today have focussed on exploiting protein–protein interaction data to their fullest extent within static interaction models. These have revealed structure–function relationships within the protein network, uncovered new candidate genes for genetic studies and drug research and development and finally provided a means to study the evolution of the system. The rapid maturation of medium and high-throughput biochemical technologies means that dissecting the synapse proteome's dynamic complexity is fast becoming a reality. Here we look at these new challenges and explore rule-based modelling as a basis for a new generation of synaptic models.

1 Introduction

Brains vary widely in their complexity from the simplest of organisms having a few hundreds or thousands of interconnected cells to the massively complex human brain with an estimated 10^{12} neurons with some 10^{15} connections between them [30]. Systems analysis of brains requires researchers to consider the biology at many different levels from molecular signalling complexes through to the networks of neuronal connections both within the brain and beyond with external sensory

J.D. Armstrong (✉) • O. Sorokina
School of Informatics, University of Edinburgh, Edinburgh, UK
e-mail: douglas.armstrong@ed.ac.uk; oksana.sorokina@ed.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_6,

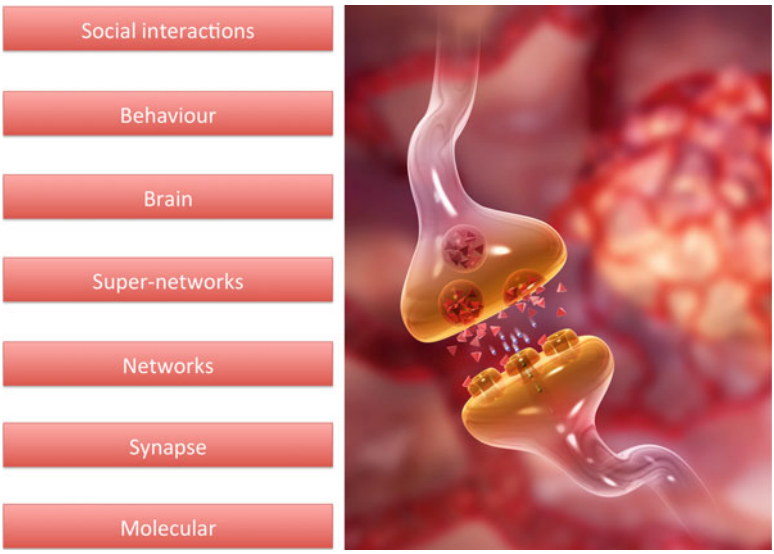


Fig. 6.1 Multiple levels of complexity in the nervous system. Systems biology has to handle multiple levels (*left*) of spatio-temporal complexity in the nervous system from molecular to behaviour. While we focus on the molecular events at the post-synaptic density, we must always bear in mind the other molecular signalling events pre-synaptically in the adjoining neurons and elsewhere in the cell (*right*). *Right panel image*: "synaptic junction" is reproduced with kind permission of Gary Carlson at gcarlson.com

and motor systems. Ultimately, a systems level approach needs to consider the phenotype, or behaviour, of the animal, which in many cases occurs within a social context (Fig. 6.1).

Although all these levels of organisation have a major role in brain function, it is the molecular level at which there is the closest correlation to health. Both genetic evidence and pharmacological basis of existing treatments provide strong evidence for a general molecular basis for almost all human neurological disorders (although environmental factors are also extremely important). However, the vast majority of human brain disorders are very complex and the genetic association data points to multiple molecular targets and pathways in most disorders [20]. This is backed up by evidence that many of the most effective CNS drugs are actually fairly promiscuous in their target specificity [3]. The scale of this problem cannot be underestimated: take for example neurodegenerative diseases [54]. These disorders are seriously disabling, chronic conditions that devastate individuals, their families and have become an ever increasing burden on all societies. Combined neurological disorders have accounted for the largest single cost of healthcare budgets in the developed world for many years. When combined with secondary care costs, the burden to the EU alone soared to an astronomical €160 billion in 2008 and is estimated in excess of $600 billion globally per year. For many neurodegenerative diseases, the molecular complexity is a problem but it is often regarded as manageable with small

numbers of important, interacting proteins implicated in the disease mechanism [34]. However for psychiatric disorders, which affect a significant proportion of the population at some point during their lives, the molecular complexity is much higher and gaining statistically significant genetic associations has required the largest of studies [39]. Better understanding of the molecular mechanisms underpinning these diseases is vital for the development of new treatments that are so urgently needed.

This presents a huge challenge to modern drug discovery, which is aimed at the identification of validated targets upon which drug screening and design can be based. How can we resolve disease mechanisms in such complex diseases? Moreover, the very cell type we often need to target is in itself complex. Neurons feature many connected compartments, each with their own proteome and often also with their own translational machinery. Molecular systems biology approaches provide the route, by which these diverse molecular targets and pathways can be resolved into complex models [24]. Those models firstly capture the biology and can then be used to make predictions that can help inform disease research, diagnoses and onwards to drug discovery. Ultimately, we need to integrate molecular systems models at neuronal synapses with cellular level models [47], through networks and into brain function and disorder. Here we discuss progress towards the molecular models of neuronal synapses with consideration of their evolutionary history, their link to cognition in healthy individuals and their role in disease. Critically, we examine where current models and methodologies have taken us, what we have learned, what their limitations are and finally present a new modelling framework that may help resolve some of the important issues with current methods.

2 Models of the Synaptic Proteome

Models of the synaptic proteome have largely been developed from proteomic analysis of neural tissue with a focus on either the pre-synaptic machinery [7, 37] or the post-synaptic density (PSD) [10, 28]. These studies have used a variety of fractionation or immunoprecipitation-based approaches to isolate protein complexes from brain tissue samples that are enriched for either pre- or post-synaptic proteins or alternatively proteins closely linked with one or more key molecular baits.

Large-scale fractionation experiments are generally based on separating out the synapses (synaptosomes) using careful centrifugation in a density gradient. This approach has been particularly useful in obtaining a holistic view of what proteins are found at the synapse with successful studies in both rodent [10, 28] and human [2] brain samples. Pleasingly, the total number of proteins reported by these studies appears to be slowing. Initially there was very little overlap between these studies (approximately 35% reported by a comparison analysis performed by Collins et al. [10]). While the number has increased dramatically in recent studies (mostly due to increased sensitivity in mass spectrometry approaches), the overlap between parallel studies has started to increase significantly. When combined and taking into consideration the convergence of overlap, there is already evidence for some

3,000 proteins at the synapse with a probable total in the vicinity of 4,000 or so [6, 10, 41, 51, 52]. While these studies are based on mass spectroscopy-based evidence from peptide fragments, the signatures obtained are mapped onto gene models and so the numbers found in existing studies do not reflect any additional complexity that may be provided by alternative transcripts.

While the purification and cell specificity are quite crude and there are clearly technical issues with trying to identify low abundance proteins, these studies do provide a useful initial parts list. Connecting these lists together to form the first networks can be performed using a variety of approaches. In early modelling studies [43], the quality of the pioneering yeast-2-hybrid interaction studies was felt to be questionable and so literature-based approaches were adopted that used initial high-throughput text mining to identify candidate publications that described protein–protein interaction studies. These initial hits were then manually curated to ensure the biochemistry evidence was robust and that the gene synonyms did indeed refer to proteins of interest. In more recent studies [16], on-line protein–protein interaction databases were more heavily used, but they still retained the manual curation step, although more as a validation exercise. While the bulk of interactions in on-line databases are now of high quality, there remain a small number of examples where the evidence for interactions is very indirect, often based on co-expression of mRNA that has leaked into the protein–protein interaction datasets.

Increasingly, high-throughput protein–protein interaction screening technologies are getting more repeatable. Modern approaches feature extra, more accurate controls to help minimize false-positives [42] and there are now related assays in place that work much better with proteins that are, for example, membrane associated [29]. As a result high-throughput screens are rapidly becoming more comprehensive in their coverage, and use of these data to construct interaction models is much more routine. An important advance is the improvement in data provenance where it is now much easier to check exactly which sequence was used to generate the interaction data point compared to the manual and often frustrating process of digging through interactions in the literature where the quality of the materials and methods section varies widely [22].

While synaptosome level proteomics provides us with a useful global framework for developing synapse-signalling models, the majority of studies focus in on specific identified complexes, which can be purified in a number of ways. The most commonly used approach exploits naturally occurring antigens in the complex. Proteins are carefully extracted from brain tissue samples (it should be recognized that extracting membrane-associated protein complexes is a very specialist area) and then complexes are pulled down using antibodies attached to beads or columns. The samples are then washed and then finally the complex eluted for analysis. Affinity to synthetic peptides has also been used extensively to purify NMDA/glutamate receptors [28]. The principle here is that the C-terminal hexapeptide of the NR2B (Grin2B) subunit of the NMDA receptor is the key part of the protein that interacts with its major scaffolding protein PSD-95. Thus, the hexapeptide can be used to pull down PSD-95 and its interactors. Pioneered in the rodent brain, it as now also been applied to Drosophila brain tissues providing the first direct evolutionary

comparison of the synaptic proteome [15]. Finally, there is a clear move towards the use of transgenically inserted antigens that can function as highly effective affinity tags. These are often combined with expression level markers and can be inserted as synthetic, over expression constructs or alternatively inserted into the endogenous gene, thus, retaining (as closely as possibly) the natural spatiotemporal control of expression. The system (known as tap-tag) in mice has been used to dissect protein complexes associated with PSD-95 in the brain [16] and provided a useful counterpoint to the hexapeptide studies [28]. In *Drosophila*, the approach taken to date has been (rather typically for the field) more stochastic with the use of a randomly inserting mobile protein-traps that splice into the mRNA transcript both expression markers (GFP variants) as well as high-affinity epitope tags (e.g. STREP and Flag). Many proteins have been tagged this manner with expression in various tissues covered [27, 44], including a specific screen for brain expression patterns [32].

As mentioned above, comparable proteomic techniques have now been applied to brains of multiple species, providing a window on the evolutionary origins of synapses. There are two basic methods for evolutionary comparison, computational and proteomic. Both approaches were employed by Emes et al. [15] who used the rodent (mouse) post-synaptic density as a base for comparative bioinformatics. This computational approach suggested that all vertebrate genomes had more or less the full complement of genes to support a similar synaptic proteome whereas invertebrate genomes could only account for roughly 50% of the molecular diversity. They hypothesized that the synaptic complexity in the smaller invertebrate brain was simpler. Further, some 20–25% of the complement of proteins required had orthologues in unicellular organisms (that have no nervous systems), suggesting neuronal signalling complexes evolved from cell surface receptors found in primitive unicellular ancestors. Down lineages with nervous systems, gene duplication would appear to account the increased molecular complexity and this is biased towards the receptor and scaffold proteins, which show the largest increase in numbers. The null hypothesis for the invertebrate brain is that their synapses are actually equally complex but use a different complement of proteins that would, therefore, be missed by an entirely one-sided bioinformatics approach. Emes et al. [15] then tested this hypothesis by performing proteomics analysis in an invertebrate brain (Drosophila) and confirmed that the synaptic complexity was indeed reduced (by around 50%) in comparison with the mouse studies.

The reduced molecular diversity observed in the smaller invertebrate brain in comparison with the larger vertebrate brain with increased molecular complexity is an appealing story. However, as is often the case, it is much more complex when one examines the evidence in more detail. Core to the complex (at least in rodents) is the interaction between the NMDA receptor and the four membrane associated guanylate kinase (MAGUK) proteins. Invertebrates have a single orthologue for these four proteins and this can be presented a typical example of where gene family expansion accounts for much of the increased molecular diversity observed in the rodent synapse relative to the fly (i.e. 4× complexity in the mammal). Yet looking at the gene models for these proteins, one finds that the single Drosophila gene

(Dlg) is understood to give rise to at least 15 known polypeptides [25]. Combined, the four mammalian genes account for up to 26 potential transcripts substantially reducing the gap in complexity when one converts the gene models into potential proteins [26]. The potential for next generation sequencing to dissect gene models is particularly exciting and in a few years we can expect a much clearer picture about the structure and spatiotemporal expression of splice variants in the brain [21].

Current datasets from proteomics studies do not routinely provide clear indications of exactly which of the potential protein variants are present in any specific sample. These data will become more readily available as the proteomic technologies advance. Thus, neither the current models of the PSD nor the wet-lab approaches capture this issue satisfactorily at present. We need to develop both to fully understand what the real molecular differences are between these vastly different brains.

3 Capturing Dynamic Complexity

The brain and its underpinning neuronal and molecular structures are highly dynamic (plastic) in nature and this feature is vital to its function – the site where information is processed, dissected and stored away for the purpose of modulating the behaviour of the animal. Computational neuroscience, which looks at modelling neuronal processes at the cellular level has a long and distinguished history and is now well-established as a strong complementary partner to experimental neuroscience methods [8]. At this level, the dynamics of information flow and modulation are modelled carefully, compared to physiological recordings and used to generate new hypotheses and inform experimental studies as well as a framework for capturing domain understanding [14]. However, little of this has, to date, extended much below the cellular compartment level with models of receptor effects on ion conductance, etc. largely dissociated from the underlying molecular machinery [19].

One thing we certainly do know about molecular complexes at the synapse is that they are not static. Therefore, an obvious question emerges – what is the point in producing static models of an inherently dynamic structure? Obviously, it has actually already proven itself as a useful framework for inferring the function of less well-annotated genes (guilt by association) and in proposing new targets for biological analysis [17]. However, it is clear that to gain a more realistic understanding of how the synapse proteome works and how its dynamics are involved in cognitive processes (and disorders), at some point one must consider its dynamics, how it interacts with the rest of the cellular environment and how information flow (activity) affects its structure and subsequent responses.

What dynamics do we actually have to consider? Perhaps the simplest dynamic consideration is that most current proteomic methods provide an average peptide list gathered from many different cell types. In any single cell (or synapse), we expect that only a subset of the possible molecules is present. It has already been

shown that the expression of many molecules, typically of more recent evolutionary origin and most closely related to receptor specificity, shows the highest variability of expression with brain region [15]. Knowles-Barley et al. [32] also demonstrated that in certain cases proteins that can be found in synaptic proteomes and are known to interact biochemically in other tissues are not actually co-expressed in neurons.

Through increasing sensitivity in mass spec machines, the availability of better, high-resolution co-localized expression data and combined with mRNA expression information from identified neurons, we are rapidly approaching the time when we can start to dissect the global protein interaction framework by cell types. The potential for new insights through integration of proteomics with differential mRNA expression was demonstrated in mechanistic models of Alzheimer's that combined both [40]. Clearly these are important differences and the presence/absence of molecules in different neurons will have large impacts on the functional pathways each synapse can support. Binary presence or absence can actually be handled in the static networks but this does not capture the whole story. The relative abundance of these proteins is not binary rather on a wide spectrum and with new quantitative proteomics tools rapidly developing, the first datasets are starting to emerge. The combined use of several methods, namely, the electron microscopy with quantitative immunoblotting [5], quantitative MS [6] and green fluorescent protein (GFP)-based quantitative fluorescence calibration [50] uncovered quantitative information on the stoichiometric ratio of the main proteins that comprise the PSD [48].

Beyond quantitative proteomics, there are also complex regulatory mechanisms at play that are currently not well-served by existing protein interaction model approaches. Important examples include phosphodynamics, which regulates protein activity, protein binding and potentially cleavage sites within the complex as a whole. Local translation and RNA binding proteins are located at the synapse [36,53] and can induce novel protein production at the synapse independently of the nucleus. Finally, there is of course the pattern of synapse–nucleus communication, with synapse level events known to regulate nucleus level transcription (and one would assume splicing) [9].

In the early days of protein interaction modelling in neuroscience, it was fairly obvious that these limiting factors would emerge but there were simply too few data to constrain models with too few practical methods to test any model generated in a realistic timeframe. Some notable exceptions do exist with high-quality dynamic models of some signalling pathways that are used widely by biological cells (not neuronal specific) and where extensive data and validation are more feasible (e.g. MAP kinase pathway models [31, 45]). However, with rapidly maturing technologies enabling data capture at all these levels (SILAC, Y2H, Y3H, etc.) [46, 56, 57], in the imminent future, we clearly need to move towards a modelling framework that is fit for purpose. We need to balance flexibility and descriptive power with combinatorial complexity, especially since in the first instance many parameters will need to be estimated or sampled. In fact, given the scale of the problem, models may be able to highlight the most important parameters for initial wet-lab measurement or validation.

4 Rule-based Modelling

With these challenges in mind, we surveyed a range of modelling frameworks used in system biology. The correct balance of computational tractability with increased descriptive power for our purposes appears to be well-met by a relatively recent modelling approach known as rule-based modelling (in this instance the kappa framework) [11, 12, 23, 33]. This provides a relatively simple syntax to describe protein interactions and their properties and dependencies. Each model component can be formalized as an agent, with binding sites (binding domains and motifs), which in turn are subjected to modifications/states (phosphorylation, ubiquitination, etc.) [33]. Each protein–protein interaction can be formalized as a rule, which includes only the information that is relevant for the given interaction (particular domain in particular state) and omits all the irrelevant information (other domains and states) (Fig. 6.2). In other words, it can capture not only the binary interaction logic but also the new parameters we need to include such as binding site data, affinities, effect of post-translational modification (e.g. phosphorylation state), competitive binding and protein concentrations. Notably, the design of these rule-based approaches acknowledges computational complexity and allows for generic rules to be defined that encompass a class of interactions rather than forcing every one to be treated (and computationally optimized) independently.

In the past five years, several methodologies for rule-based modelling have been developed: StochSim, MCell, Smoldyn and ChemCell, BioNetGen (BNGL) and kappa language [1, 11, 18, 23, 38]. Each language implements its own specific spectrum of features based on practically the same principles. The rule-based approach was successfully implemented in the set of receptor signalling models, each designed with different rule-based techniques. This includes Tar-receptor–mediated hemotaxis, FcεRI- and TCR (T-cell receptor)-mediated responses in immunoreactivity, GPCR (G-protein coupled receptors)-signalling and many others [4,23,35,55]. The main advantage of all rule-based techniques is that the calculation efficiency does not depend upon the size of the network implied by the set of rules. That makes possible simulating the formation of the multi-subunit signalling complex simultaneously with the receptor-mediated phosphorylation dynamics.

Most previous examples of rule-based models were applied to dynamic modelling of specific signalling cascades. However, kappa formalism was used already for theoretical analysis of “liquidity” of the protein agglomerate at equilibrium [13]. Following this approach, we took the first steps towards quantitative model development by examining the steady states that are reachable by the system rather than on the detailed dynamics of transition processes. The analysis of the topological properties of final complex is, therefore, a natural and step-wise extension of the topology analysis of original PPI network models but which now considers protein abundance and interaction affinities.

The modelling framework is still inherently graph based (Fig. 6.3) and, therefore, we can use pre-existing interaction data and visualization methods to jump-start the activity. With the addition of known or estimated parameters, rule-based systems

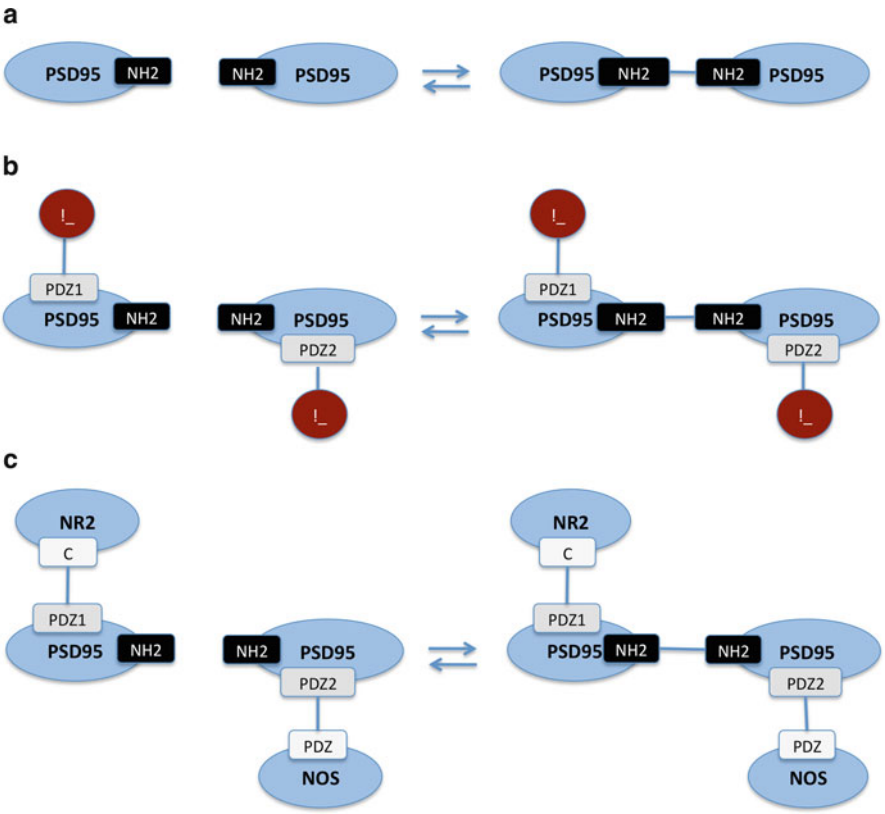


Fig. 6.2 Rule examples with different levels of contextualization. Depending on reaction knowledge and modelling purpose, the binding could be modelled as unconditional (**a**) or depending on specific conditions (**b** and **c**). *Rule A: PSD95(NH2), PSD95(NH2)* → *PSD95(NH2!0)PSD95 (NH2!0)*. This is the most basic type of the rule, used most often in the model. Two molecules of PSD95 make a dimer irrespective of the states of any other domains of PSD95. Therefore, this rule covers all the possibilities, including the more specific cases B and C. *Rule B: PSD95(NH2, PDZ1!_), PSD95(NH2, PDZ2!_)*→ *PSD95(NH2!0, PDZ!_), PSD95 (NH2!0, PDZ2!_)*. This rule adds constrains: PSD95 molecules have to be bound through their PDZ1/PDZ2 domains to bind each other. !_ means that the identity of the binding partner is not specified. *Rule C:* PSD95(NH2, PDZ1!0), NR2(c!0), PSD95(NH2, PDZ2!1), NOS(PDZ!1) → PSD95(NH2!2, PDZ!0), NR2(c!0), PSD95(NH2!2, PDZ2!1), NOS(PDZ!1). This example contains the most specific constrains: only those molecules of PSD95 can bind each other that are already bound through their PDZ1/PDZ2 domains to C-terminus of NR2 receptor molecule and PDZ domain of NOS

can then be simulated using either deterministic (ODE) or, more often, stochastic methods. That approach allows us to test parameter ranges and the effects these have on models that can be supported. We can also extend this and perform sensitivity analysis on individual parameters to measure how influential each is on the global network architecture.

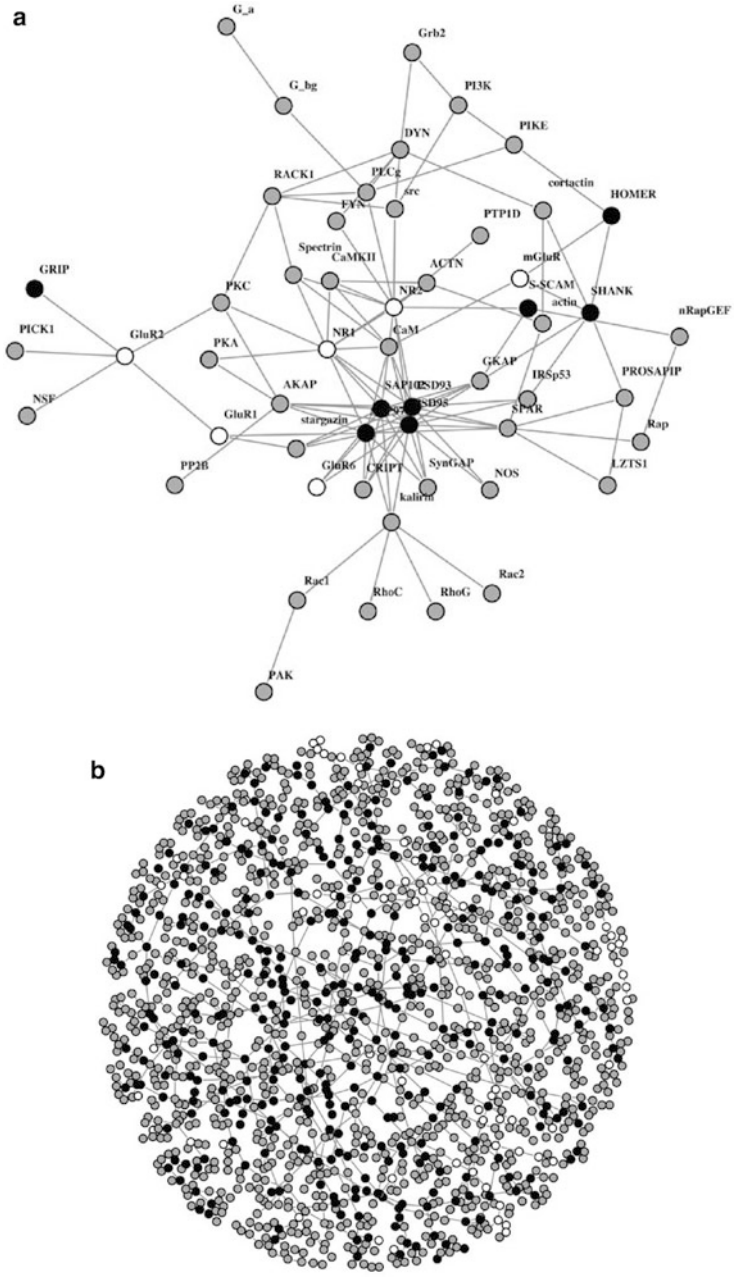


Fig. 6.3 **Panel a** shows a typical static PPI network model where each molecular species is represented a single time and is fully connected with its interaction partners. This specific example is a simplified interaction network containing the core molecules of the NMDA receptor in the post-synaptic density. **Panel b** shows the result of a rule-based model simulation of the same network in **Panel a**. In this example, the PDZ-domain containing scaffold proteins are *black*, the receptor molecules are in *white* and all other molecule types are in *grey*

5 A Kappa Model of the Post-synaptic Density

As a proof of concept, we recently engineered a kappa model of a core set of proteins from the PSD (Fig. 6.3, for details, see [49]). In summary, we selected 54 key proteins isolated in a range of proteomic studies of the PSD and for which we have at least a general understanding of the role they are likely to play in synaptic mechanics. These include key classes of molecule from the cell membrane receptors through scaffolding proteins, in particular, the MAGUKs, GTPases, kinases, phosphatases and structural proteins.

The 54 proteins in the model would require ∼150 reversible rules (e.g. Fig. 6.2) to describe their interactions with a total of ∼300 parameters that require definition. However, taking into account that although functionally different, proteins of PSD are generally enriched with several classes of domain, which comprise complementary interaction pairs, the multiple interactions within PSD could be divided into subclasses according to this feature. Rule decontextualisation, then, allows us to generalize the interaction logic, so that we can define a single parameter for entire subclass of domain interactions within the complex, such as the common interactions between PDZ domains of the MAGUKs and the C-terminal domains of the NMDA receptor 2b subunit. Therefore, the number of parameters could be substantially reduced (to 84 rate constants in the case of our model). The model comprises association and dissociation rules for 54 agents (domain–domain interactions), drawn from the literature.

$$\text{'DLG\_GKAP'DLG(GK), GKAP(GKBD)} \rightarrow \text{DLG(GK1!1), GKAP(GKBD!1)@k20} \quad (6.1)$$

$$\text{'DLG\_GKAP\_diss'DLG(GK!1), GKAP(GKBD!1)} \rightarrow \text{DLG(GK), GKAP(GKBD)@k\_20} \quad (6.2)$$

The example rule above, in kappa syntax rules, describes the reactions for association (6.1) and dissociation (6.2) for members of MAGUKS/DLG family of proteins with their interactor GKAP. Here, the association and dissociation happen via GK (guanylate kinase) and GKBD (guanylate kinase binding domain) domains, where the rate of forward reaction is k20 and the rate of backward reaction is k_20. In accordance with above, the pair of constants k20/k_20 could be substituted not only for all 4 PSD-95 family members but also for other model agents that carry GK domain. Therefore, list of rules formulated for specific subclasses of domain–domain interactions could be easily applied to any other protein–protein interaction network, which is based on the same principle domain association. The rule could include more or less context, depending on our knowledge of protein–protein interaction and type of the model (Fig. 6.3).

Stochastic simulation then gives an indication of the capacity of the model PSD to generate complexes of different size and composition when system attains the steady state (Fig. 6.3). The stability of these complexes is directly influenced by the parameter values, e.g. by dissociation constant (K_d), and we can examine these

effects. The numerical analysis of the relative steady state distribution of protein complexes and their sizes allows comparison of the molecular structure of the PSD model under different perturbations. Importantly, we can vary protein stoichiometry and obtain different compositions of protein agglomerations. For example, we can now easily simulate a knockout phenotype for each model element and look how this affects the structure and the size of the protein complexes in the equilibrium. The example of PSD95 mutant presented in [49] shows that changing the initial concentration of PSD95 from 300 copies to 0 significantly reduces the size of the average complex (from 300 molecules to 80) in the relative steady state and decreases the number of protein types in the average complex from 40 to 15. This kind of analysis gives new insights for future linking of physiological phenomena to underlying molecular restructuring mechanisms.

For any first model of this sort, we actually have access to very few laboratory verified parameters. Therefore, we can provide biologically relevant constraints to the model and use quasi-random sampling methods to pinpoint parameter values that tend to satisfy these constrains. We can further use sensitivity testing to rank order the influence each parameter has over the global architecture. For this first PSD model, we looked for parameter sets that produced diverse complexes with molecular weight ranges tending towards the known size of the rodent post-synaptic density [48].

The first (proof of concept) model does not include protein/domain cooperatively effects and only minimally touches upon the dynamic signalling connected with post-translational modifications. However, it has enough predictable power for interrogating the effects of different perturbations, such as change of protein concentration (mutants) and domain availability and affinity (introducing of splice variants and drugs), on the structural properties of the system.

6 Conclusions

“Each generation imagines itself to be more intelligent that the one that went before it, and wiser than the one that comes after it.” George Orwell.

In some respects, this captures the nature of the research we are faced with. As our understanding of the synaptic proteome grows, we require more complex methods to faithfully reproduce the biology within. We are clearly entering a step-change both in terms of biochemical analysis and model complexity for the synaptic proteome. We do not claim that the rule-based modelling approach we described here will be the ideal solution, rather that it is an evolutionary step and allows us to address the next generation of research questions as new technological developments permit. Specifically it balances the paucity of biochemical detail with the need to capture some dynamic information in larger, more complex interaction networks. We have already found it to be useful for simulating the stoichiometry in the complex in both natural and in mutated states, data that is already starting to become available from more quantitative proteomic studies. Further, it provides a

set of tools that allow us to extend into the analysis of interaction logic covering post-translational modifications including phosphodynamics, ubiquitination and competitive binding. In other words, as our understanding of the evolution of the molecular complex that underpins cognition is growing, the modelling frameworks we need to explore and describe these complexes also have to evolve.

Acknowledgements We acknowledge Anatoly Sorokin for help with simulation implementation. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk). The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement nos. HEALTH-F2–2009–241498 ("EUROSPIN" project) and HEALTH-F2–2009–242167 ("SynSys-project").

References

1. Andrews SS, Addy NJ, Brent R, Arkin AP (2010) Detailed simulations of cell biology with smoldyn 2.1. PLoS Comput Biol 6(3):e1000705
2. Bayés À, Lagemaat LNvd, Collins MO, Croning MDR, Whittle IR, Choudhary JS, Grant SGN (2011) Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat Neurosci 14:19–21
3. Bianchi MT, Botzolakis EJ (2010) Targeting ligand-gated ion channels in neurology and psychiatry: is pharmacological promiscuity an obstacle or an opportunity? BMC Pharmacol 10:3
4. Bray D, Bourret RB (1995) Computer analysis of the binding reactions leading to a transmembrane receptor-linked multiprotein complex involved in bacterial chemotaxis. Mol Biol Cell 6(10):1367–1380
5. Chen X, Vinade L, Leapman RD, Petersen JD, Nakagawa T, Phillips TM, Sheng M, Reese TS (2005) Mass of the postsynaptic density and enumeration of three key molecules. Proc Natl Acad Sci USA 102(32):11551–11556
6. Cheng D, Hoogenraad CC, Rush J, Ramm E, Schlager MA, Duong DM, Xu P, Wijayawardana SR, Hanfelt J, Nakagawa T, Sheng M, Peng J (2006) Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. Mol Cell Proteom 5:1158–1170
7. Chua JJE, Kindler S, Boyken J, Jahn R (2010) The architecture of an excitatory synapse. J Cell Sci 123:819–823
8. Churchland PS, Koch C, Sejnowski TJ (1993) What is computational neuroscience? In: Computational neuroscience. MIT, Cambridge
9. Cohen S, Greenberg ME (2008) Communication between the synapse and the nucleus in neuronal development, plasticity, and disease. Annu Rev Cell Dev Biol 24:183–209
10. Collins MO, Husi H, Yu L, Brandon JM, Anderson CNG, Blackstock WP, Choudhary JS, Grant SGN (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. J Neurochem 97:16–23
11. Danos V, Feret J, Fontana W, Harmer R, Krivine J (2009) Rule-based modelling and model perturbation. Trans Comput Syst Biol XI, Lecture Notes in Computer Science 5750:116–137
12. Danos V, Feret J, Fontana W, Krivine J (2007) Scalable simulation of cellular signaling networks. Proceedings of APLAS
13. Danos V, Schumacher LJ (2009) How liquid is biological signalling? Theor Comput Sci 410(11):1003–1012

14. Druckmann S, Banitt Y, Gidon A, Schürmann F, Markram H, Segev I (2007) A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. Front Neurosci 1(1):7–18
15. Emes RD, Pocklington AJ, Anderson CNG, Bayes A, Collins MO, Vickers CA, Croning MDR, Malik BR, Choudhary JS, Armstrong JD, Grant SGN (2008) Evolutionary expansion and anatomical specialization of synapse proteome complexity. Nat Neurosci 11:799–806
16. Fernández E, Collins MO, Uren RT, Kopanitsa MV, Komiyama NH, Croning MDR, Zografos L, Armstrong JD, Choudhary JS, Grant SGN (2009) Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. Mol Syst Biol 5(269)
17. Frank R, McRae A, Pocklington A, Lagemaat Lvd, Navarro P, Croning M, Komiyama N, Bradley S, Challiss R, Armstrong J, Finn R, Malloy M, MacLean A, Harris S, Starr J, Bhaskar S, Howard E, Hunt S, Coffey A, Ranganath V, Deloukas P, Rogers J, Muir W, Deary I, Blackwood D, Visscher P, Grant S (2011) Clustered coding variants in the synaptic receptor complexes of individuals with schizophrenia and bipolar disorder. PloS One 6(4):e19011
18. Franks KM, Bartol TM, Sejnowski TJ (2001) An mcell model of calcium dynamics and frequency-dependence of calmodulin activation in dendritic spines. Neurocomputing 38(40): 9–16
19. Goldwyn JH, Imennov NS, Famulare M, Shea-Brown E (2011) Stochastic differential equation models for ion channel noise in Hodgkin–Huxley neurons. Phys Rev E Stat Nonlin Soft Matter Phys 83(4 Pt 1):041908
20. Grant SGN, Marshall MC, Page K-L, Cumiskey MA, Armstrong JD (2005) Synapse proteomics of multiprotein complexes: en route from genes to nervous system diseases. Human Mol Genetics 14(Suppl 2):R225–R234
21. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genetics 11:476–486
22. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, Mering Cv, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat Biotechnol 22:177–183
23. Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W (2006) Rules for modeling signal-transduction systems. Sci STKE 344:re6
24. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–643
25. http://flybase.org/reports/FBgn0001624.html http://flybase.org/reports/FBgn0001624.html
26. http://www.ensembl.org/Homo_sapiens/Info/Index
27. http://www.flyprot.org/
28. Husi H, Ward MA, Choudhary JS, Blackstock WP, Grant SGN (2000) Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. Nat Neurosci 3:661–669
29. Johnsson N, Varnavsky A (1994) Split ubiquitin as a sensor of protein interactions in vivo. Proc Natl Acad Sci USA 91(22):10340–10344
30. Kandel ER, Schwartz JH (1985) Principles of neural science. McGraw-Hill, New York.
31. Kholodenko BN (2006) Cell signalling dynamics in time and space. Nat Rev Mol Cell Biol 7(3):165–176
32. Knowles-Barley S, Longair M, Armstrong JD (2010) Braintrap: a database of 3D protein expression patterns in the Drosophila brain. Database (Oxford) 2010:baq005
33. Krivine J, Danos V, Benecke A (2009) Modelling epigenetic information maintenance: a kappa tutorial. Comput Aided Verification, Lecture Notes in Computer Science 5643/2009:17–32
34. Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B, Letenneur L, Bettens K, Berr C, Pasquier F, Fiévet N, Barberger-Gateau P, Engelborghs S, Deyn PD, Mateo I, Franck A, Helisalmi S, Porcellini E, Hanon

O, Investigators tEAsDI, Pancorbo MMd, Lendon C, Dufouil C, Jaillard C, Leveillard T, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Bossù P, Piccardi P, Annoni G, Seripa D, Galimbert D, Hannequin D, Licastro F, Soininen H, Ritchie K, Blanché H, Dartigues J-F, Tzourio C, Gut I, Broeckhoven CV, Alpérovitch A, Lathrop M, Amouyel P (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 41:1094–1099

35. Lee K-H, Dinner AR, Tu C, Campi G, Raychaudhuri S, Varma R, Sims TN, Burack WR, Wu H, Wang J, Kanagawa O, Markiewicz M, Allen PM, Dustin ML, Chakraborty AK, Shaw AS (2003) The immunological synapse balances t cell receptor signaling and degradation. Science 302(5648):1218–1222
36. Mikl M, Vendra G, Doyle M, Kiebler MA (2010) RNA localization in neurite morphogenesis and synaptic regulation: current evidence and novel approaches. J Comp Physiol A 196(5): 321–334
37. Morciano M, Beckhaus T, Karas M, Zimmermann H, Volknandt W (2009) The proteome of the presynaptic active zone: from docked synaptic vesicles to adhesion molecules and maxi-channels. J Neurochem 108(3):662–675
38. Novère NL, Shimizu TS (2001) STOCHSIM: modelling of stochastic biomolecular processes. Bioinformatics 17(6):575–576
39. Owen MJ, Craddock N, O'Donovan MC (2010) Suggestion of roles for both common and rare risk variants in genome-wide studies of Schizophrenia. Arch Gen Psychiatry 67(7):667–673
40. Papassotiropoulos A, Fountoulakis M, Dunckley T, Stephan DA, Reiman EM (2006) Genetics, transcriptomics and proteomics of Alzheimer's disease. J Clin Psychiatry 67(4):652–670
41. Peng J, Kim MJ, Cheng D, Duong DM, Gygi SP, Sheng M (2004) Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. J Biol Chem 279:21003–21011
42. Petschnigg J, Snider J, Stagljar I (2011) Interactive proteomics research technologies: recent applications and advances. Curr Opin Biotechnol 22(1):50–58
43. Pocklington AJ, Cumiskey M, Armstrong JD, Grant SGN (2006) The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. Mol Syst Biol 2:0023
44. Rees JS, Lowe N, Armean IM, Roote J, Johnson G, Drummond E, Spriggs H, Ryder E, Russell S, Johnston DS, Lilley KS (2011) In vivo analysis of proteomes and interactomes using parallel affinity capture (ipac) coupled to mass spectrometry. Mol Cell Proteomics 10(6):M110.002386
45. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G (2004) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol 20:370–375
46. Seay D, Hook B, Evans K, Wickens M (2006) A three-hybrid screen identifies mRNAs controlled by a regulatory protein. RNA 12(8):1594–1600
47. Sejnowski T, Koch C, Churchland P (1998) Computational neuroscience. Science 241(4871):1299–1306
48. Sheng M, Hoogenraad CC (2007) The postsynaptic architecture of excitatory synapses: a more quantitative view. Ann Rev Biochem 76:823–847
49. Sorokina O, Sorokin A, Armstrong JD (2011) Towards a quantitative model of the post synaptic proteome. Mol BioSyst 7(10):2813–2823
50. Sugiyama Y, Kawabata I, Sobue K, Okabe S (2005) Determination of absolute protein numbers in single synapses by a GFP-based calibration technique. Nat Methods 2:677–684
51. Trinidad JC, Thalhammer A, Specht CG, Lynn AJ, Baker PR, Schoepfer R, Burlingame AL (2008) Quantitative analysis of synaptic phosphorylation and protein expression. Mol Cell Proteom 7:684–696
52. Trinidad JC, Thalhammer A, Specht CG, Schoepfer R, Burlingame AL (2005) Phosphorylation state of postsynaptic density proteins. J Neurochem 92(6):1306–1316
53. Wang DO, Martin KC, Zukin RS (2010) Spatially restricting gene expression by local translation at synapses. Trends Neurosci 33(4):173–182

54. Wimo A, Jönsson L, Gustavsson A, McDaid D, Ersek K, Georges J, Gulácsi L, Karpati K, Kenigsberg P, Valtonen H (2010) The economic impact of dementia in Europe in 2008 – cost estimates from the Eurocode project. Int J Geriatric Psychiatry 26(8):825–832
55. Woolf PJ, Linderman JJ (2004) An algebra of dimerization and its implications for G-protein coupled receptor signaling. J Theor Biol 229(2):157–168
56. Young KH (1998) Yeast two-hybrid: so many interactions, (in)so littletime. . .. Biol Rep 58:302–311
57. Zhang G, Neubert TA (2009) Use of stable isotope labeling by amino acids in cell culture (silac) for phosphotyrosine protein identification and quantitation. Meth Mol Biol 527(II):79–92

Chapter 7
Molecular Systems Biology of Sic1 in Yeast Cell Cycle Regulation Through Multiscale Modeling

Matteo Barberis

Abstract Cell cycle control is highly regulated to guarantee the precise timing of events essential for cell growth, i.e., DNA replication onset and cell division. Failure of this control plays a role in cancer and molecules called cyclin-dependent kinase (Cdk) inhibitors (Ckis) exploit a critical function in cell cycle timing. Here we present a multiscale modeling where experimental and computational studies have been employed to investigate structure, function and temporal dynamics of the Cki Sic1 that regulates cell cycle progression in *Saccharomyces cerevisiae*. Structural analyses reveal molecular details of the interaction between Sic1 and Cdk/cyclin complexes, and biochemical investigation reveals Sic1 function in analogy to its human counterpart $p27^{Kip1}$, whose deregulation leads to failure in timing of kinase activation and, therefore, to cancer. Following these findings, a bottom-up systems biology approach has been developed to characterize modular networks addressing Sic1 regulatory function. Through complementary experimentation and modeling, we suggest a mechanism that underlies Sic1 function in controlling temporal waves of cyclins to ensure correct timing of the phase-specific Cdk activities.

Abbreviations

Cdk:	Cyclin-dependent kinase
Cki:	Cyclin-dependent kinase inhibitor
ODE:	Ordinary differential equation
KID:	Kinase inhibitory domain

M. Barberis (✉)
Humboldt University Berlin, Institute for Biology, Invalidenstr. 42, 10115 Berlin, Germany
e-mail: matteo.barberis@biologie.hu-berlin.de

Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany
e-mail: barberis@molgen.mpg.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_7,

IDP: Intrinsically disordered protein
FRET: Förster resonance energy transfer
FLIM: Fluorescence lifetime imaging microscopy

1 Timing in Cell Cycle Regulation

Cell cycle regulation is governed by sequential activation of a family of serine–threonine cyclin-dependent kinases (Cdks), whose activities rise and fall are being controlled by a complex regulatory network. Since timely regulation of Cdk/cyclin complexes is critical for proper completion of cell cycle phases, multiple signals have to be integrated to control their activity. Besides cyclin accumulation, localization, and phosphorylation/dephosphorylation [1, 2], Cdk regulation is mediated by cyclin-dependent kinase inhibitors (Ckis), which ensure the correct timing of its activation in different cell cycle phases [3]. Cki inhibitors have been proposed to define thresholds for Cdk/cyclin activity by setting levels that Cdk/cyclin complexes must exceed to become active [4]. Accordingly, cell cycle progression or arrest would depend on relative concentration of inhibitors and cyclins: a decrease in Cdk/cyclin components or an increase in inhibitor levels would prevent the accumulation of inhibitor-free Cdk/cyclin complexes, therefore inhibiting cell cycle progression. The Cki $p27^{Kip1}$ of the Kip/Cip family is a key protein establishing the threshold that Cdk/cyclin complexes must overcome in order to progress into S phase in mammalian cells [4, 5]. The amount of $p27^{Kip1}$ is rate limiting for cell cycle progression and alters the balance between proliferation or arrest. In fact, its misregulation is found in various cancer types [6–11] due to an abnormal activation of Cdk/cyclin inhibited by $p27^{Kip1}$. Reduction of $p27^{Kip1}$ activity increases the proliferation rate in tumor cells [12], and desensitizes cells to antimitogenic signals, thus preventing their apoptosis [13]. Yet, the molecular context in which Cki inhibitors activate and regulate cell cycle progression is not fully understood. More specifically, it is of relevant interest to investigate the molecular mechanisms that lead to deregulation of Ckis like $p27^{Kip1}$ – and therefore to the failure in precise timing of kinase activation – in several tumors including breast, colon, prostate, lung, esophageal, and gastric cancers [8, 9, 14].

1.1 Failure of Cki Control and Cancer Development

$p27^{Kip1}$ functions throughout all cell cycle phases by interacting with different Cdk/cyclin complexes. It has a crucial role at the G1/S transition by interacting with and inhibiting Cdk2/cyclin E and Cdk2/cyclin A activities, thus blocking cell cycle progression. High protein levels lead to cell cycle arrest in G1 phase [15, 16], whereas cell cycle re-enter requires $p27^{Kip1}$ downregulation, resulting in Cdk activation [17]. $p27^{Kip1}$ binds also to Cdk4,6/cyclin D, being both a Cdk4

inhibitor and a non-inhibitor depending on the growth state of the cell [18]. In early G1 phase, $p27^{Kip1}$ promotes assembly and nuclear import of Cdk4,6/cyclin D, increasing cyclin D stability, without inhibiting Cdk4 activity [19, 20]. This ternary complex functions as a reservoir for $p27^{Kip1}$, which through Cdk4/cyclin D binding is displaced from its principal target, Cdk2/cyclin E.

$p27^{Kip1}$ is a dosage-dependent tumor suppressor genes whose functional loss leads to tumor development, being its reduced dosage contributing to cancer susceptibility. Mice lacking one copy of *CDKN1B* the gene encoding for $p27^{Kip1}$, display increased tumor frequency [21] and $p27^{Kip1} -/-$ mice display a further increase in tumor rate developing tumors in multiple tissues, including adenomas and adenocarcinomas of the intestine and the lung, granulosa cell tumors of the ovary and uterine tumors [21]. The dosage effect might be functioning in human tumors as well, where loss or decrease of $p27^{Kip1}$ expression is frequently observed in human cancer and correlates with poor patient survival. The first human cases reported to have abnormally low amounts of nuclear $p27^{Kip1}$ were associated with increased tumor aggressiveness and a relatively poor clinical outcome for breast and colon cancer [22]. Moreover, the correlation of cytoplasmic localization of $p27^{Kip1}$ with high tumor grade and poor prognosis has been reported [23, 24]. It has been now recognized that $p27^{Kip1}$ deregulation can be a prognostic indicator for a variety of tumors [7–9, 14].

1.2 Sic1, the Cki Regulating Cell Cycle Timing in Budding Yeast

The budding yeast cell cycle is driven by periodic changes in kinase activities, regulated by different cyclin subunits that associate with the Cdk1 kinase in successive waves: Cln1, Cln2, and Cln3 in G1 phase; Clb5 and Clb6 in S phase; Clb1, Clb2, Clb3, and Clb4 in G2/M phase [25, 26]. Despite their redundancy, cyclins are expressed at a different timing and appear sequentially in specific cell cycle phases, resulting in a significant divergence of function [27–29]. Besides accumulation and degradation, specific Ckis contribute to the regulation of cyclins: Far1 inhibits Cdk1/Cln complexes [30, 31] and Sic1 inhibits Cdk1/Clb complexes [32, 33]. The logic of a Cki/cyclin threshold that drives phase-specific events has been proposed in basic models of cell cycle progression in budding yeast, for the entrance into S phase by activating waves of cyclins that set the timing for mitosis onset and cell division [34, 35].

Sic1 and Cdk1/Clb complexes that drive S and M phases are locked in mortal combat over control of the cell cycle: Sic1 inactivates Cdk1/Clb complexes and promotes Clb degradation, whereas Cdk1/Clbs antagonize Sic1 transcription and promote Sic1 degradation [36, 37]. Sic1 is synthesized at the end of mitosis [38–40] and persists throughout G1 phase preventing the precocious DNA synthesis by inhibiting Cdk1/Clb5,6 activity [33, 41]. Sic1 is largely degraded at the onset

of S phase [33, 42, 43] with a switch-like mechanism via multisite phosphorylation [44] mediated by Cdk1/Cln1,2 [45–48], thus relieving Cdk1/Clb5,6 inhibition and allowing cells to enter into S phase. Moreover, all Cdk1/Clb complexes can maintain Sic1 proteolysis during S and G2 phases until anaphase [26]. As cells exit mitosis, Cdk1/Clb2 activity declines and Sic1 is produced. Similarly to $p27^{Kip1}$, which stably associate with Cdk4/cyclin D1 to assemble them into active complexes, Sic1 has been shown to promote nuclear import of Cdk1/Clb5, since cytoplasmic Clb5 accumulation is observed upon inactivation of *SIC1* gene [49].

Sic1 is not an essential gene but it plays an essential role in setting the correct timing of DNA replication onset by maintaining a temporal window free from Cdk1/Clb activity, critical requirement for origin licensing. In a *sic1Δ* mutant, DNA replication initiates prematurely from fewer origins, S phase is extended and sister chromatids are inefficiently separated during anaphase [50]. In addition, chromosome combing showed that the distance between replicons is 1.5 times longer in *sic1Δ* cells compared to wild type [51]. As a consequence, chromosomes break and rearrange at a high frequency, and cells exhibit a 100-fold increase in minichromosome loss and gross chromosomal rearrangements compared to wild type [50, 52]. The precocious Cdk1/Clb5,6 activation causes severe genome instability through its inhibitory effect on pre-RC formation in late G1 phase Similarly, $p27^{Kip1}$-deficient cells activate the G2/M checkpoint and show an increased number of chromatid breaks, leading to chromosomal instability, a hallmark of cancers with poor prognosis [53]. Thus, by inhibiting any residual Cdk1/Clb activity in G1 phase, Sic1 promotes efficient origin licensing, probably also activating dormant origins [54]. As it has been underlined for $p27^{Kip1}$, the molecular mechanism by which Sic1 regulates the Cdk1/Clb activities is not fully understood.

2 Cell Cycle Regulation and Computational Modeling

To understand biological processes, biomolecules are generally investigated in the framework of molecular networks [55] and molecular systems biology is the integrative discipline that aims to explain properties of biological systems in terms of their molecular components and interactions [56, 57]. The structure of these networks can vary over time and space generating network dynamics [58] and modularity permits to dissect complex biological networks in small modules and provides functional and mechanistic insights [59]. Molecules contributing to a particular phenotype are usually connected to each other to form functional modules [60, 61] and have similar biological functions, as suggested from the dynamically organized modularity in the budding yeast interactome network [62, 63]. Although some modules, such as stable protein complexes, are constantly present in various cellular conditions, other modules are dynamically assembled and disassembled.

Cell cycle-dependent protein complexes undergo this temporal dynamics during different phases of the cell cycle [64]. In mammalian cells, their formation has been addressed by non-linear differential equations [65], qualitative modeling [66], or

hybrid approaches that combine continuous differential equations and discrete Boolean networks [67]. Detailed models focused on the G1/S transition have been reported, with particular focus on Ckis regulating timing of cell cycle progression [68–74]. Cell cycle regulators are conserved between eukaryotes and mathematical models have also been developed for budding yeast by using systems of ordinary differential equations (ODEs) [65, 75–78], logical [79–81], stochastic [78, 82–84], or numerical [85] modeling. Networks focused on the exit from mitosis are available [86–90] and detailed analyses of the G1/S transition highlight the specific role of the Cki Sic1 in regulating the timing of DNA replication onset [91, 92].

These models capture essential molecular events occurring during temporal dynamics of cell cycle progression. In this way, components of modular cell cycle networks and their interactions can be identified following an iterative process, in which molecular investigation and mathematical modeling account for the system behavior. This approach has been pursued in our laboratories following a bottom-up systems biology, starting from constitutive parts of a network by formulating their interactions, the kinetic equations, and then predicting the system behavior [93]. Multiscale modeling integrating structural analyses, biochemical investigation, and protein dynamics into non-linear and stochastic modeling has been employed to address regulatory functional modules centered around the Cki Sic1. Models have been tested for internal consistency by computational analyses and external consistency by experimental validation [84, 92, 94–97]. Despite the well-recognized role of Sic1 in regulating cell cycle progression, there is a controversy about the specific phases where this Cki functions. Sic1 is largely degraded at the G1/S transition to permit DNA synthesis [33, 42, 43]; however, *SIC1* transcription [38, 40] and Sic1 levels [98, 99] are observed throughout the entire cell cycle, and recent data claim that Sic1 contributes to cell cycle robustness [100]. Here we summarize how literature evidence are reconciled through a molecular systems biology approach suggesting a role of Sic1 in regulating the timing of cyclin waves.

3 Structural Modeling of the Sic1-Cdk1/Clb Interaction

Comprehension of the interactions between molecules involved in biological systems is crucial to predict their behavior by a systems biology approach. A full understanding of how molecules interact comes only from three-dimensional (3D) structures, as they provide atomic details about binding. Although a huge number of interactions is known, precise molecular details are available for few of them due to limitations in studying large protein complexes, for which obtaining sufficient purified material for X-ray studies can be difficult. In fact, assembly of two or more macromolecules in complex requires precise control and timing in the cell, and this is not easy to reproduce in a laboratory setting [101]. For both human and budding yeast, there is a large gap between the number of complexes detected in yeast two-hybrid [102–106] or affinity purification [107, 108] assays and the number for which experimental 3D structures are available. Therefore, methods to

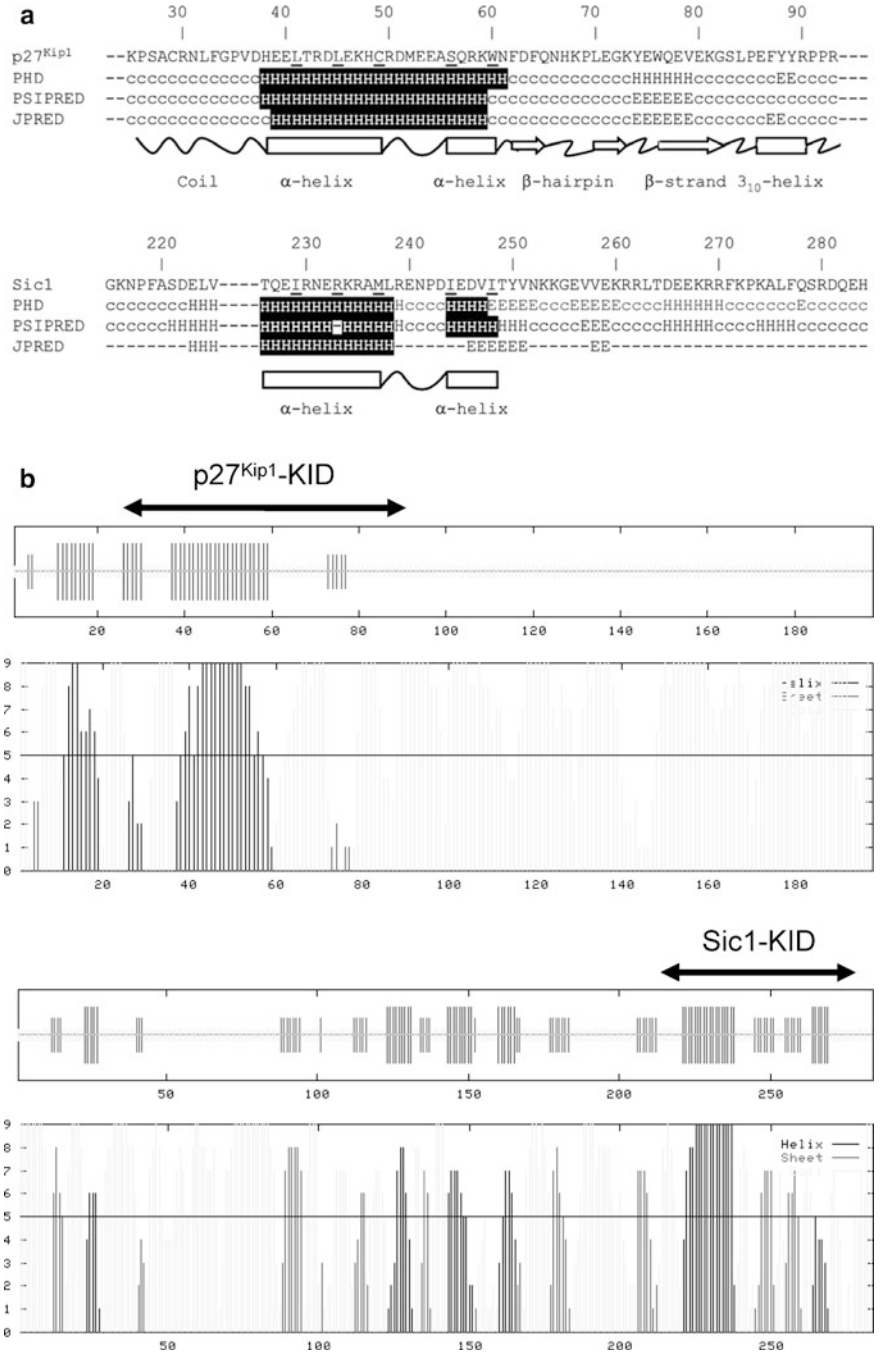


Fig. 7.1 (overleaf) (**a**) Sequence alignment and secondary structure predictions for $p27^{Kip1}$ and Sic1 KIDs. Predictions computed by PHD, PSIPRED, and JPRED programs for $p27^{Kip1}$-KID

predict atomic details for interacting proteins have been developed. The structures of interacting proteins can be modeled computationally if structures have been previously determined for homologous proteins. There are many interactions for which structural data are available [109, 110] and homology modeling is used to test whether interactions between homologous proteins can be modeled on the basis of an interaction of known structure [111–114]. The accuracy of models built by homology depends on the degree of sequence identity between target and template sequences. When sequence similarity is high, i.e., greater than 25–30% of sequence identity, proteins are likely to interact in the same way [110].

Despite efforts from different groups, attempts to crystallize Sic1 have failed. To investigate molecular interactions between the Cki Sic1 and Cdk1/Clb complex, we therefore employed a protein–protein complex for which coordinate data are available to model interactions between their analogous in budding yeast. We assessed whether potential homologous sequences fit onto a previously determined structure of a complex. Sic1 is a functional homologue of the Cki Rum1 in fission yeast [115] and a potential functional homologue of mammalian Ckis of the Kip/Cip family [116], thus local properties at the interaction surface with Cdk/cyclin could be conserved. Conventional homology modeling is not applicable to Sic1 due to its very low sequence similarity to $p27^{Kip1}$; however, secondary structure predictions of kinase inhibitory domains (KIDs) of Sic1 (C-terminal, amino acids 215–284) [117] and $p27^{Kip1}$ (N-terminal, amino acids 25–93) [118] suggest a similarity in this region as compared with the secondary structure deduced from the X-ray structure of $p27^{Kip1}$-KID [118] (Fig. 7.1a). Sequence alignment of $p27^{Kip1}$ and Sic1 KIDs highlighted a long α-helix predicted in Sic1 (amino acids 226–248) which shares a similar amphiphilic profile with the corresponding α-helix of $p27^{Kip1}$ (amino acids 38–60) [94]. Besides the conserved α-helix, other common secondary structure elements were not catch. However, prediction of Sic1-KID secondary structure with PHD, which provides about 70% accuracy [121–123], reveals two β-sheets and a short α-helix in addition to the long amphiphilic α-helix (Fig. 7.1b) [119]. These elements are indeed present in $p27^{Kip1}$-KID structure, suggesting that KIDs of $p27^{Kip1}$ and Sic1 might fold in a similar manner.

On this basis, Sic1-KID was built by homology modeling by using $p27^{Kip1}$-KID as a template, abridged from X-ray structure of the $p27^{Kip1}$/Cdk2/cyclin

Fig. 7.1 (continued) (amino acids 25–93) and Sic1-KID (amino acids 215–284) are shown (H, α-helix; E, β-sheet; c, coil). Residues involved in binding to either Cdk2 or cyclin A are *underlined*. Residues predicted to be in an α-helix within the Cdk2/cyclin A-interacting region are shown as *white-on-black* characters. Secondary structures deduced from X-ray structure of $p27^{Kip1}$-KID and expected for Sic1-KID are shown below the alignments. Reproduced with permission from Barberis et al. (2005) Biochem J 387(Pt 3):639–647. © the Biochemical Society [94]. (**b**) Secondary structure predictions of $p27^{Kip1}$ and Sic1 computed by PHD [119]. α-Helix, β-sheet, and coil are shown in *black*, *gray*, and *light gray*, respectively. The criteria for determining the secondary structure is the reliability index reported in the *Y*-axis (range 0–9): amino acids with a value ≥ 5 are predicted with a confidence of 82% [123]. KIDs of $p27^{Kip1}$ and Sic1 are indicated with a *double arrow*

A ternary complex (PDB entry 1JSU), and the amphiphilic α-helix of $p27^{Kip1}$ was mutated in silico to generate the predicted α-helix of Sic1. The Sic1-KID model was then docked onto Cdk2/cyclin A, refined by steps of conjugate gradients energy minimization [94] by using the GROMOS force field [120], and molecular interactions have been analyzed. Five amino acids within the amphiphilic α-helix of $p27^{Kip1}$-KID and Sic1-KID establish hydrophobic contacts with Cdk2/cyclin A [94, 118]. Interestingly, amino acid Leu41 of $p27^{Kip1}$-KID, part of the LF motif that recognizes cyclin A [118], is conserved in Sic1 (Ile229), suggesting that both Ckis share structural elements in the Cdk/cyclin-interacting region [94]. To build a yeast Sic1/Cdk/Clb complex by homology, Cdk1 kinase and Clb5 cyclin have been considered due to their functional homology to Cdk2 and cyclin A in driving entrance into S phase. The high sequence similarity of yeast and mammalian Cdks (77%) and cyclins (51%) allowed us to build a model of Cdk1/Clb5 by using Cdk2/cyclin A as a template and then to dock the Sic1-KID model. The Sic1/Cdk1/Clb5 complex was geometrically optimized by conjugate gradients energy minimization and molecular dynamics by using the consistent valence force field [124]. The interface between Sic1-KID, Cdk1, and Clb5 is characterized by steric and electronic contacts that allow the formation of a stable complex, as judged from hydrophobic contact analysis. The contacts between the LF domain of $p27^{Kip1}$-KID and cyclin A are conserved between an LV domain of Sic1-KID and Clb5 (Fig. 7.2 Table 7.1) [95, 119], in agreement with the fact that $p27^{Kip1}$ and Clb5 interact in vivo [125]. The interactions between the amphiphilic α-helix of $p27^{Kip1}$-KID and Cdk2/cyclin A are also observed between the Sic1-KID α-helix and Cdk1/Clb5 (Fig. 7.3) [95, 118, 119]. Here, amino acid Arg233 of Sic1-KID, although not hydrophobic, is threaded within Clb5 structure, making use of the alkyl moiety to effect specific hydrophobic interactions. Finally, amino acid Leu276 of Sic1-KID is located in a hydrophobic pocket of Cdk1 and have steric features to displace the ATP molecule bound and inhibit the kinase activity as it has been shown for the amino acid Tyr88 of $p27^{Kip1}$-KID on Cdk2 (Fig. 7.2 Table 7.1) [95,119]. The analyses indicate that, despite a low sequence similarity, KIDs of Sic1 and $p27^{Kip1}$ are structurally related and suggest that recruitment of a Cki on a hydrophobic pocket of a cyclin might be a conserved mechanism to realize Cdk/cyclin inhibition [125, 126].

4 Sic1 is a Functional Homologue to the Cki $p27^{Kip1}$

The structural findings underlie the role of kinase inhibitory domains (KIDs) in the regulation of Cdk/cyclin activity. The 3D structure of $p27^{Kip1}$-KID revealed that it is extended over the surface of Cdk2/cyclin A by forming hydrophobic contacts with regions on both cyclin and kinase [118, 127]. Moreover, isothermal titration calorimetry (ITC) [128] to determine thermodynamic parameters of $p27^{Kip1}$ binding to Cdk2/cyclin A and surface plasmon resonance (SPR) [129] to analyze kinetics of $p27^{Kip1}$ association/dissociation with/from Cdk2/cyclin A indicated that

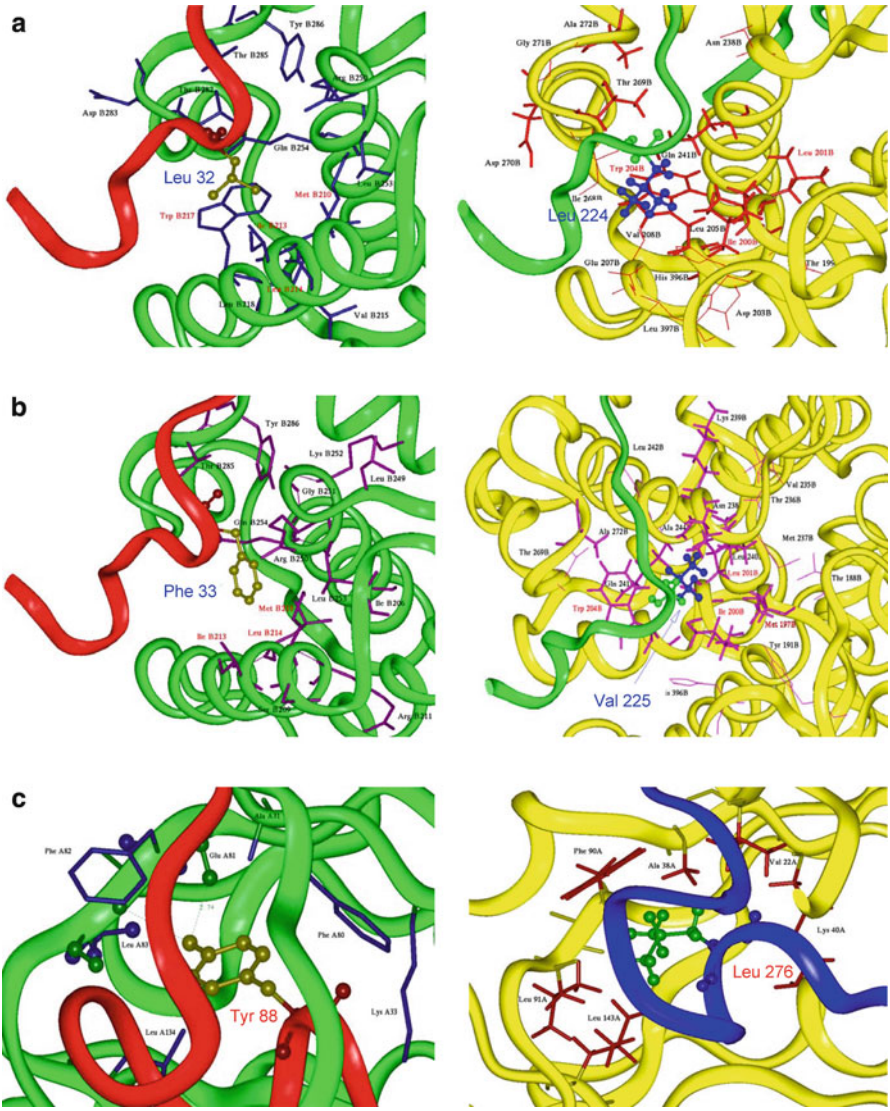


Fig. 7.2 (**a**) Interface contacts between the amino acid Leu of the $p27^{Kip1}$-KID LFG domain or Sic1-KID and the corresponding Cdk/cyclin complexes. (**b**) Interface contacts between the amino acid Phe of the $p27^{Kip1}$-KID LFG domain or Sic1-KID and the corresponding Cdk/cyclin complexes. (**c**) Interface contacts between Tyr88 of $p27^{Kip1}$-KID or Leu62 of Sic1-KID and the corresponding Cdk/cyclin complexes. Structural analysis was carried out using the InsightII software package (Biosym) and interactions evaluated within a range of 5 Å are shown ($p27^{Kip1}$-KID, red; Cdk2/cyclin A, *green*; Sic1-KID, *blue*; Cdk1/Clb5, *yellow*) [119]

Table 7.1 Summary of the interface contacts between (1) $p27^{Kip1}$-KID and Cdk2/cyclin A and (2) Sic1-KID and Cdk1/Clb5. Interactions between LF domain of $p27^{Kip1}$-KID (Leu32, Phe33) or LV domain of Sic1-KID (Leu224, Val225) and the cyclins and as well as interactions between Ckis and Cdks have been evaluated within a range of 5 Å (top) or 7 Å (bottom) [119]. A high conservation in amino acid type is observed between Ckis, cyclins, and Cdks: bold, essential contacts; italic, stabilizing interactions; other amino acids, other interactions

$p27^{Kip1}$-KID/cyclin A		Sic1-KID/Clb5		$p27^{Kip1}$-KID/Cdk2	Sic1-KID/Cdk1
Leu32	Phe33	Leu224	Val225	Tyr88	Leu276
Met210	**Met210**		**Met197**	**Lys33**	**Lys40**
Ile213	**Ile213**	**Leu200**	**Ile200**	*Ala31*	*Ala38*
Leu214	**Leu214**	**Ile201**	**Ile201**	*Phe80*	*Phe88*
Trp217		**Trp204**	**Leu204**	**Glu81**	**Glu89**
Arg250	Arg250		*Leu240*	*Phe82*	*Phe90*
Leu253	*Leu253*	*Gln241*	*Gln241*	*Leu83*	*Leu91*
Gln254	*Gln254*			*Leu134*	*Leu143*
Leu218	Gly251	Leu205	Asn238		
Thr282	Lys252	Thr269	Lys239		
Asp283		Asp270			
Thr285	Thr285	Ala272	Ala272		

$p27^{Kip1}$ tightly binds Cdk2/cyclin A (nanomolar) via a sequential mechanism. In fact, it occupies a conserved hydrophobic pocket for substrate recruitment on cyclin A [119, 130], then it binds to the N-terminal lobe of Cdk2 flattening it out and disrupting the active site, finally inserting itself into the ATP binding pocket and blocking ATP binding to Cdk2 [131]. The relevance of KID is highlighted by analysis of knock-in mice with a $p27^{Kip1}$ variant that lacks the Cdk inhibitory function, $p27^{Kip1}$ (CK−), which revealed that $p27^{Kip1}$ (CK−/CK−) displayed tumor development as the $p27^{Kip1}$ null (−/−) mice and a range of hyperplasia and neoplasia suggesting that $p27^{Kip1}$ (CK−), which localizes in the cytoplasm, functions as an oncogenic protein [132]. Therefore, addressing the molecular mechanism that Cki develops to inhibit Cdk/cyclin activity is undoubtedly challenging to understand how the timing of cell cycle regulation is accomplished.

To test the hypothesis that Sic1-KID is able to interact productively with Cdk/cyclin complexes, as predicted by structural analysis [94, 119], interactions of Sic1 with mammalian Cdk2 and cyclin A (alone or in complex) purified from baculovirus have been tested by SPR. Sic1 protein was covalently coupled to a carboxymethylated dextran surface by using amine-coupling chemistry [133] and association/dissociation of Sic1 was determined by fluxing several concentrations of Cdk2, cyclin A, and Cdk2/cyclin A. The analysis indicated that the affinity of Sic1 for Cdk2 was very low (dissociation equilibrium constant, $K_D = 10^{-5}$), while binding of Sic1 to cyclin A and to Cdk2/cyclin A was favorable [95]. In particular, a strong interaction was observed between Sic1 and Cdk2/cyclin A ($K_D = 10^{-7}$) compared to Sic1 and cyclin A ($K_D = 10^{-6}$). These findings suggest that Sic1 realize its inhibitory function by interacting first with the cyclin and then extending on the surface of the Cdk/cyclin complex to reach and inhibit the binding site on the kinase [95]. Consistently, Sic1 was shown to strongly inhibit both yeast and

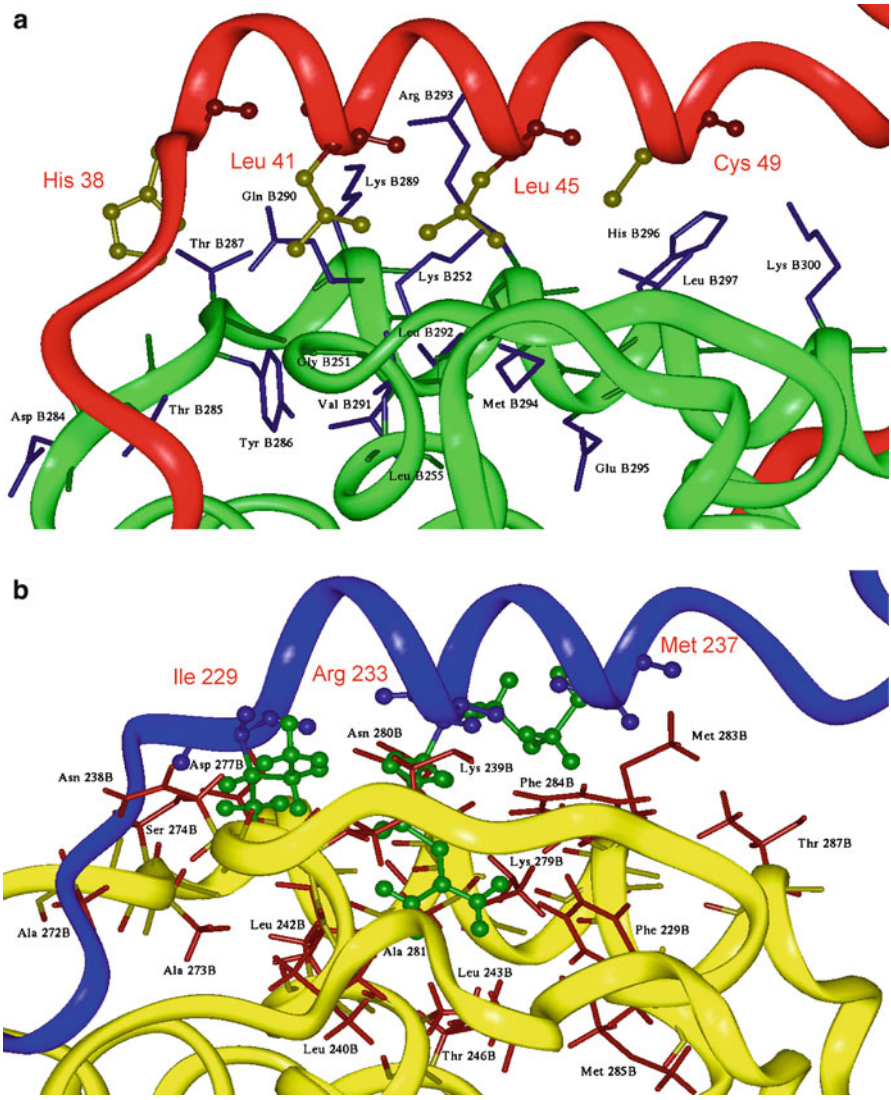


Fig. 7.3 (**a**) Interface contacts between the amphiphilic α-helix of $p27^{Kip1}$-KID (amino acids 38–60) and Cdk2/cyclin A. (**b**) Interface contacts between the amphiphilic α-helix of Sic1-KID (amino acids 226–248) and Cdk1/Clb5. Structural analysis was carried out using the InsightII software package (Biosym) and interactions within a range of 7 Å (**a** and **b**) or 5 Å (**c**) are shown ($p27^{Kip1}$-KID, *red*; Cdk2/cyclin A, *green*; Sic1-KID, *blue*; Cdk1/Clb5, *yellow*) [119]

mammalian Cdk/cyclin activities by a similar double-step inhibitory mechanism [95,96]. The physiological relevance of these results has been addressed by showing that both Sic1 and the mammalian Cki $p27^{Kip1}$ rescued the phenotype of a *sic1Δ* strain [95]. Altogether, these findings indicate that Sic1 is functionally related to $p27^{Kip1}$, employing a conserved mechanism of inhibition on Cdk/cyclin activity.

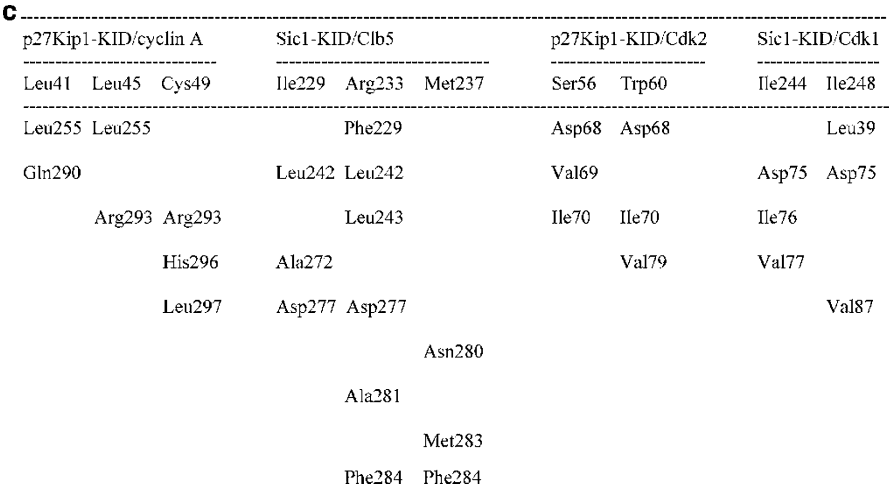


Fig. 7.3 (continued)

5 Cki Phosphorylation: Hallmark of Cell Cycle Timing

Despite $p27^{Kip1}$-KID reveals secondary structure elements that tightly insinuate within Cdk2/cyclin A architecture, in particular the nascent amphiphilic α-helix [131, 134], secondary structure and disorder prediction indicate that the Cki is mainly disordered [135]. Heat-resistant assay, hydrodynamic analysis, proteolysis, ITC, circular dichroism (CD) and NMR spectroscopy showed that $p27^{Kip1}$ is largely disordered [135–139]. In particular, the latter revealed that portions of the isolated $p27^{Kip1}$ sequence show secondary structure in solution [138]. Thus, $p27^{Kip1}$ belongs to the intrinsically disordered proteins (IDPs). Many proteins that play important cellular functions, i.e., regulation of cell division, transcription and translation, phosphorylation, signal transduction [140, 141], are characterized by sequence domains lacking secondary or tertiary structure and, thus, classified as IDPs. Furthermore, 79% of human cancer-associated proteins have been classified as IDPs [142]. Considering that IDP sequences are generally exposed to the solvent, a large number of sites are accessible for post-translational modification, which regulate function, localization, and stability. $p27^{Kip1}$ is mainly regulated by phosphorylation directed by various signal transduction pathways, which controls timing of Cdk/cyclin activity by weakening the Cki inhibitory activity [143–145]. Moreover, tumorigenesis associated to $p27^{Kip1}$ has been described due to its phosphorylationinduced cytoplasmic localization [146–150]. Therefore, disorder and flexibility of $p27^{Kip1}$ enable structural fluctuations and phosphorylation events that regulate its turnover at the G1/S transition during cell cycle control.

5.1 Regulation of Sic1 Activity by Phosphorylation

As aforementioned, despite efforts from different groups, attempts to crystallize Sic1 have failed. However, Sic1 has been shown to be a disordered protein both in its free state and when bound to Cdc4 [151], which acts as ubiquitin–protein ligase directing ubiquitination of the phosphorylated Sic1 [45, 46]. The complex showed a mixture of different conformations shifting around in a dynamic equilibrium, and each of the six phosphate groups on Sic1 needed for its recognition from the proteasome have been found to occupy the single Cdc4 pocket, one after the other [151]. Further analyses highlighted that although Sic1 is disordered when bound to Cdc4, it maintains a compact structure that keeps the phosphate groups close together to form an electrostatic field that glues Sic1 to Cdc4 [152]. Complementary biophysical methods have also been applied to the study the isolated Sic1 in solution. Sequence analysis, gel filtration, CD, electrospray-ionization mass spectrometry (ESI-MS), and limited proteolysis showed that Sic1 is mainly disordered with an intrinsic propensity for ordered structure in its C-terminal region in correspondence of the kinase inhibitory domain (KID) [153]. Sic1 can, therefore, be classified as an IDP like $p27^{Kip1}$. Recently, studies of limited proteolysis, CD, NMR, and nano-ESI-MS analysis showed a modular organization for Sic1 being its C-terminal region is relatively more compact than the N-terminal one, with the boundary of the C-terminal lying close to the amino acid Trp186, suggesting that it is possible to recognize structural domains in an IDP [154, 155]. Moreover, Fourier-transform infrared (FT-IR) spectroscopy and ion-mobility (IM) measurements revealed that the isolated Sic1-KID retains dynamic helical structure and populates collapsed states of different compactness [156].

As shown for $p27^{Kip1}$, post-translational modification via phosphorylation can play a role modulating Sic1 conformational transitions [157]. Several signaling pathways promote Sic1 phosphorylation regulating its stability, thus timing of the G1/S transition. Activation of the Hog1 pathway due to high osmolarity results in a cell cycle arrest in G1 by phosphorylation and, thus, stabilization of Sic1 [158, 159]. Moreover, inhibition of the TOR pathway by rapamycin leads to phosphorylation and stabilization of Sic1, as shown for $p27^{Kip1}$ [160], which accumulates into the nucleus both in glucose and ethanol-grown cells [161]. In addition, other kinases are involved in Sic1 phosphorylation: Pho85 is required for the prompt degradation of Sic1 [162] and Ime2 is necessary but not sufficient to promote Sic1 destruction during sporulation [163]. An important regulator of cell cycle progression is CK2, a ubiquitous, highly pleiotropic and constitutively active serine–threonine kinase conserved in all eukaryotes [164], which phosphorylates both Sic1 and $p27^{Kip1}$ [99, 165, 166]. Moreover, Sic1 accumulation is observed following CK2 inactivation, inhibiting the Cdk1/Clb5 complex, therefore effectively blocking the G1/S transition [167].

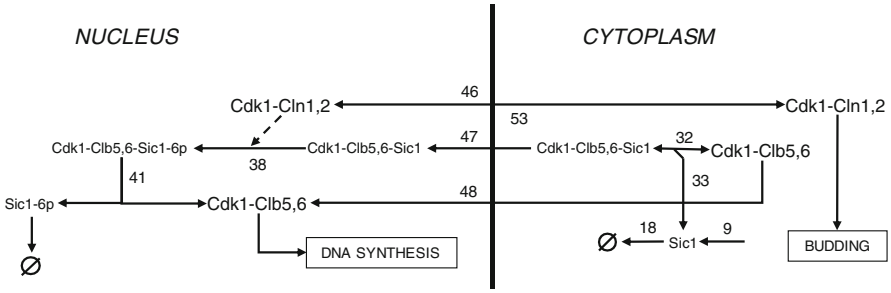


Fig. 7.4 Schematic representation of Cdk1/Clb5 activation regulating the G1/S transition in budding yeast. Sic1 binds to Cdk1/Clb5,6 (32) and triggers the complex into the nucleus (47), where it is degraded after Cdk1/Cln1,2-mediated phosphorylation (38, 41). Modified from Barberis et al. (2007) PLoS Comput Biol 3(4):e:64 [92]

5.2 Sic1 Phosphorylation by CK2: Mechanism of S Phase Timing

Ckis carry out the inhibitory function by formation of ternary complexes with their target cyclin and kinase. The temporal dynamic of assembly and disassembly of these complexes during cell cycle progression determines at which time and in which cellular compartment regulatory phosphorylation events take place. As aforementioned, subcellular localization of Cki and other cell cycle proteins is recognized to be a major factor that regulates cell cycle transitions, since altered localization of Ckis is linked to cancer aggressiveness. Nevertheless, for reasons of simplicity, models of cell cycle regulation do not generally consider this aspect. The need to incorporate this fundamental regulatory feature stimulated us to generate a computational model considering the localization of $p27^{Kip1}$ and kinase complexes involved in the regulation of the G1/S transition in mouse fibroblasts [74], following the mechanism of Cdk/cyclin inhibition by $p27^{Kip1}$ and phosphorylated $p27^{Kip1}$ [168]. The model recapitulates events from growth factor stimulation to S phase onset following phosphorylation states associated to activation or deactivation of $p27^{Kip1}$ and kinase complexes in nucleus or cytoplasm [74].

In parallel to the mammalian network, we developed a detailed mathematical model of the G1/S transition in budding yeast [92], taking into account the nucleo/cytoplasmic localization of key players and the carbon source regulation of Sic1 to promote nuclear import of the Cdk1/Clb5 complex [49]. As for the G1/S network in mouse fibroblasts, the model was implemented by a set of 34 ODEs, 32 species and 67 kinetic parameters [169] describing the temporal change in concentration of key players and as well as phosphorylation states of Sic1 and kinase complexes regulating entrance into S phase (Fig. 7.4) [49]. CK2-mediated phosphorylation on amino acid Ser201 of Sic1 has been recognized to alter timing of the G1/S transition by affecting Sic1 affinity for Cdk1/Clb5 [95, 99, 166, 167],

and mutations that impair (Ser201/Ala) or mimic (Ser201/Glu) phosphorylation by CK2 affect the coordination between cell growth and cell cycle progression in vivo [99, 119]. However, analyses of the mutants did not reveal appreciable effects on the conformation of isolated Sic1 [153], as observed instead for $p27^{Kip1}$ [165]. These contradictory data motivated to investigate the physiological role of Sic1 phosphorylation by CK2 not in the isolated form but when bound to Cdk1/Clb5,6. Many proteins lack rigid 3D structure, existing as dynamic ensembles of inter-converting conformations and acquiring an ordered structure when binding to specific intracellular partners [170] or by functional regulation via post-translational modifications, as shown for $p27^{Kip1}$ [135]. Therefore, more detailed structural information is needed to interpret the effect of this phosphorylation on the interaction between Sic1 and Cdk1/Clb5,6. In addition, difficulty to estimate in vivo phosphorylation kinetics encouraged us to estimate realistic values to include as kinetic constants in the mathematical model of the G1/S transition. Real-time measurement by SPR using immobilized Sic1 showed that it interacts with catalytic (α) and regulatory (β) subunits of CK2, being the strength of the binding in the same range as compared to the CK2β/$p27^{Kip1}$ interaction [95, 165].

Moreover, Sic1 was phosphorylated by the CK2 with an apparent K_M = 460 nM, value comparable to the K_M for the CK2-mediated phosphorylation of $p27^{Kip1}$ (467 nM) [165]. In order to test the hypothesis that Ser201 phosphorylation on Sic1 could be relevant for interaction with Cdk/cyclin complexes, a model peptide encompassing amino acids 192–216 of Sic1 was synthesized either as such or with Ser201 replaced by phosphoserine. Both peptides were covalently coupled to carboxymethylated dextran surfaces by using amine-coupling chemistry and binding with mammalian Cdk2/cyclin A was examined. Interestingly, Sic1 peptide encompassing Ser201 was bound more strongly to Cdk2/cyclin A in its phosphorylated than in its nonphosphorylated form [95]. Consistently, Sic1 fully phosphorylated on Ser201 by CK2 was shown a stronger inhibitor of both yeast and mammalian Cdk/cyclin activities than the unphosphorylated protein, suggesting a possible regulatory role of CK2 phosphorylation on Sic1 activity. The very high negative charge density of the Sic1 phosphor-acceptor site prompted us to investigate whether basic patches might be present on the surface of Cdk1/Clb5 in positions compatible with a direct interaction. Homology modeling techniques assume that proteins interact using two relatively large interfaces, however, it is well-established that many interactions, particularly those of lower affinity, are mediated by one domain binding to a small stretch of polypeptide in another protein, i.e., small sequences characteristic of a consensus phosphorylation site. These interaction are difficult to detect and study computationally or experimentally because they often involve unstructured parts of the polypeptide chain that become ordered only on binding [170]. Therefore, restrained molecular dynamics have been carried out to dock a 23-amino acid-long Sic1 region comprising the CK2 consensus sequence QES^{201}EDEED (amino acids 192–214) of the modeled Sic1-KID on Cdk1/Clb5 (Fig. 7.5a), and interactions with Cdk1/Clb5 were investigated by energy minimization [95, 119]. Analysis of the surface electrostatic potential of Cdk1/Clb5 allowed localization of clusters of highly positively charged residues

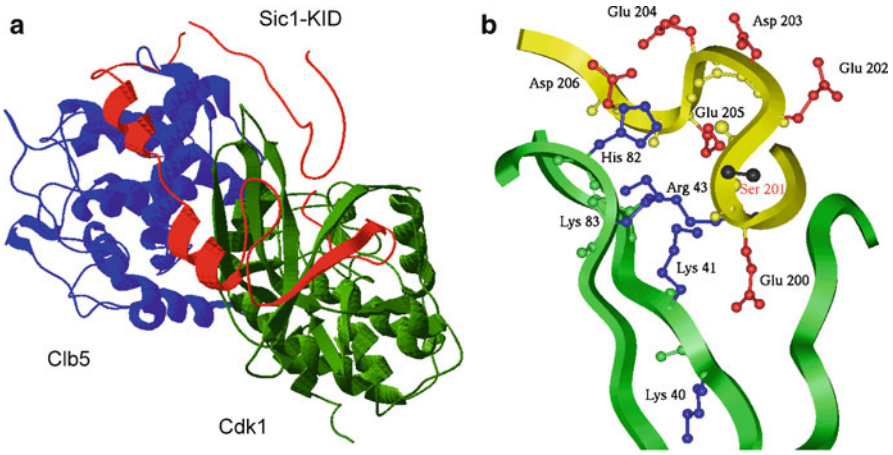


Fig. 7.5 (**a**) Sic1-KID 3D model was docked on Cdk1/Clb5 by using the $p27^{Kip1}$/Cdk2/cyclin A X-ray structure as a template and optimized using energy minimization (Sic1-KID, *red*; Cdk1, *green*; Clb5, *blue*). (**b**) Interface contacts between the CK2 consensus site on Sic1 and Cdk1. Structural analysis was carried out using the InsightII software package (Biosym) and interactions within a range of 5 Å are shown (Sic1-KID: backbone – *yellow*, amino acidic lateral chains – *red*; Cdk1: backbone – *green*, amino acidic lateral chains – *blue*). Reproduced with permission from Barberis et al. (2005) Biochem Biophys Res Commun 336(4):1040–1048 [95, 119]

on Cdk1, which have proper electrostatic characteristics to interact productively with the negatively charged CK2 consensus sequence centered on Sic1 (Fig. 7.5b) [95, 119].

Biochemical and structural analyses suggest that CK2 may play a role in the regulation of Sic1 activity by phosphorylation of amino acid Ser201. The phosphorylation could induce long-term rear-rangements of the 3D structure of Sic1-KID, as reported in the literature [171, 172], remodeling Cdk1 surface and altering the interaction with Cdk1/Clb5, ultimately affecting the G1/S transition and, thus, entrance into S phase. To investigate dynamic consequences of change in the affinity of Sic1 for Cdk1/Clb5 for the timing of S phase onset, different kinetic constant values for this binding have been tested in the mathematical model of the G1/S transition in different nutritional setups. Considering that binding between two proteins can be affected by cellular growth conditions and that protein phosphorylation alters binding to another protein [171], we assumed that a poor carbon source (i.e., ethanol) is associated with a low level of phosphorylation, whereas a rich carbon source (i.e., glucose) to a high level of phosphorylation, and kinetic parameters have been chosen to obtain simulated dynamics close to the one measured experimentally [92]. To assess the effect of changing growth conditions from glucose to ethanol media, input parameters such as growth rate and initial levels of network key players have been altered. However, to obtain a good fitting between experimental and computational dynamics, the affinity observed for unphosphorylated Sic1 to Cdk1/Clb5 has been introduced in the simulated ethanol condition (i.e., reduction by two orders of magnitude compared to

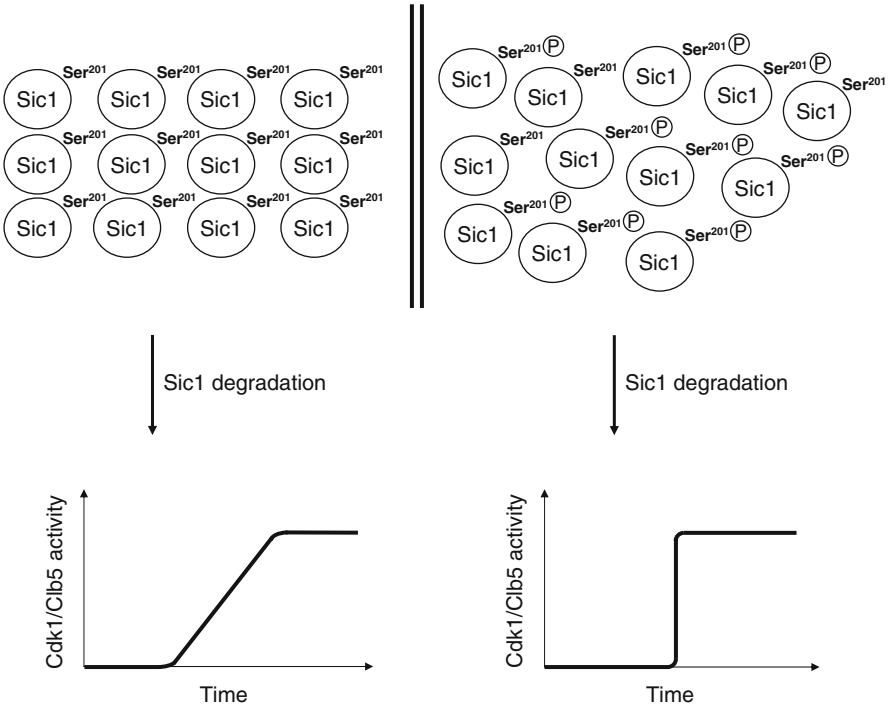


Fig. 7.6 Model of DNA replication onset. Increased Sic1 affinity to Cdk1/Clb5, which mimics CK2-mediated phosphorylation on Ser201 of Sic1, leads to a delay in the entrance into S phase due to abolishment of the kinase activity and to a larger accumulation of Cdk1/Clb5. When Sic1 is degraded, a huge amount of Cdk1/Clb5 is promptly available to activate DNA synthesis

the simulated glucose dynamics) [92, 95]. Taken together, biochemical, structural, and computational analyses generated the prediction, to be tested experimentally, that Sic1 might have a lower binding affinity for Cdk1-Clb5,6 in ethanol-grown cells compared with the one in glucose-grown ones, and that phosphorylation in rich medium is dependent on CK2. Interestingly, evidence that this can be true can be surmised from a recent study on the *ck1Δck2Δ* temperature-sensitive (ts) double mutant [167]. This mutant grows normally at 25°C (functional CK2), Sic1 level is low, and Cdk1-Clb5 activity is high (due to low Sic1). Contrarily, at 37°C the mutant, and thus CK2, is inactive, Sic1 level is high, and Cdk1-Clb5 activity is abolished (due to high Sic1), even if comparable levels of Clb5 are observed. This situation suggests that the condition at 37°C could be comparable to growth in ethanol medium, where high Sic1 levels observed experimentally [49] might be due to decrease in CK2 kinase activity on Sic1. This scenario implies that the phosphorylated state of Sic1 could influence its localization and, therefore, timing in which the S phase onset is accomplished, ensuring that no premature origin licensing takes place by strongly inhibiting Cdk1/Clb5. Licensed origins could be then activated on schedule by providing higher Cdk activity to start DNA replication after Sic1 proteolysis (Fig. 7.6).

6 Sic1 Regulates Timing of Cdk1/Clb Activities

Both in budding yeast and in higher eukaryotes, genomic instability occurs when the G1/S transition is deregulated and cells enter into S phase prematurely. This acquired mutability is critical since a majority of genes mutated in human cancers influence the G1/S transition [173]. The initiation of DNA replication in budding yeast is regulated by an irreversible switch in which Sic1 is degraded at the S phase onset [33]. The activation on schedule of Cdk1/Clb5,6 and of other waves of Cdk1/Clb activity, i.e., Cdk1/Clb3,4 and Cdk1/Clb1,2, from S to M phases is strictly related on disappearance of Sic1 [34,35]. The precocious activation of Cdk1/Clb5,6 observed in a *sic1Δ* mutant initiates prematurely DNA replication from fewer origins, phenomena called sparse origin firing [50], and severe genome instability and chromosome rearrangements occurs [50, 52]. However, the mechanism by which control of cell cycle timing is lost is not clear.

To investigate genomic instability in budding yeast, we studied balance between Sic1 and Cdk1/Clb5,6 in activating replication origins at the entrance into S phase [92,96], and considered Sic1 both a stoichiometric inhibitor of Cdk1/Clb complexes [33,94] and a promoter of Cdk1/Clb5,6 entry into the nucleus [92], as shown experimentally [49]. We described origins activation with a stochastic model considering the rate of firing dependent on nuclear Cdk1/Clb5,6 availability and observed an early firing of replication origins in a *sic1Δ* mutant compared to the wild type due to a precocious activation of Cdk1/Clb5 [92, 96], as experimentally observed [50]. This suggests that appearance and disappearance of Sic1 regulates replication origins by controlling the timing of Cdk1/Clb5,6 activity. Whether the role of Sic1 as a timer of cell cycle transitions is realized mainly through inhibition of Cdk1/Clb5,6 activity or via direct binding to or regulating other kinase activities or components of the DNA replication machinery is still not fully understood. Thus, to investigate whether Sic1 may function as a timer in coordinating the staggering behavior of phase-specific Cdk1/Clb complexes during cell cycle progression, a combined computational and experimental approach has again been employed. Interactions of Sic1 with one or more Cdk1/Clb complexes have been drown with CellDesigner [174] and implemented by a set of 11 ODEs describing the dynamic behavior of Cdk1/Clb complexes in time (Fig. 7.7) [97], and the characteristic pattern know as waves of cyclins [25,175,176] investigated. This modular network is small enough for an accurate mathematical modeling. In fact, when a sufficient small network is considered and kinetic parameters are available, or when parameters are unknown but components ad reactions are known, kinetic models have been successfully used to predict signaling properties [177]. Computational analysis revealed that temporal coordination of Clb cyclins appearance, and their oscillation-like behavior, is observed only when Sic1 binds to all Cdk1/Clb complexes [97]. Accordingly, associations of Sic1 with all Clb cyclins have been detected in high throughput genome-wide screenings for complexes [47,125,178–184]. Therefore, models have been tested for internal consistency by computational analyses, i.e., sensitivity analysis, and for external consistency by experimental validation via protein–protein

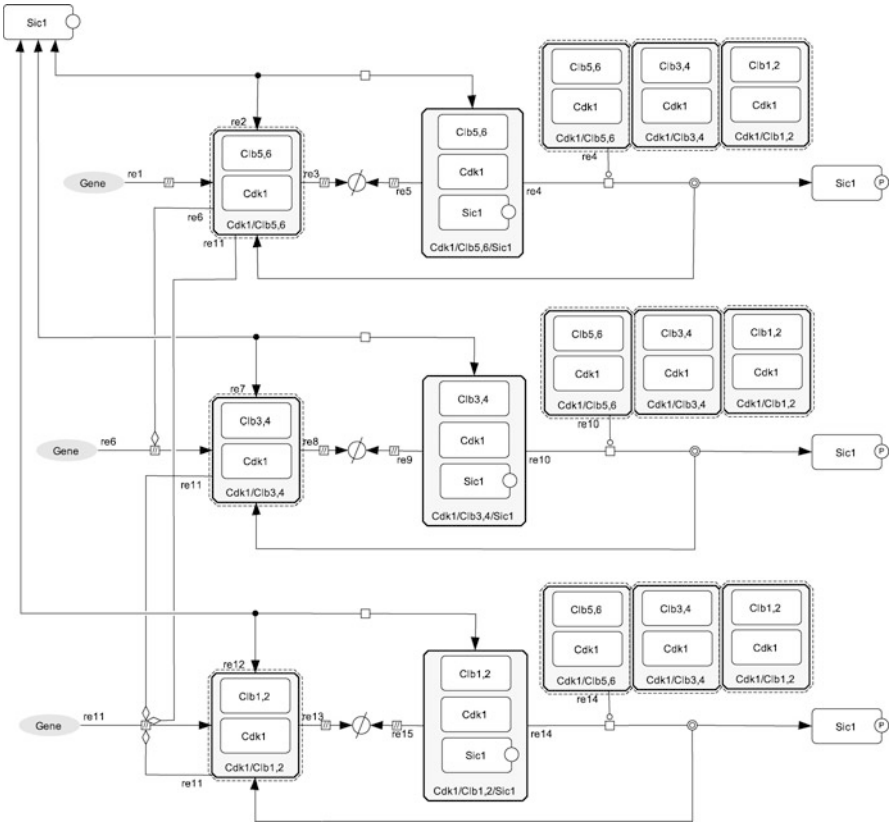


Fig. 7.7 Schematic model of Cdk1/Clb regulation. After production of Cdk1/Clb5,6 (re1), Sic1 binds forming the Cdk1/Clb5,6/Sic1 complex (re2). Sic1 is degraded primarily by Cdk1/Cln1,2 (not shown) and by Cdk1/Clb activities (re4) and Clb5,6 is degraded in both Cdk1/Clb5,6 (re3) and Cdk1/Clb5,6/Sic1 (re5) complexes. Cdk1/Clb5,6 activates Cdk1/Clb3,4, in addition to its basal production (re6), and Sic1 binds to Cdk1/Clb3,4 forming the Cdk1/Clb3,4/Sic1 complex (re7). Sic1 is degraded by Cdk1/Clb activities (re10) and Clb3,4 is degraded in both Cdk1/Clb3,4 (re8) and Cdk1/Clb3,4/Sic1 (re9) complexes. Cdk1/Clb3,4 activates Cdk1/Clb1,2 together with Cdk1/Clb5,6, in addition to its basal production (re11). Sic1 binds to Cdk1/Clb1,2 forming the Cdk1/Clb1,2/Sic1 complex (re12). Sic1 is degraded by Cdk1/Clb activities (re14) and Clb1,2 is degraded in both Cdk1/Clb1,2 (re13) and Cdk1/Clb1,2/Sic1 (re15) complexes. Cdk1/Clb1,2 activates itself by a positive feedback loop (re11) [97]

interaction techniques. Global sensitivity analysis with a Monte Carlo approach has been employed to investigate whether kinetic parameter values influence time delay between Clb cyclins. Random sampling with 10,000 kinetic parameter sets has been carried out by varying them between 0.1 and 10-fold of their initial values. By comparing networks where Sic1 binds to one or more Cdk1/Clb complexes, any change of parameters affects the delay of Clb appearance only when Sic1 binds to all kinase complexes [97].

Considering that regulation of time delays between Clb cyclins is apparently triggered by interaction of Sic1 with all Cdk1/Clb complexes, binding of Sic1 with all Clb cyclins has been, therefore, investigated experimentally. Interactions with Clb2 and Clb5 are well-established, but association to Clb3 (180–182) and Clb4 (183) has been shown only in high throughput genome-wide screenings for complexes and never validated independently. In vitro analyses, yeast two-hybrid and GST pull-down, revealed interactions between Sic1 and all Clb cyclins [97]. Moreover, Förster resonance energy transfer (FRET) via fluorescence lifetime imaging microscopy (FLIM) has been employed to investigate Sic1/Clb interactions in living yeast cells. By using this powerful technique that detects close co-localization of fluorescent proteins and provides high spatial and temporal resolution (nanoseconds), occurrence of FRET was measured by monitoring the change in Sic1 lifetime in the presence and absence of Clb cyclins [185, 186]. Association of Sic1 to each Clb cyclins subtype has been observed and different FRET efficiencies were measured [187]. These findings, together with the fact that Sic1 is a substrate of Clb3-associated kinase activity, as shown for both Clb5 and Clb2 [45, 46, 188], support the hypothesis that Sic1 interacts with all Cdk1/Clb complexes throughout cell cycle progression. However, despite their homology [189, 190], distinct Clb cyclins might target Sic1 preferentially to enable its function [187].

The above results and the fact that Sic1 levels are observed throughout the cell cycle [98, 99] inspired to follow Sic1 and Clb cyclins levels in G1-synchronized yeast cells by elutriation, to demonstrate that Sic1 does not interact only with Clb5,6 at the G1/S transition and with Clb1,2 regulating mitotic exit [33, 39, 191] but also with Clb3,4 during the temporal window in which its levels should decrease. Temporal dynamics of wild type cells showed the characteristic periodicity of Clb cyclins levels, with their on schedule appearance and disappearance one after the other, and the coexistence of Sic1 and all Clb cyclins including Clb3,4 overall cell cycle progression [97]. However, an interesting result has been shown perturbing the structure of the mathematical model by testing Cdk1/Clb regulation in the absence of Sic1, mimicking a *sic1Δ* mutant. In this scenario, computational simulations predicted an abolishment of Clb cyclins waves with their levels reaching a different plateau over the simulation time. Strikingly, the prediction finds its validation in elutriated *sic1Δ* cells, which completely loose timing and regulated periodicity of Clb cyclins appearance although proceeding into the replicative state, revealing that both Clb3 and Clb2 arise at the beginning of G1 phase as observed for Clb5 with levels that progressively increase to reach a different plateau [97]. This result agrees with the fact that a *sic1Δ* strain accumulates Clb5 in early G1 phase generating high Cdk1/Clb5,6 activity, therefore promoting precocious DNA replication [50]. Consequently, an uncontrolled temporal pattern of Clb cyclins may lead to cells that segregate not completely replicated chromosomes, resulting in extensive chromosome loss [52]. Altogether, these findings suggest that Sic1, through a feed-forward regulation, triggers waves of Clb cyclins and timing of their appearance, therefore controlling Clb-associated kinase activities. Moreover, the hypothesis that heterodimer formation of various Sic1/Clb pairs can differ according

to the abundance of a certain Clb cyclin [192, 193] is likely to be relevant for their localization and temporal window of activity.

Further computational and experimental analyses that we performed have shown that waves of Clb2, Clb3, and Clb5 levels can be still observed, although temporally delayed, after constitutive expression of a non-degradable form of Sic1 (*SIC1-OP*) [92, 97], as recently envisioned [100]. These data suggest that stable Sic1 transiently blocks Cdk1/Clb activation, but that ultimately the total level of these complexes increases above the Sic1 level. This is due to the fact that oscillations in Sic1 level are enough to trigger the feed-forward loop necessary for the switching of Cdk1/Clb complexes between states of high and low concentrations [194]. Computational simulations abolishing virtually the degradation of Sic1 by any of the Cdk1/Clbs (Fig. 7.7, reaction rates 4, 10, and 14) reproduce the scenarios in which cells carrying *SIC1-OP* are lethal in the absence of either *CLB2*, *CLB3*, or *CLB5* genes [100], showing a lethal phenotype of Sic1 over-expression in backgrounds with single deleted Clb cyclins – compared to the viable phenotype associated to strains with the sole deletion of each Clb subtype – due to a not proper timing of accumulation of the remaining ones (Fig. 7.8). This results clearly reflects the specific activity that Clb cyclins play at various cell cycle stages [192, 195] and indeed indicates that our computational predictions are valid, supporting a role of Sic1 in the regulation of Cdk1/Clb complexes.

7 Conclusions and Outlook

The aim of systems biology is to obtain a quantitative description of cellular functions to elucidate complex human diseases such as cancer. Due to the complexity of human cells, model systems, e.g., budding yeast, are used for medical research. In this organism, complete understanding of cell cycle regulation is not trivial and many detailed molecular mechanism are still unknown.

The observation that cells replicating their chromosomes from a sub-optimal number of origins are karyotypically unstable is important to understand tumorigenesis, in agreement with the fact that G1/S regulators are mutated in cancer [173]. The modular bottom-up systems biology approach here presented has been useful to investigate the role of cell cycle players whose deregulation leads to abnormal replication dynamics. We have employed a multiscale modeling to elucidate a mechanism by which the Cki Sic1 controls Cdk1/Clb activities in budding yeast by integrating results derived from structural, biochemical, cell biological, and computational studies. Biochemical studies and molecular dynamics simulations helped us to decipher the role of Sic1 intrinsic structure in molecular recognition, and computational modeling predicted physiological properties related to Sic1 function, which have been successfully validated experimentally. Thus, our approach can be valuable for determining the specific mechanism of Cdk1/Clb regulation. A more complete description of Sic1 role at the G1/S transition will not only require to resolve the molecular details of Sic1/Clb interactions in living

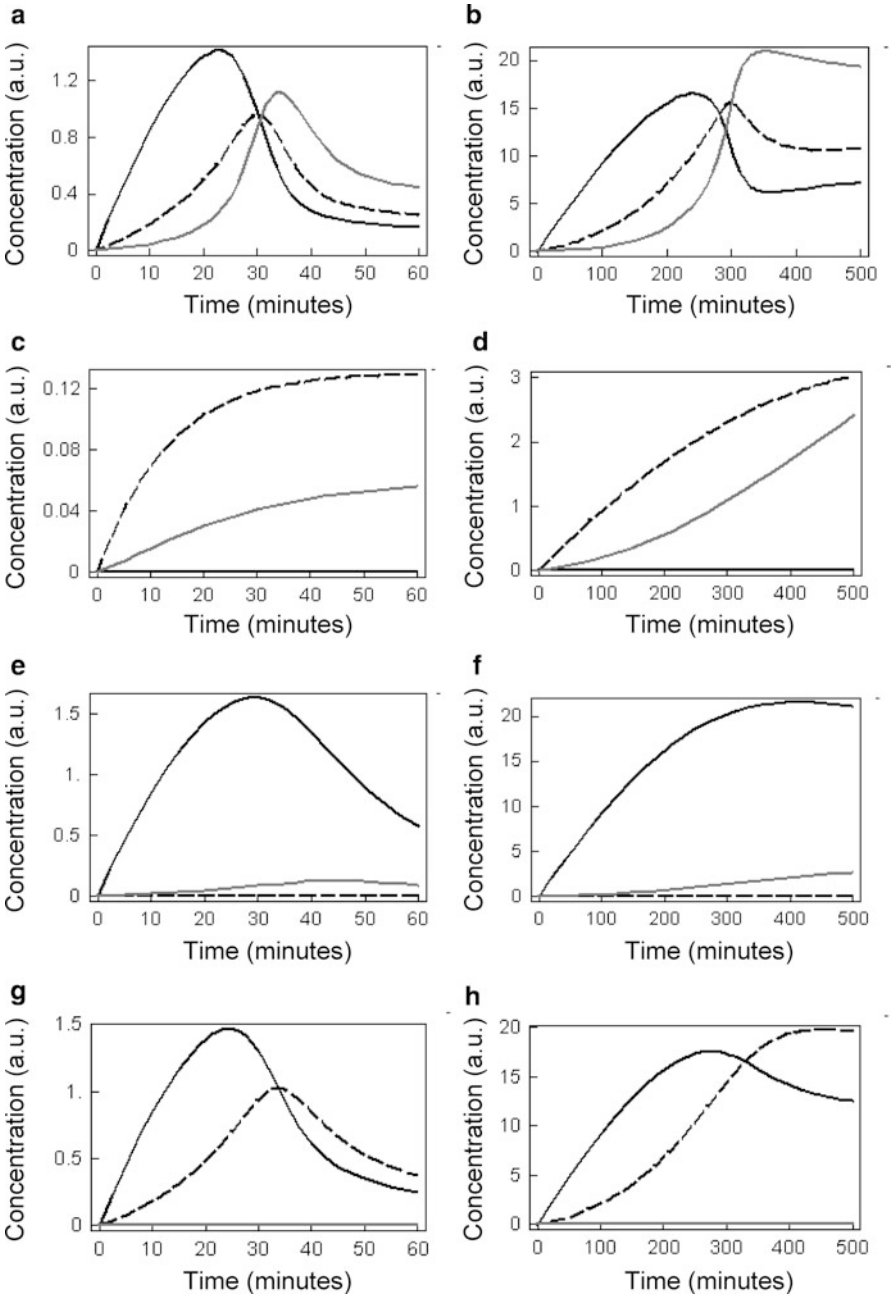


Fig. 7.8 Over-expression of *SIC1-OP* leads to lethality in backgrounds with deletion of each Clb cyclins subtype. Simulations of Clb cyclins wave formation in a wild type background without (**a**) or with *SIC1-OP* (**b**) are shown at different timing (wild type, 60 min; *SIC1-OP*, 500 min). Single deletions in each Clb cyclins subtype without (*clb5*, *6Δ*, C; *clb3*, *4Δ*, E; *clb1*, *2Δ*, G) or with *SIC1-OP* (*clb5*, *6Δ*, D; *clb3*, *4Δ*, F; *clb1*, *2Δ*, H) are also shown. Protein levels are marked in different colors (Clb5,6, *black*; Clb3,4, *dotted black*; Clb1,2, *gray*)

cells but also the extension to the transcriptional regulation of the phase-specific Clb cyclins, currently under investigation. Moreover, a stochastic model that addresses Sic1 transcription and the resulting noise on Sic1/Clb5 balance at the G1/S transition has been developed [84]. Computational simulation revealed that an increased amount of *SIC1* mRNA leads to an amplified dispersion of Sic1 protein levels, suggesting that both Sic1 protein and mRNA levels are critical to set the timing of Sic1 downregulation and, therefore, S phase onset [84].

At the molecular level, considering that Clb cyclins have a high sequence similarity, structural studies could be pursued to investigate whether stabilizing residues predicted are conserved in all Sic1/Clb interactions. However, interaction details have to be necessarily considered within the cellular context, where Clb cyclins are expressed at different times during cell cycle progression, at variable protein levels and in different cellular compartments [192, 193]. Computational and experimental results have to take into account that an in vitro interaction might have no in vivo meaning. Therefore, strength in the affinity of protein–protein interactions is functionally relevant for a physiological cellular response. For example, FLIM–FRET technique provides insights into binding affinities, however accurate values are difficult to obtain experimentally and to be predicted theoretically. The development of systems to measure kinetic parameters for protein–protein interactions is certainly a critical challenge in systems biology, to combine structural details, affinity data, and computational network analyses. Structures can also give information on the order of events in a network, by indicating for example which interactions cannot occur simultaneously due to a common binding interface, e.g., the binding of Sic1 to the phase-specific Clb cyclins. As soon as more structural details become available, cell cycle networks centered around the regulation of Clb cyclins by Sic1 will be more realistic. However, small modular networks such as the one we have described that satisfy inherent properties or explain physiological behaviors can predict biochemical activities and new functional interactions, e.g., the binding of Sic1 to Clb3,4, before carrying out an experimental validation. Identification of the molecular structure of a functional module requires hypothesis-driven experiments and computational modeling to elucidate design principles and to describe its temporal dynamics, therefore being a powerful strategy to improve understanding of cell cycle regulation.

Acknowledgements MB is supported by grants from the European Commission ENFIN (contract number LSHGCT-2005–518254) and UNICELLSYS (contract number HEALTH-2007–201142) to EK. I would like to thank Edda Klipp and Lilia Alberghina for their constant scientific support in the course of my research, and Marco Vanoni, Luca De Gioia, and Francesc Posas for stimulating discussions.

References

1. Obaya AJ, Sedivy JM (2002) Regulation of cyclin-Cdk activity in mammalian cells. Cell Mol Life Sci 59(1):126–142
2. Morgan DO (1995) Principles of CDK regulation. Nature 374(6518):131–134

3. De Clercq A, Inze D (2006) Cyclin-dependent kinase inhibitors in yeast, animals, and plants: a functional comparison. Crit Rev Biochem Mol Biol 41(5):293–313
4. Sherr CJ, Roberts JM (1995) Inhibitors of mammalian G1 cyclin-dependent kinases. Genes Dev 9(10):1149–1163
5. Sherr CJ, Roberts JM (1999) CDK inhibitors: positive and negative regulators of G1-phase progression. Genes Dev 13(12):1501–1512
6. Tsihlias J, Kapusta L, Slingerland J (1999) The prognostic significance of altered cyclin-dependent kinase inhibitors in human cancers. Annu Rev Med 50:401–423
7. Besson A, Dowdy SF, Roberts JM (2008) CDK inhibitors: cell cycle regulators and beyond. Dev Cell 14(2):159–169
8. Chu IM, Hengst L, Slingerland JM (2008) The Cdk inhibitor p27 in human cancer: prognostic potential and relevance to anticancer therapy. Nat Rev Cancer 8(4):253–267
9. Abukhdeir AM, Park BH (2008) P21 and p27: roles in carcinogenesis and drug resistance. Expert Rev Mol Med 10:e19
10. Hershko DD (2008) Oncogenic properties and prognostic implications of the ubiquitin ligase Skp2 in cancer. Cancer 112(7):1415–1424
11. Mishra A, Godavarthi SK, Jana NR (2009) UBE3A/E6-AP regulates cell proliferation by promoting proteasomal degradation of p27. Neurobiol Dis 36(1):26–34
12. Slingerland J, Pagano M (2000) Regulation of the cdk inhibitor p27 and its deregulation in cancer. J Cell Physiol 183(1):10–17
13. Lloyd RV, Erickson LA, Jin L, Kulig E, Qian X, Cheville JC, Scheithauer BW (1999) p27kip1: a multifunctional cyclin-dependent kinase inhibitor with prognostic significance in human cancers. Am J Pathol 154(2):313–323
14. Belletti B, Nicoloso MS, Schiappacassi M, Chimienti E, Berton S, Lovat F, Colombatti A, Baldassarre G (2005) p27(kip1) functional regulation in human cancer: a potential target for therapeutic designs. Curr Med Chem 12(14):1589–1605
15. Polyak K, Lee MH, Erdjument-Bromage H, Koff A, Roberts JM, Tempst P, Massagué J (1994) Cloning of p27Kip1, a cyclin-dependent kinase inhibitor and a potential mediator of extracellular antimitogenic signals. Cell 78(1):59–66
16. Toyoshima H, Hunter T (1994) p27, a novel inhibitor of G1 cyclin-Cdk protein kinase activity, is related to p21. Cell 78(1):67–74
17. Coats S, Flanagan WM, Nourse J, Roberts JM (1996) Requirement of p27Kip1 for restriction point control of the fibroblast cell cycle. Science 272(5263):877–880
18. Ray A, James MK, Larochelle S, Fisher RP, Blain SW (2009) p27Kip1 inhibits cyclin D-cyclin-dependent kinase 4 by two independent modes. Mol Cell Biol 29(4):986–999
19. LaBaer J, Garrett MD, Stevenson LF, Slingerland JM, Sandhu C, Chou HS, Fattaey A, Harlow E (1997) New functional activities for the p21 family of CDK inhibitors. Genes Dev 11(7):847–862
20. Cheng M, Olivier P, Diehl JA, Fero M, Roussel MF, Roberts JM, Sherr CJ (1999) The p21(Cip1) and p27(Kip1) CDK 'inhibitors' are essential activators of cyclin D-dependent kinases in murine fibroblasts. EMBO J 18(6):1571–83
21. Fero ML, Randel E, Gurley KE, Roberts JM, Kemp CJ (1998) The murine gene p27Kip1 is haplo-insufficient for tumour suppression. Nature 396(6707):177–180
22. Loda M, Cukor B, Tam SW, Lavin P, Fiorentino M, Draetta GF, Jessup JM, Pagano M (1997) Increased proteasome-dependent degradation of the cyclin-dependent kinase inhibitor p27 in aggressive colorectal carcinomas. Nat Med 3(2):231–234
23. Slingerland J, Pagano M (2000) Regulation of the cdk inhibitor p27 and its deregulation in cancer. J Cell Physiol 183(1):10–17
24. Coqueret O (2003) New roles for p21 and p27 cell-cycle inhibitors: a function for each cell compartment? Trends Cell Biol 13(2):65–70
25. Futcher B (1996) Cyclins and the wiring of the yeast cell cycle. Yeast 12(16):1635–1646
26. Nasmyth K (1996) At the heart of the budding yeast cell cycle. Trends Genet 12(10): 405–412

27. Cross FR, Yuste-Rojas M, Gray S, Jacobson MD (1999). Specialization and targeting of B-type cyclins. Mol Cell 4(1):11–9
28. Murray AW (2004) Recycling the cell cycle: cyclins revisited. Cell 116(2):221–234
29. Bloom J, Cross FR (2007) Multiple levels of cyclin specificity in cell-cycle control. Nat Rev Mol Cell Biol 8(2):149–160
30. Chang F, Herskowitz I (1990) Identification of a gene necessary for cell cycle arrest by a negative growth factor of yeast: FAR1 is an inhibitor of a G1 cyclin, CLN2. Cell 63(5):999–1011
31. Peter M, Herskowitz I (1994) Direct inhibition of the yeast cyclin-dependent kinase Cdc28-Cln by Far1. Science 265(5176):1228–1231
32. Mendenhall MD (1993) An inhibitor of p34CDC28 protein kinase activity from *Saccharomyces cerevisiae*. Science 259(5092):216–219
33. Schwob E, Bohm T, Mendenhall MD, Nasmyth K (1994) The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in *S. cerevisiae*. Cell 79(2):233–244
34. Alberghina L, Martegani E, Mariani L, Bortolan G (1983–1984) A bimolecular mechanism for the cell size control of the cell cycle. Biosystems 16(3–4):297–305
35. Alberghina L, Porro D, Cazzador L (2001) Towards a blueprint of the cell cycle. Oncogene 20(9):1128–1134
36. Deshaies RJ (1997) Phosphorylation and proteolysis: partners in the regulation of cell division in budding yeast. Curr Opin Genet Dev 7(1):7–16
37. Zachariae W, Nasmyth K (1999) Whose end is destruction: cell division and the anaphase-promoting complex. Genes Dev 13(16):2039–2058
38. Knapp D, Bhoite L, Stillman DJ, Nasmyth K (1996) The transcription factor Swi5 regulates expression of the cyclin kinase inhibitor p40SIC1. Mol Cell Biol 16(10):5701–5707
39. Toyn JH, Johnson AL, Donovan JD, Toone WM, Johnston LH (1997) The Swi5 transcription factor of *Saccharomyces cerevisiae* has a role in exit from mitosis through induction of the cdk-inhibitor Sic1 in telophase. Genetics 145(1):85–96
40. Aerne BL, Johnson AL, Toyn JH, Johnston LH (1998) Swi5 controls a novel wave of cyclin synthesis in late mitosis. Mol Biol Cell 9(4):945–956
41. Schneider BL, Yang QH, Futcher AB (1996) Linkage of replication to start by the Cdk inhibitor Sic1. Science 272(5261):560–562
42. Verma R, Annan RS, Huddleston MJ, Carr SA, Reynard G, Deshaies RJ (1997) Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase. Science 278(5337):455–460
43. Thornton BR, Toczyski DP (2003) Securin and B-cyclin/CDK are the only essential targets of the APC. Nat Cell Biol 5(12):1090–1094
44. Nash P, Tang X, Orlicky S, Chen Q, Gertler FB, Mendenhall MD, Sicheri F, Pawson T, Tyers M (2001) Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. Nature 414(6863):514–521
45. Skowyra D, Craig KL, Tyers M, Elledge SJ, Harper JW (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin–ligase complex. Cell 91(2): 209–219
46. Feldman RM, Correll CC, Kaplan KB, Deshaies RJ (1997) A complex of Cdc4p, Skp1p, and Cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. Cell 91(2):221–230
47. Verma R, Feldman RM, Deshaies RJ (1997) SIC1 is ubiquitinated in vitro by a pathway that requires CDC4, CDC34, and cyclin/CDK activities. Mol Biol Cell 8(8):1427–1437
48. Verma R, McDonald H, Yates JR 3rd, Deshaies RJ (2001) Selective degradation of ubiquitinated Sic1 by purified 26S proteasome yields active S phase cyclin-Cdk. Mol Cell 8(2):439–448
49. Rossi RL, Zinzalla V, Mastriani A, Vanoni M, Alberghina L (2005) Subcellular localization of the cyclin dependent kinase inhibitor Sic1 is modulated by the carbon source in budding yeast. Cell Cycle 4(12):1798–1807

50. Lengronne A, Schwob E (2002) The yeast CDK inhibitor Sic1 prevents genomic instability by promoting replication origin licensing in late G(1). Mol Cell 9(5):1067–1078
51. Caburet S, Conti C, Bensimon A (2002) Combing the genome for genomic instability. Trends Biotechnol 20(8):344–350
52. Nugroho TT Mendenhall MD (1994) An inhibitor of yeast cyclindependent protein kinase plays an important role in ensuring the genomic integrity of daughter cells. Mol Cell Biol 14(5):3320–3328
53. See WL, Miller JP, Squatrito M, Holland E, Resh MD, Koff A (2010) Defective DNA double-strand break repair underlies enhanced tumorigenesis and chromosomal instability in p27-deficient mice with growth factor-induced oligodendrogliomas. Oncogene 29(12):1720–1731
54. Schwob E (2004) Flexibility and governance in eukaryotic DNA replication. Curr Opin Microbiol 7(6):680–690
55. Han JD (2008) Understanding biological functions through molecular networks. Cell Res 18(2):224–237
56. Kitano H (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet 41(1):1–10
57. Westerhoff HV, Alberghina L (2005) Systems biology: did we know it all along? In: Alberghina L, Westerhoff HV (eds) Systems biology definitions and perspectives. Springer Berlin
58. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signaling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111
59. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(6761 Suppl):C47–C52
60. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147
61. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38(3):285–293
62. Bertin N, Simonis N, Dupuy D, Cusick ME, Han JD, Fraser HB, Roth FP, Vidal M (2007) Confirmation of organized modularity in the yeast interactome. PLoS Biol 5(6):e153
63. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430(6995):88–93
64. de Lichtenberg U, Jensen LJ, Brunak S, Bork P (2005) Dynamic complex formation during the yeast cell cycle. Science 307(5710):724–727
65. Csikasz-Nagy A, Battogtokh D, Chen KC, Novak B, Tyson JJ (2006) Analysis of a generic model of eukaryotic cell-cycle regulation. Biophys J 90(12):4361–4379
66. Fauré A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22(14):e124–131
67. Singhania R, Sramkoski RM, Jacobberger JW, Tyson JJ (2011) A hybrid model of mammalian cell cycle regulation. PLoS Comput Biol 7(2):e1001077
68. Kohn KW (1998) Functional capabilities of molecular network components controlling the mammalian G1/S cell cycle phase transition. Oncogene 16(8):1065–1075
69. Aguda BD, Tang Y (1999) The kinetic origins of the restriction point in the mammalian cell cycle. Cell Prolif 32(5):321–335
70. Qu Z, Weiss JN, MacLellan WR (2003) Regulation of the mammalian cell cycle: a model of the G1-to-S transition. Am J Physiol Cell Physiol 284(2):C349-C364

71. Swat M, Kel A, Herzel H (2004) Bifurcation analysis of the regulatory modules of the mammalian G1/S transition. Bioinformatics 20(10):1506–1511
72. Novak B, Tyson JJ (2004) A model for restriction point control of the mammalian cell cycle. J Theor Biol 230(4):563–579
73. Haberichter T, Madge B, Christopher RA, Yoshioka N, Dhiman A, Miller R, Gendelman R, Aksenov SV, Khalil IG, Dowdy SF (2007) A systems biology dynamical model of mammalian G1 cell cycle progression. Mol Syst Biol 3:84
74. Alfieri R, Barberis M, Chiaradonna F, Gaglio D, Milanesi L, Vanoni M, Klipp E, Alberghina L (2009) Towards a systems biology approach to mammalian cell cycle: modeling the entrance into S phase of quiescent fibroblasts after serum stimulation. BMC Bioinformatics 10 (Suppl 12):S16
75. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol Biol Cell 11(1):369–391
76. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15(8):3841–3862
77. Allen NA, Chen KC, Shaffer CA, Tyson JJ, Watson LT (2006) Computer evaluation of network dynamics models with application to cell cycle control in budding yeast. Syst Biol (Stevenage) 153(1):13–21
78. Barik D, Baumann WT, Paul MR, Novak B, Tyson JJ (2010) A model of yeast cell-cycle regulation based on multisite phosphorylation. Mol Syst Biol 6:405
79. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci USA 101(14): 4781–4786
80. Irons DJ (2009) Logical analysis of the budding yeast cell cycle. J Theor Biol 257(4):543–559
81. Fauré A, Naldi A, Lopez F, Chaouiya C, Ciliberto A, Thieffry D (2009) Modular logical modelling of the budding yeast cell cycle. Mol BioSyst 5(12):1787–1796
82. Braunewell S, Bornholdt S (2007) Superstability of the yeast cell-cycle dynamics: ensuring causality in the presence of biochemical stochasticity. J Theor Biol 245(4):638–643
83. Palmisano A, Mura I, Priami C (2009) From ODES to language-based, executable models of biological systems. Pac Symp Biocomput 14:239–250
84. Barberis M, Beck C, Amoussouvi A, Schreiber G, Diener C, Herrmann A, Klipp E (2011) Low number of SIC1 mRNA molecules ensures low noise level in cell cycle progression of budding yeast. Mol BioSyst 7(10):2804–2812
85. Lovrics A, Csikasz-Nagy A, Zsely IG, Zador J, Turanyi T, Novak B (2006) Time scale and dimension analysis of a budding yeast cell cycle model. BMC Bioinform 7:494
86. Ciliberto A, Lukács A, Tóth A, Tyson JJ, Novák B (2005) Rewiring the exit from mitosis. Cell Cycle 4(8):1107–1112
87. Queralt E, Lehane C, Novak B, Uhlmann F (2006) Downregulation of PP2A(Cdc55) phosphatase by separase initiates mitotic exit in budding yeast. Cell 125(4):719–732
88. Tóth A, Queralt E, Uhlmann F, Novák B (2007) Mitotic exit in two dimensions. J Theor Biol 248(3):560–573
89. Vinod PK, Freire P, Rattani A, Ciliberto A, Uhlmann F, Novak B (2011) Computational modelling of mitotic exit in budding yeast: the role of separase and Cdc14 endocycles. J R Soc Interface 8(61):1128–1141
90. Ball DA, Ahn TH, Wang P, Chen KC, Cao Y, Tyson JJ, Peccoud J, Baumann WT (2011) Stochastic exit from mitosis in budding yeast: model predictions and experimental observations. Cell Cycle 10(6):999–1009
91. Alarcón T, Tindall MJ (2007) Modelling cell growth and its modulation of the G1/S transition. Bull Math Biol 69(1):197–214
92. Barberis M, Klipp E, Vanoni M, Alberghina L (2007) Cell size at S phase initiation: an emergent property of the G1/S network. PLoS Comput Biol 3(4):e64
93. Bruggeman FJ, Westerhoff HV (2007) The nature of systems biology. Trends Microbiol 15(1):45–50

94. Barberis M, De Gioia L, Ruzzene M, Sarno S, Coccetti P, Fantucci P, Vanoni M, Alberghina L (2005) The yeast cyclindependent kinase inhibitor Sic1 and mammalian p27Kip1 are functional homologues with a structurally conserved inhibitory domain. Biochem J 387(Pt 3):639–647
95. Barberis M, Pagano MA, Gioia LD, Marin O, Vanoni M, Pinna LA, Alberghina L (2005) CK2 regulates in vitro the activity of the yeast cyclin-dependent kinase inhibitor Sic1. Biochem Biophys Res Commun 336(4):1040–1048
96. Barberis M, Klipp E (2007) Insights into the network controlling the G1/S transition in budding yeast. Genome Inform 18:85–99
97. Barberis M, Linke C, Adrover MA, Lehrach H, Posas F, Krobitsch S, Klipp E (2011) Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins. Biotechnol Adv, doi:10.1016/j.biotechadv.2011.09.004
98. Archambault V, Li CX, Tackett AJ, Wasch R, Chait BT, Rout MP, Cross FR (2003) Genetic and biochemical evaluation of the importance of Cdc6 in regulating mitotic exit. Mol Biol Cell 14(11):4592–4604
99. Coccetti P, Rossi RL, Sternieri F, Porro D, Russo GL, di Fonzo A, Magni F, Vanoni M, Alberghina L (2004) Mutations of the CK2 phosphorylation site of Sic1 affect cell size and S-Cdk kinase activity in *Saccharomyces cerevisiae*. Mol Microbiol 51(2):447–460
100. Cross FR, Schroeder L, Bean JM (2007) Phosphorylation of the Sic1 inhibitor of B-type cyclins in *Saccharomyces cerevisiae* is not essential but contributes to cell cycle robustness. Genetics 176(3):1541–1555
101. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7(3):188–197
102. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062):1173–1178
103. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein–protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968
104. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403(6770): 623–627
105. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98(8): 4569–4574
106. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabási AL, Tavernier J, Hill DE, Vidal M (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322(5898):104–110
107. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB (2004) Structure-based assembly of protein complexes in yeast. Science 303(5666):2026–2029
108. Aloy P, Pichaud M, Russell RB (2005) Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol 15(1):15–22

109. Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22(10):1317–1321
110. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332(5):989–998
111. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 99(9):5896–5901
112. Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins 49(3):350–364
113. Lu L, Arakaki AK, Lu H, Skolnick J (2003) Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. Genome Res 13(6A):1146–1154
114. Mosca R, Pons C, Fernández-Recio J, Aloy P (2009) Pushing structural information into the yeast interactome by high-throughput protein docking experiments. PLoS Comput Biol 5(8):e1000490
115. Sánchez-Díaz A, González I, Arellano M, Moreno S (1998) The Cdk inhibitors p25rum1 and p40SIC1 are functional homologues that play similar roles in the regulation of the cell cycle in fission and budding yeast. J Cell Sci 111(Pt 6):843–851
116. Peter M, Herskovitz I (1994) Joining the complex: cyclin-dependent kinase inhibitory proteins and the cell cycle. Cell 79(2):181–184
117. Hodge A, Mendenhall M (1999) The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. Mol Gen Genet 262(1):55–64
118. Russo AA, Jeffrey PD, Patten AK, Massagué J, Pavletich NP (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature 382(6589):325–331
119. Barberis M (2000) Phenotypic analysis of Ser201/Glu and Ser201/Ala mutants in *Saccharomyces cerevisiae* and their functional interpretation on the basis of a three-dimensional model of Clb5/Cdc28/p40Sic1 complex built by homology. Dissertation, University of Milano-Bicocca, Milan
120. Heiner AP, Berendsen HJ, van Gunsteren WF (1992) MD simulation of subtilisin BPN' in a crystal environment. Proteins 14(4):451–464
121. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy . J Mol Biol 232(2):584–599
122. Rost B, Sander C, Schneider R (1994) PHD – an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10(1):53–60
123. Rost B (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525–539
124. Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M, Hagler AT (1988) Structure and energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. Proteins 4(1): 31–47
125. Cross FR, Jacobson MD (2000) Conservation and function of a potential substrate-binding domain in the yeast Clb5 B-type cyclin. Mol Cell Biol 20(13):4782–4790
126. Schulman BA, Lindstrom DL, Harlow E (1998) Substrate recruitment to cyclin-dependent kinase 2 by a multipurpose docking site on cyclin A. Proc Natl Acad Sci USA 95(18): 10453–10458
127. Morgan DO (1996) Under arrest at atomic resolution. Nature 382(6589):295–296
128. Spolar RS, Record MT Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. Science 263(5148):777–784
129. Malmqvist M (1999) BIACORE: an affinity biosensor system for characterization of biomolecular interactions. Biochem Soc Trans 27(2):335–340
130. Lacy ER, Wang Y, Post J, Nourse A, Webb W, Mapelli M, Musacchio A, Siuzdak G, Kriwacki RW (2005) Molecular basis for the specificity of p27 toward cyclin-dependent kinases that regulate cell division. J Mol Biol 349(4):764–773

131. Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki RW (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. Nat Struct Mol Biol 11(4):358–364
132. Besson A, Hwang HC, Cicero S, Donovan SL, Gurian-West M, Johnson D, Clurman BE, Dyer MA, Roberts JM (2007) Discovery of an oncogenic activity in p27Kip1 that causes stem cell expansion and a multiple tumor phenotype. Genes Dev 21(14):1731–1746
133. Johnsson B, Löfås S, Lindquist G (1991) Immobilization of proteins to a carboxymethyldextran-modified gold surface for biospecific interaction analysis in surface plasmon resonance sensors. Anal Biochem 198(2):268–277
134. Bienkiewicz EA, Adkins JN, Lumb KJ (2002) Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). Biochemistry 41(3):752–759
135. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW (2008) Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 47(29):7598–7609
136. Hengst L, Dulic V, Slingerland JM, Lees E, Reed SI (1994) A cell cycle-regulated inhibitor of cyclin-dependent kinases. Proc Natl Acad Sci USA 91(12):5291–5295
137. Bowman P, Galea CA, Lacy E, Kriwacki RW (2006) Thermodynamic characterization of interactions between p27(Kip1) and activated and non-activated Cdk2: intrinsically unstructured proteins as thermodynamic tethers. Biochim Biophys Acta 1764(2):182–189
138. Sivakolundu SG, Bashford D, Kriwacki RW (2005) Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. J Mol Biol 353(5):1118–1128
139. Galea CA, Nourse A, Wang Y, Sivakolundu SG, Heller WT, Kriwacki RW (2008) Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. J Mol Biol 376(3):827–838
140. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6(5):1882–1898
141. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN (2007) Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J Proteome Res 6(5):1899–1916
142. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323(3):573–584
143. Bloom J, Pagano M (2003) Deregulated degradation of the cdk inhibitor p27 and malignant transformation. Semin Cancer Biol 13(1):41–47
144. Chu IM, Hengst L, Slingerland JM (2008) The Cdk inhibitor p27 in human cancer: prognostic potential and relevance to anticancer therapy. Nat Rev Cancer 8(4):253–267
145. Lee J, Kim SS (2009) The function of p27 KIP1 during tumor development. Exp Mol Med 41(11):765–771
146. Blagosklonny MV (2002) Are p27 and p21 cytoplasmic oncoproteins? Cell Cycle 1(6): 391–393
147. Sicinski P, Zacharek S, Kim C (2007) Duality of p27Kip1 function in tumorigenesis. Genes Dev 21(14):1731–1746
148. Chu I, Sun J, Arnaout A, Kahn H, Hanna W, Narod S, Sun P, Tan CK, Hengst L, Slingerland J p27 phosphorylation by Src regulates inhibition of cyclin E-Cdk2. Cell 128(2): 281–294
149. Grimmler M, Wang Y, Mund T, Cilensek Z, Keidel EM, Waddell MB, Jäkel H, Kullmann M, Kriwacki RW, Hengst L (2007) Cdk-inhibitory activity and stability of p27Kip1 are directly regulated by oncogenic tyrosine kinases. Cell 128(2):269–280
150. Hidaka T, Hama S, Shrestha P, Saito T, Kajiwara Y, Yamasaki F, Sugiyama K, Kurisu K (2009) The combination of low cytoplasmic and high nuclear expression of p27 predicts a better prognosis in high-grade astrocytoma. Anticancer Res 29(2):597–603

151. Mittag T, Orlicky S, Choy WY, Tang X, Lin H, Sicheri F, Kay LE, Tyers M, Forman-Kay JD (2008) Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. Proc Natl Acad Sci USA 105(46):17772–17777
152. Mittag T, Marsh J, Grishaev A, Orlicky S, Lin H, Sicheri F, Tyers M, Forman-Kay JD (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 18(4):494–506
153. Brocca S, Samalíková M, Uversky VN, Lotti M, Vanoni M, Alberghina L, Grandori R (2009) Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins 76(3):731–746
154. Brocca S, Testa L, Samalikova M, Grandori R, Lotti M (2011) Defining structural domains of an intrinsically disordered protein: Sic1, the cyclin-dependent kinase inhibitor of Saccharomyces cerevisiae. Mol Biotechnol 47(1):34–42
155. Testa L, Brocca S, Samalikova M, Santambrogio C, Alberghina L, Grandori R (2011) Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. Biotechnol J 6(1):96–100
156. Brocca S, Testa L, Sobott F, Šamalikova M, Natalello A, Papaleo E, Lotti M, De Gioia L, Doglia SM, Alberghina L, Grandori R (2011) Compact conformations of an intrinsically disordered protein: the kinase-inhibitor domain of Sic1. Biophys J 100(9):2243–2252
157. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32(3):1037–1049
158. Escote X, Zapater M, Clotet J, Posas F (2004) Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. Nat Cell Biol 6(10):997–1002
159. Zapater M, Clotet J, Escoté X, Posas F (2005) Control of cell cycle progression by the stress-activated Hog1 MAPK. Cell Cycle 4(1):6–7
160. Bjornsti MA, Houghton PJ (2004) The TOR pathway: a target for cancer therapy. Nat Rev Cancer 4(5):335–348
161. Zinzalla V, Graziola M, Mastriani A, Vanoni M, Alberghin L (2007) Rapamycin-mediated G1 arrest involves regulation of the Cdk inhibitor Sic1 in *Saccharomyces cerevisiae*. Mol Microbiol 63(5):1482–1494
162. Nishizawa M, Kawasumi M, Fujino M, Toh-e A (1998) Phosphorylation of Sic1, a cyclin-dependent kinase (Cdk) inhibitor, by Cdk including Pho85 kinase is required for its prompt degradation. Mol Biol Cell 9(9):2393–2405
163. Sedgwick C, Rawluk M, Decesare J, Raithatha S, Wohlschlegel J, Semchuk P, Ellison M, Yates J 3rd, Stuart D (2006) *Saccharomyces cerevisiae* Ime2 phosphorylates Sic1 at multiple PXS/T sites but is insufficient to trigger Sic1 degradation. Biochem J 399(1):151–160
164. Meggio F, Pinna LA (2003) One-thousand-and-one substrates of protein kinase CK2? FASEB J 17(3):349–368
165. Tapia JC, Bolanos-Garcia VM, Sayed M, Allende CC, Allende JE (2004) Cell cycle regulatory protein p27KIP1 is a substrate and interacts with the protein kinase CK2. J Cell Biochem 91(5):865–879
166. Coccetti P, Zinzalla V, Tedeschi G, Russo GL, Fantinato S, Marin O, Pinna LA, Vanoni M, Alberghina L (2006) Sic1 is phosphorylated by CK2 on Ser201 in budding yeast cells. Biochem Biophys Res Commun 346(3):786–793
167. Tripodi F, Zinzalla V, Vanoni M, Alberghina L, Coccetti P (2007) In CK2 inactivated cells the cyclin dependent kinase inhibitor Sic1 is involved in cell-cycle arrest before the onset of S phase. Biochem Biophys Res Commun 359(4):921–927
168. Xu X, Nakano T, Wick S, Dubay M, Brizuela L (1999) Mechanism of Cdk2/Cyclin E inhibition by p27 and p27 phosphorylation. Biochemistry 38(27):8713–8722
169. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H (2005) Systems biology in practice. Concepts, implementation and application. Wiley, KGaA
170. Bracken C, Iakoucheva LM, Romero PR, Dunker AK (2004) Combining prediction, computation and experiment for the characterization of protein disorder. Curr Opin Struct Biol 14(5):570–576

171. Du JT, Li YM, Ma QF, Qiang W, Zhao YF, Abe H, Kanazawa K, Qin XR, Aoyagi R, Ishizuka Y, Nemoto T, Nakanishi H (2005) Synthesis and conformational properties of phosphopeptides related to the human tau protein. Regul Pept 130(1–2):48–5
172. Suenaga A, Kiyatkin AB, Hatakeyama M, Futatsugi N, Okimoto N, Hirano Y, Narumi T, Kawai A, Susukita R, Koishi T, Furusawa H, Yasuoka K, Takada N, Ohno Y, Taiji M, Ebisuzaki T, Hoek JB, Konagaya A, Kholodenko BN (2005) Tyr-317 phosphorylation increases Shc structural rigidity and reduces coupling of domain motions remote from the phosphorylation site as revealed by molecular dynamics simulations. J Biol Chem 279(6):4657–4662
173. Sidorova JM, Breeden LL (2003) Precocious G1/S transitions and genomic instability: the origin connection. Mutat Res 532(1–2):5–19
174. Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23(8):961–966
175. Fitch I, Dahmann C, Surana U, Amon A, Nasmyth K, Goetsch L, Byers B, Futcher B (1992) Characterization of four B-type cyclin genes of the budding yeast *Saccharomyces cerevisiae*. Mol Biol Cell 3(7):805–818
176. Koch C, Nasmyth K (1994) Cell cycle regulated transcription in yeast. Curr Opin Cell Biol 6(3):451–459
177. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7(3):165–176
178. Bailly E, Reed SI (1999) Functional characterization of rpn3 uncovers a distinct 19S proteasomal subunit requirement for ubiquitin-dependent proteolysis of cell cycle regulatory proteins in budding yeast. Mol Cell Biol 19(10):6872–6890
179. Honey S, Schneider BL, Schieltz DM, Yates JR, Futcher B (2001) A novel multiple affinity purification tag and its use in identification of proteins associated with a cyclin–CDK complex. Nucleic Acids Res 29(4):E24
180. Archambault V, Chang EJ, Drapkin BJ, Cross FR, Chait BT, Rout MP (2004) Targeted proteomic study of the cyclin–Cdk module. Mol Cell 14(6):699–711
181. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084):631–636
182. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440(7084):637–643
183. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteom 6(3):439–450
184. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, Ahn J, Dewar-Darch D, Reguly T, Tang X, Almeida R, Qin ZS, Pawson T, Gingras AC, Nesvizhskii AI, Tyers M (2010) A global protein kinase and phosphatase interaction network in yeast. Science 328(5981):1043–1046
185. Nair DK, Jose M, Kuner T, Zuschratter W, Hartig R (2006) FRET-FLIM at nanometer spectral resolution from living cells. Opt Express 14(25):12217–12229
186. Rizzo MA, Springer GH, Granada B, Piston DW (2004) An improved cyan fluorescent protein variant useful for FRET. Nat Biotechnol 22(4):445–449

187. Schreiber G, Barberis M, Scolari S, Klaus C, Herrmann A, Klipp E (2012) Unraveling interactions of cell cycle-regulating proteins Sic1 and B-type cyclins in living yeast cells: a FLIMFRET approach. FASEB J 26, doi:10.1096/fj.11-192518
188. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO (2003) Targets of the cyclin-dependent kinase Cdk1. Nature 425(6960):859–864
189. Stuart D, Wittenberg C (1998) CLB5 and CLB6 are required for premeiotic DNA replication and activation of the meiotic S/M checkpoint. Genes Dev 12(17):2698–2710
190. Richardson H, Lew DJ, Henze M, Sugimoto K, Reed SI (1992) Cyclin-B homologs in *Saccharomyces cerevisiae* function in S phase and in G2. Genes Dev 6(11):2021–2034
191. López-Avilés S, Kapuy O, Novák B, Uhlmann F (2009) Irreversibility of mitotic exit is the consequence of systems-level feedback. Nature 459(7246):592–595
192. Cross FR, Archambault V, Miller M, Klovstad M (2002) Testing a mathematical model of the yeast cell cycle. Mol Biol Cell 13(1):52–70
193. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425(6959): 737–741
194. Thornton BR, Toczyski DP (2003) Securin and B-cyclin/CDK are the only essential targets of the APC. Nat Cell Biol 5(12):1090–1094
195. Miller ME, Cross FR (2001) Cyclin specificity: how many wheels do you need on a unicycle? J Cell Sci 114 (Pt 10):1811–1820

Chapter 8
Proteome-Wide Screens in *Saccharomyces cerevisiae* Using the Yeast GFP Collection

Yolanda T. Chong*, Michael J. Cox*, and Brenda Andrews

Abstract The budding yeast is a simple and genetically tractable eukaryotic organism. It remains a leading system for functional genomic work and has been the focus of many pioneering efforts, including the systematic construction and analysis of gene deletion mutants. Over the past decade, many large-scale studies have made use of the deletion and other mutant collections to assay genetic interactions, chemical sensitivities, and other phenotypes, contributing enormously to our understanding of gene function. The deletion mutant collection has also been used in cell biological surveys to identify genes that control cell and organelle morphology. One valuable approach for systematic definition of gene function and biological pathways involves global assessment of the localization patterns of the proteins they encode and how these patterns are altered in response to environmental or genetic perturbation. However, proteome-wide, cell biological screens are extremely challenging, from both a technical and computational perspective. The yeast GFP collection, an elegant and unique strain set, is ideal for studying both protein localization and abundance across the proteome (http://yeastgfp.yeastgenome.org/). In this chapter, we outline how the yeast GFP collection has been used to date and discuss approaches for conducting future surveys of the proteome.

1 The Yeast GFP Collection

The first effort to comprehensively determine the subcellular localization of every protein in a eukaryotic cell exploited the *Saccharomyces cerevisiae* model system (Fig. 8.1a) [1]. Huh et al. created a collection of haploid budding yeast strains

*Authors contributed equally

Y.T. Chong • M.J. Cox • B. Andrews (✉)
The Donnelly Centre, Department of Molecular Genetics, University of Toronto, Toronto, Canada
e-mail: yolanda.chong@utoronto.ca; mike.cox@utoronto.ca; brenda.andrews@utoronto.ca

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_8,

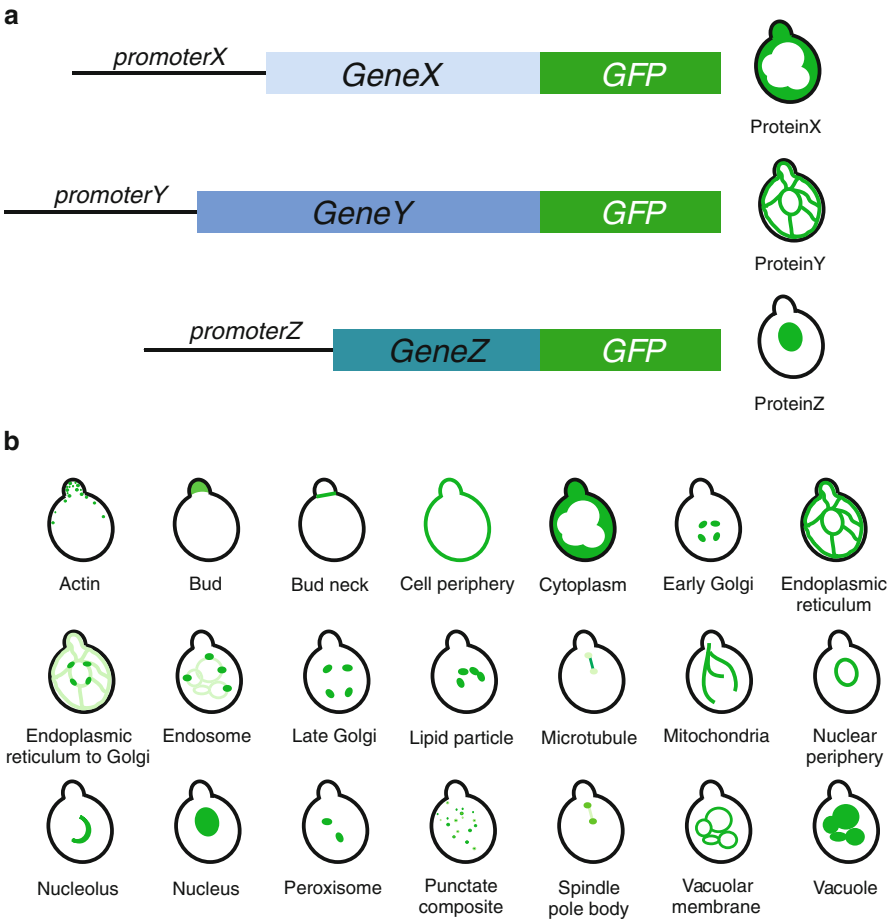


Fig. 8.1 Yeast array for systematic analysis of the proteome. (**a**) Construction of the yeast GFP collection. The GFP cassette was integrated at the 3′ end, directly upstream of the stop codon, for every open reading frame in the yeast genome. (**b**) Subcellular compartments in the yeast cell. Schematic modified from yeastgfp.yeastgenome.org. Proteins that fall into the punctate composite category did not co-localize with any of the compartment-specific markers tested in this study

in which 97% of annotated yeast open reading frames were tagged at their chromosomal loci with a cassette encoding the green fluorescent protein (GFP) and a selectable marker. Each strain in this collection produces a different full-length protein tagged with GFP at its C-terminus under the control of its endogenous promoter. After analysis using wide-field fluorescence microscopy, 4,156 strains, covering approximately 70% of the proteome, were annotated as having a visible GFP signal. Two individuals classified the collection into 11 localization categories using manual scoring and then later conducted co-localization experiments to assign proteins to another 11 subcellular patterns, defining 22 distinct locations within the

cell (Fig. 8.1b). This project assigned 70% of the previously unannotated proteome to a subcellular location and showed 80% agreement with low throughput findings. Proteins whose functions are affected by the GFP tag may account for some of the proteins whose localization patterns could not be determined. However, the GFP-tagging project revealed that 87% of essential genes can tolerate a C-terminal GFP tag suggesting that, for most proteins, the tag does not significantly impair protein function, consistent with other studies [2].

2 Measuring Protein Abundance

The yeast GFP collection can be combined with fluorescence microscopy or flow cytometry to measure protein abundances across the proteome. Both techniques are based on the principle that fluorescence of a GFP-tagged protein is proportional to its abundance [3] and measure protein abundance at single cell resolution in living cells. While GFP-tagging can potentially alter protein stability, the majority of proteins in the GFP collection are not likely to be affected [3]; however, abundance changes identified through high-throughput techniques should be verified using complementary approaches. Below, we describe innovative projects that made use of the yeast GFP collection to systematically survey changes in the abundance of the proteome in response to genetic and environmental perturbations.

3 Monitoring Protein Turnover

High-throughput fluorescence microscopy and the GFP collection were used to find new targets of the F-box protein Grr1, a specificity component of the conserved SCF E3-ubiquitin ligase complex [4]. For this screen, high-content imaging and a two-color reporter system were used to simultaneously visualize wild-type and mutant cells to avoid illumination discrepancies. A strain carrying a disruption of the *GRR1* locus was marked with RFP and crossed into the GFP collection. Wild-type and mutant (RFP-expressing) cells were mixed and imaged in the same well. To mark and identify the entire cell population, a blue fluorescent dye was used. Median GFP intensities of the wild-type population were compared to median GFP intensities of the mutant population and strains expressing greater than two-fold changes were further analyzed as potential substrates of Grr1. Subsequent experiments showed that the abundance changes identified in this screen resulted from both changes in gene expression and protein stability and revealed new targets of Grr1, illustrating the value of this approach.

4 Changes in Protein Levels in Response to Environmental Perturbations

Flow cytometry can be used as another approach to monitor fluorescence produced by GFP-fusion proteins in living cells. While this technique is not as sensitive as fluorescence microscopy and cannot be used to monitor protein localization, high-throughput flow cytometers can measure protein abundance for tens of thousands of cells in seconds [3]. Newman et al. found that ~2,500 strains in the GFP collection produced a GFP signal of sufficient strength to be reliably distinguished from cellular autofluorescence using flow cytometry. Approximately 40% of the tagged proteins in these strains showed changes in abundance when cells were grown in different media. Comparison with DNA microarrays showed that most of the changes in fluorescence could be explained by changes in mRNA levels; however, for at least 6% of the strains tested, alterations in protein abundance appeared to result from post-transcriptional mechanisms.

5 Detecting Gene Dosage Effects on Protein Abundance

Flow cytometry and the GFP collection have also been used to explore the relationship between gene copy number and protein expression levels. The abundance of a GFP-tagged protein encoded by a single copy of a gene was measured in diploid strains which contained either a wild-type copy or a deletion of this gene at the homologous chromosomal locus [5]. To minimize discrepancies in the measurement of GFP fluorescence due to experimental variation, wild-type strains that constitutively expressed mCherry were co-cultured with the heterozygous GFP-ORF deletion strains that were not marked with mCherry; thus the genotype of each cell in the sample could be determined by the presence or absence of red fluorescence. The analysis of 730 different GFP-tagged proteins showed that, in the vast majority of cases, GFP-fusion protein levels did not change in the heterozygous deletion strains to compensate for reduced gene dosage. Furthermore, while many strains showed changes in GFP-tagged protein abundance in response to different growth conditions, the set of genes exhibiting dosage compensation was unchanged under these conditions. These results suggest that if compensatory mechanisms do exist they are rarely triggered when protein dosage drops to only 50% of that normally found in the cell.

Similar experiments were used to ask whether functional compensation occurred between paralogous gene pairs in haploid cells. More than 200 strains were analyzed in which one member of a paralogous pair was deleted while the other was tagged with GFP. When grown in rich medium, 11% of these strains showed an increase in the abundance of a GFP-tagged protein when the gene encoding its paralog was deleted [6]. In most cases, the increased protein abundance reflected increased transcription of the GFP-tagged gene. However, for some genes, changes

in mRNA levels did not correlate with protein abundance, suggesting that post-transcriptional regulatory mechanisms may also be involved. Interestingly, the gene pairs that exhibited paralog responsiveness were also more likely to display synthetic lethal interactions than non-responsive pairs. This result suggests that paralog responsiveness occurs between gene pairs that share a common function required for cell growth. This hypothesis is supported by the observation that additional examples of paralog responsiveness were uncovered when cells were grown under different environmental conditions and, under these conditions, the responsive gene pairs were required for optimal cell viability.

6 Defining Protein Movement Within the Cell

Several studies have been conducted using all or part of the yeast GFP collection to identify proteins that change localization in response to perturbations. This collection is particularly useful for dynamic localization studies because the tagged genes are under the control of their endogenous promoters; therefore, the proteins they encode are likely to be produced at physiological levels, minimally disrupting intracellular transport mechanisms.

Shin and colleagues sought to identify novel targets of the TORC1 kinase signaling pathway [7]. The entire GFP collection was treated with rapamycin, a TORC1 inhibitor that induces a starvation response in yeast, and pre- and post-treated cells were manually examined for differences in protein localization. When rapamycin was applied to cells for 2 hours, 98 localization changes were observed out of the 4,156 strains tested. Further analysis of one of these potential targets revealed a previously unappreciated connection between TORC1 and Stp1, a transcription factor involved in amino acid sensing.

In a more focused survey, 1,632 strains from the collection that had been manually annotated as having only a cytoplasmic pattern were screened for new intracellular structures that may not have been identified when the collection was first examined [2]. In this study, manual scorers visually identified nine proteins capable of forming filamentous subcellular structures in the yeast cytoplasm. To exclude the possibility of filament formation resulting from the presence of the GFP tag, HA tags were substituted and filament formation was still observed for these proteins. Co-localization experiments with these proteins revealed that they assembled into four different types of novel filaments. Interestingly, different environmental conditions had varying effects on the formation of different classes of filaments. Based on their findings, the authors speculate that the formation of these structures is a potential mechanism for regulating the activity of these proteins in the cell. These observations suggest that screening the GFP collection under different conditions may enable identification of previously undefined subcellular structures.

Another study was conducted to identify condition-specific changes in the proteome on a subset of strains from the GFP collection. To observe the presence of "macro-molecular depots" in cells transitioning in and out of the stationary phase,

approximately 800 strains, all of which expressed GFP-tagged proteins annotated to be cytosolic under standard growth conditions, were screened for protein localization pattern differences between quiescent and actively proliferating cells [8]. In this study, the researchers used a unique cell array approach, where chemically fixed cells were adhered to a microscope slide, forming a cell chip [9]. In this survey, two scorers manually found 180 proteins that formed obvious foci when cells were grown to stationary phase. Most proteins associated with foci had annotated roles in stress response and metabolism, consistent with their response to the growth conditions. Formation of some foci was nutrient-specific and reversible when cells were placed in fresh media, suggesting that foci may serve as intermediary storage mechanisms during cellular stasis. The same cell chip technology was used in a systematic characterization of the yeast proteome in response to treatment with mating pheromone [10]. Two annotators visually screened the GFP collection for proteins that relocalize to the shmoo tip, a cell projection that forms in response to pheromone to facilitate mating. Up to 16,000 micrographs per chip were analyzed, a significant advancement in throughput and sample handling. However, fixation of the GFP-tagged strains increased autofluorescence, resulting in a high false negative rate, suggesting that live cell imaging may facilitate more reliable screening of the yeast GFP collection.

7 Applying a High-content Screening Approach to the Yeast GFP Collection

Although the GFP collection is a unique resource for systematic analysis of the dynamic proteome, it has been arguably underutilized, largely due to the challenges associated with analysis of large sets of cell biological data. Visually analyzing thousands of micrographs is a daunting and time consuming task, and localization assignments made by different human scorers often show poor agreement [9, 11]. These problems can be avoided by adopting a high-content screening approach, which combines high-throughput microscopy with automated image analysis.

Methods are available for automated manipulation of yeast arrays, and these methods can be readily adapted to the GFP collection. In particular, synthetic genetic array (SGA) technology allows the introduction of a marked allele of any query gene of interest, such as a fluorescent marker for a cell compartment of interest, into an arrayed yeast collection through a series of replica pinning steps [12] (Fig. 8.2a). For example, SGA has been used to introduce a fluorescent tubulin protein into the yeast deletion collection to identify mutants with defects in spindle morphology [13]. The major challenge lies not in creating cell arrays and assays compatible with automated image analysis, but rather in the computational assessment of changes in the localization of the proteome in response to genetic or environmental perturbations.

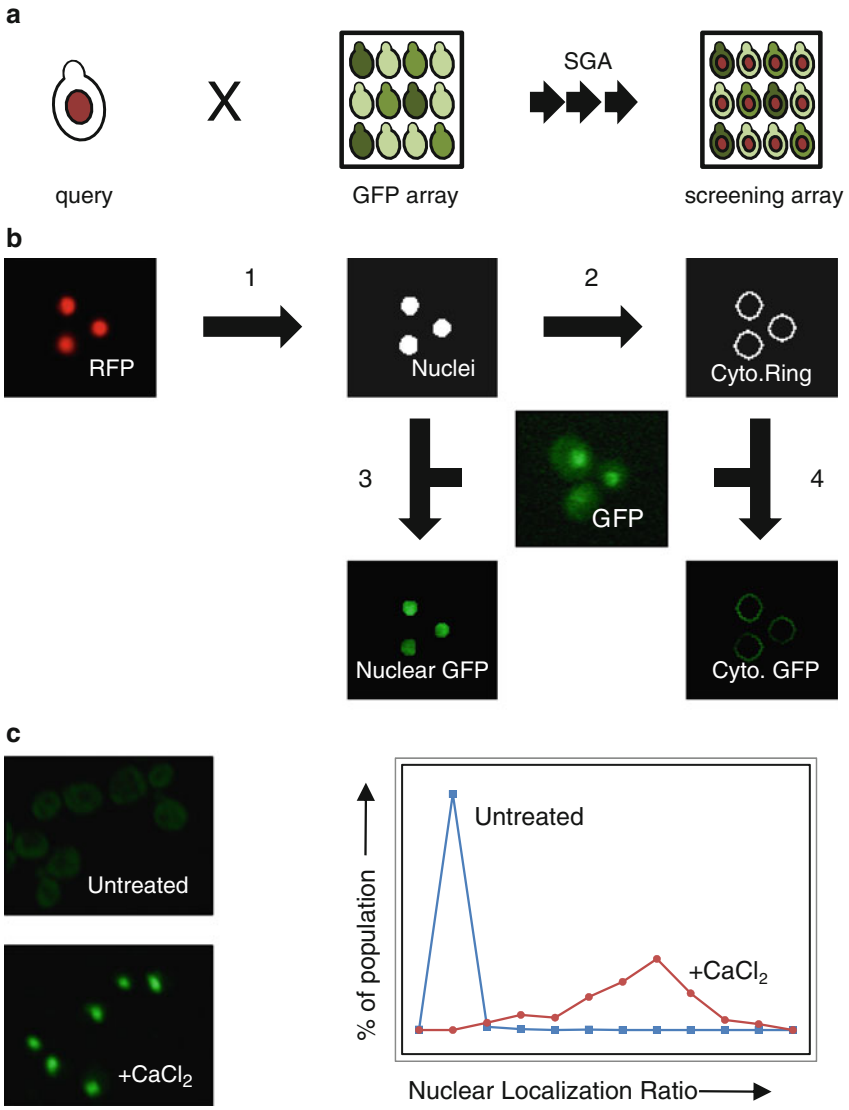


Fig. 8.2 Scheme for automated analysis of changes in nuclear protein localization. (**a**) Generation of strains for imaging. A query strain carrying an RFP-tagged nuclear marker is mated to an array of strains from the GFP collection. The synthetic genetic array or SGA method is used to create an array of haploid strains that express both of the GFP- and RFP-tagged proteins. (**b**) Segmentation of images. The RFP-tagged nuclear landmark is used to identify the location of nuclei, generating a nuclear mask (*1*). A ring is expanded around the nuclear mask into the cytoplasm (Cyto.Ring) (*2*). Intensity measurements are extracted from the GFP image using either the nuclear masks (*3*) or cytoplasmic rings (*4*). (**c**) Quantification of nuclear translocation. For each cell in an image, the nuclear localization ratio is calculated by dividing the average nuclear GFP intensity by the average cytoplasmic GFP intensity. (*Left*) Images of cells expressing a GFP-tagged derivative of Crz1, a calcium-responsive transcription factor. Untreated Crz1-GFP cells and cells treated with 0.2 M $CaCl_2$ are shown. (*Right*) Histogram showing the nuclear localization ratios for untreated (*squares*) or $CaCl_2$-treated (*circles*) Crz1-GFP samples [>190 nuclei per sample]

To use computational image analysis successfully, regions of interest (ROIs) in an image, such as cells or organelles, must first be defined using a process known as segmentation. ROIs can be identified using fluorescently-tagged proteins or commercially available dyes as landmarks for cellular compartments. For example, the GFP collection might be surveyed for proteins that shuttle in and out of the nucleus in response to a perturbation by introducing a fluorescent landmark that defines the nucleus. Images for both the GFP marker and the nuclear landmark for each strain can be acquired using an automated imaging system (available systems are reviewed in Vizeacoumar et al. [14]) (Fig. 8.2b). Image analysis software such as CellProfiler or ImageJ can then be used to segment the ROIs in the image using the nuclear landmark [11, 15]. Following segmentation, the nuclear region can be expanded by a defined distance into the surrounding cytoplasm (Fig. 8.2b). By subtracting the nuclear region from this expanded region, a ring that overlaps a portion of the cytoplasm is created. Average intensity measurements can then be extracted from the GFP image in the areas that correspond to nucleus and cytoplasmic ring, and a ratio of these numbers can be used as a measure of the relative distribution of the fusion protein between these two compartments (Fig. 8.2c). We have used this relatively simple segmentation approach to survey the movement of yeast transcription factors in response to environmental and genetic perturbations (MC and BA, unpublished).

To screen the entire GFP collection for protein localization changes under genetic or chemical perturbations, the co-localization approach described above could be feasibly applied; however, this would require the introduction of at least 22 different compartmental landmarks into the collection. Alternatively, a computational approach can be used to determine the subcellular localization of each GFP-tagged protein, based solely upon the distribution of the GFP-signal within the cell. To survey the collection in this fashion, the entire cell must be segmented using a cytoplasmic fluorescent marker or dye. Alternatively, light microscope images can be used to define the cell [16, 17], although often with less success than fluorescent images. Once the cell has been identified, numerous texture measurements, which describe the distribution of the GFP signal within these ROIs, can be used in a machine learning approach to define a set of rules (i.e., a classifier) that reliably discriminate each unique subcellular localization pattern [18]. For example, a support vector machine multi-class classification method was used to define patterns in published images for ~2,600 strains in the GFP collection with 81% accuracy when compared to the manual annotations [18]. In this study, only images for which the GFP-tagged protein had been assigned to a single localization category were assessed. So far, this approach has not been used to systematically identify changes in protein localization in response to genetic or other perturbations in budding yeast. However, several open source software tools have been developed to aid biologists in designing classifiers for defining patterns or shapes of interest [19, 20]. These computational tools ought to stimulate more researchers to incorporate automated image and data analyses in future surveys of the GFP collection.

8 Conclusions

The budding yeast GFP collection is a unique tool for systematic analysis of protein function. Clever studies have been undertaken to detect changes in protein abundance in an automated fashion using flow cytometry and high-throughput microscopy. The GFP collection has also been surveyed for protein localization changes using manual inspection of cell images. Advances in high-throughput microscopy and automated image analysis mean that the GFP collection can now be used to rapidly acquire quantitative information about proteome dynamics in response to genetic and environmental perturbations. The non-invasive nature of these techniques allows for the monitoring of the proteome at single cell resolution over time. Being able to resolve information for individuals in a population allows us to detect changes that only occur in a subpopulation of cells; these changes would otherwise be missed in techniques that only analyze whole-population data (e.g., Western blot). The ability to follow protein localization and levels on a large scale and in an automated fashion is an essential step towards a quantitative description of biological pathways and processes.

References

1. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. Nature 425(6959):686–691. doi:10.1038/nature02026 nature02026 [pii]
2. Noree C, Sato BK, Broyer RM, Wilhelm JE (2010) Identification of novel filament-forming proteins in *Saccharomyces cerevisiae* and *Drosophila melanogaster*. J Cell Biol 190(4): 541–551. doi:jcb.201003001 [pii] 10.1083/jcb.201003001
3. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 441(7095):840–846. doi:nature04785 [pii] 10.1038/nature04785
4. Benanti JA, Cheung SK, Brady MC, Toczyski DP (2007) A proteomic screen reveals SCFGrr1 targets that regulate the glycolytic-gluconeogenic switch. Nat Cell Biol 9(10):1184–1191. doi:ncb1639 [pii] 10.1038/ncb1639
5. Springer M, Weissman JS, Kirschner MW (2010) A general lack of compensation for gene dosage in yeast. Mol Syst Biol 6:368. doi:msb201019 [pii] 10.1038/msb.2010.19
6. DeLuna A, Springer M, Kirschner MW, Kishony R (2010) Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLoS Biol 8(3):e1000347. doi:10.1371/journal.pbio.1000347
7. Shin CS, Kim SY, Huh WK (2009) TORC1 controls degradation of the transcription factor Stp1, a key effector of the SPS amino-acid-sensing pathway in *Saccharomyces cerevisiae*. J Cell Sci 122(Pt 12):2089–2099. doi:122/12/2089 [pii] 10.1242/jcs.047191
8. Narayanaswamy R, Levy M, Tsechansky M, Stovall GM, O'Connell JD, Mirrielees J, Ellington AD, Marcotte EM (2009) Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. Proc Natl Acad Sci USA 106(25):10147–10152. doi:0812771106 [pii] 10.1073/pnas.0812771106

9. Narayanaswamy R, Niu W, Scouras AD, Hart GT, Davies J, Ellington AD, Iyer VR, Marcotte EM (2006) Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. Genome Biol 7(1):R6. doi:gb-2006–7–1-r6 [pii] 10.1186/gb-2006–7–1-r6
10. Narayanaswamy R, Moradi EK, Niu W, Hart GT, Davis M, McGary KL, Ellington AD, Marcotte EM (2009) Systematic definition of protein constituents along the major polarization axis reveals an adaptive reuse of the polarization machinery in pheromone-treated budding yeast. J Proteome Res 8(1):6–19. doi:10.1021/pr800524g 10.1021/pr800524g [pii]
11. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7(10):R100. doi:gb-2006–7–10-r100 [pii] 10.1186/gb-2006–7–10-r100
12. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294(5550):2364–2368. doi:10.1126/science.1065810 294/5550/2364 [pii]
13. Vizeacoumar FJ, van Dyk N, F SV, Cheung V, Li J, Sydorskyy Y, Case N, Li Z, Datti A, Nislow C, Raught B, Zhang Z, Frey B, Bloom K, Boone C, Andrews BJ (2010) Integrating high-throughput genetic interaction mapping and high-content screening to explore yeast spindle morphogenesis. J Cell Biol 188(1):69–81. doi:jcb.200909013 [pii] 10.1083/jcb.200909013
14. Vizeacoumar FJ, Chong Y, Boone C, Andrews BJ (2009) A picture is worth a thousand words: genomics to phenomics in the yeast *Saccharomyces cerevisiae*. FEBS Lett 583(11):1656–1661. doi:S0014–5793(09)00272–5 [pii] 10.1016/j.febslet.2009.03.068
15. Collins TJ (2007) ImageJ for microscopy. Biotechniques 43(1 Suppl):25–30. doi:000112517 [pii]
16. Kuijper A, Heise B (2008) An automatic cell segmentation method for differential interference contrast microscopy. IEEE Pattern Recognition, 2008 ICPR 2008 19th International Conference: 1–4, Tampa, FL
17. Obara B, Veeman M, Choi JH, Smith W, Manjunath BS (2010) Segmentation of ascidian notochord cells in DIC timelapse images. Microsc Res Technol 74(8):727–734. doi:10.1002/jemt.20950
18. Chen SC, Zhao T, Gordon GJ, Murphy RF (2007) Automated image analysis of protein localization in budding yeast. Bioinformatics 23(13):i66–71. doi:23/13/i66 [pii] 10.1093/bioinformatics/btm206
19. Jones TR, Kang IH, Wheeler DB, Lindquist RA, Papallo A, Sabatini DM, Golland P, Carpenter AE (2008) CellProfiler analyst: data exploration and analysis software for complex image-based screens. BMC Bioinform 9:482. doi:1471–2105–9–482 [pii] 10.1186/1471–2105–9–482
20. Misselwitz B, Strittmatter G, Periaswamy B, Schlumberger MC, Rout S, Horvath P, Kozak K, Hardt WD (2010) Enhanced CellClassifier: a multi-class classification tool for microscopy images. BMC Bioinform 11:30. doi:1471–2105–11–30 [pii] 10.1186/1471–2105–11–30

Chapter 9
Unraveling the Complex Regulatory Relationships Between Metabolism and Signal Transduction in Cancer

Michelle L. Wynn, Sofia D. Merajver, and Santiago Schnell

Abstract Cancer cells exhibit an altered metabolic phenotype, known as the Warburg effect, which is characterized by high rates of glucose uptake and glycolysis, even under aerobic conditions. The Warburg effect appears to be an intrinsic component of most cancers and there is evidence linking cancer progression to mutations, translocations, and alternative splicing of genes that directly code for or have downstream effects on key metabolic enzymes. Many of the same signaling pathways are routinely dysregulated in cancer and a number of important oncogenic signaling pathways play important regulatory roles in central carbon metabolism. Unraveling the complex regulatory relationship between cancer metabolism and signaling requires the application of systems biology approaches. Here we discuss computational approaches for modeling protein signal transduction and metabolism as well as how the regulatory relationship between these two important cellular processes can be combined into hybrid models.

M.L. Wynn
Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
e-mail: mlwynn@umich.edu

S.D. Merajver
Department of Internal Medicine and Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
e-mail: smerajve@umich.edu

S. Schnell (✉)
Department of Molecular and Integrative Physiology, Center for Computational Medicine and Bioinformatics and Brehm Center for Diabetes Research, University of Michigan Medical School, Ann Arbor, MI, USA
e-mail: schnells@umich.edu

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_9, © Springer Science+Business Media, LLC 2012

1 Background

1.1 Cancer Systems Biology

Systems biology is the integration of theoretical and experimental methods to build a predictive model of a complex biological system. Tumor environments are extremely complex and encompass a large number of cells interacting with a changing microenvironment across a variety of spatial and temporal scales. Cancer systems biology, then, aims to understand the interactions that occur across microscopic and macroscopic scales in a tumor and, importantly, aims to exploit these interactions in a predictive way. Ideally, cancer models built using systems biology methods will have translational significance and can, for example, be used to predict rational therapeutic targets.

1.2 Cancer Signaling and Metabolism

Cancer cells exhibit an altered metabolic phenotype characterized by high rates of glucose uptake and glycolysis, even under aerobic conditions. This altered metabolism, first described by Warburg [1], is referred to as the Warburg effect and is so pervasive among cancers that it is routinely leveraged in the clinic with fluorodeoxyglucose-positron emission tomography (FDG-PET). In general, high tumor glucose uptake observed in FDG-PET scans correlates with poor prognostic outcome [2,3]. There is evidence to suggest that reliance on non-oxidative glycolytic metabolism sustains the biosynthetic requirements of rapid proliferation [2].

While the Warburg effect appears to be an intrinsic component of most cancer progressions, a precise etiology remains elusive. Both oncogenic signaling [4, 5] and interactions with the tumor microenvironment [6] play important roles in the induction of the malignant metabolic phenotype. For example, the activity of the M2 isoform of pyruvate kinase (PKM2), an important glycolytic enzyme, has been linked to the induction of the Warburg effect via tyrosine kinase signaling [7, 8].

Despite the enormous amount of genetic diversity found within a single tumor and across different cancers, many of the same signaling pathways are routinely dysregulated in cancer cells [9]. Importantly, many of these pathways have important downstream effects on metabolic behavior. For example, the phosphatidylinositol 3-kinase AKT pathway is commonly dysregulated in many human cancers [10]. AKT, a key component of this pathway, is known to play a critical role in stimulating glycolysis [11, 12]. In addition, there is evidence linking cancer progression to mutations, translocations, and alternative splicing of genes that directly code for or have downstream effects on key metabolic enzymes [13, 14].

It should be noted that there is some debate about whether increased glucose uptake translates into increased glycolytic flux and net glycolytic ATP gain in cancer cells [15]. It is possible that a significant amount of the glucose uptake

in cancer cells is shunted to pathways other than glycolysis (e.g., to the pentose phosphate pathway). Metabolic transformation, however, is increasingly recognized as an important hallmark of cancer [2, 16].

2 Modeling Intracellular Biochemical Processes

Because it is not practical to create models that are exact replicas of a complex system, trade-offs must be made between the scope and level of detail included in a model [17]. Complex cellular processes are commonly modeled with systems of continuous ordinary (ODE) or partial (PDE) differential equations. ODE and PDE models are built from underlying biophysical principles and, as a consequence, are inherently predictive. The use of continuous ODE-based approximations is justified when the system is assumed to be well mixed and the number of molecules of a given reactant ranges from 100 to 1,000 [18].

ODE-based systems, which are commonly applied to models of protein signal transduction and metabolism, are generally based on mass action and Michaelis–Menten (MM) kinetics [17, 19–21]. MM kinetics depends on the quasi-steady-state approximation, which assumes that the formation of the complex occurs on a much faster timescale than that of the other reactants. It is important, therefore, to recognize when these assumptions are invalid [22, 23].

An alternative to ODE-based kinetic models are stoichiometric models where the known structure of a chemical pathway is used to understand the state of the system under a set of specific conditions. Stoichiometric models have demonstrated predictive power using data from prokaryotes. The methods assume an optimization function (e.g., the goal of bacteria is continual production of biomass). Because these methods do not include any regulatory or kinetic information in the model formulation [24], they lack predictive power for multifunctional mammalian cells [25]. In our view, it would be extremely difficult to define an optimization function that adequately captures the complexity of a mammalian cell. Kinetic ODE models will, therefore, tend to be more predictive than stoichiometric methods because they can describe temporal dynamics. Kinetic ODE models require more knowledge *a priori* [24] than stoichiometric models, however, and this information is not always readily available.

At the other extreme are discrete logic-based Boolean models which provide a good approximation of the qualitative behavior of a biochemical system [26]. The motivation behind these models comes from the sigmoidal or hyperbolic dependence between regulatory molecules and the compounds they affect that can be thought of as having two states: saturated (“on”) and non-saturated (“off”), approximating a Boolean switch. In their simplest form, Boolean models are interaction networks where each biochemical species is represented as a node in one of two possible states: expressed (“on” or 1) or non-expressed (“off” or 0). Transfer functions between states are derived from biochemical interactions using logical operators (e.g., *AND*, *OR*, and *NOT*). In the transfer functions, there is

no notion of reaction rate and, hence, no need to estimate kinetic parameters. Despite this advantage, Boolean models have a major limitation: time is unrelated to physiological time and can provide only a qualitative chronology of molecular activations [27]. None the less, Boolean models can be important predictive tools in the absence of reliable kinetic data.

2.1 Modeling Metabolism

Many ODE-based models of glucose metabolism exist in the literature [28–31]. In general, metabolism is considered to be the set of chemical reactions catalyzed by enzymes operating in a living cell that are involved in catabolism or anabolism [19]. Enzymes regulate metabolism by catalyzing reactions [32]. Specifically, an enzyme reacts selectively with a substrate and transforms it into a product. In experimental studies of metabolism, enzyme concentrations are generally assumed to be constant during the catalyzed transformation of substrates into products [33, 34]. The majority of ODE-based metabolic models have focused on the dynamical behavior of subsets of central carbon metabolism (e.g., glycolysis or the pentose phosphate pathway). In our view, predictive models (especially in the context of cancer) should also consider the nature of the control mechanisms that regulate metabolism.

The most widely used theories of metabolic regulation are biochemical systems theory [35–37], metabolic control theory [38–40], and flux-oriented theory [41–43]. All three of these theories are in essence a form of sensitivity analysis applied to biochemical reaction models. The models consist of coupled ODEs based on the law of mass action. Sensitivity analysis is used to investigate the effects of parameter value changes on model behavior [44]. It is not surprising, then, that the primary difference between these theories is the choice of which parameters to vary when evaluating model sensitivity [44, 45].

In biochemical systems theory, the rate constants for the synthesis and degradation of metabolites are usually the parameters chosen for the sensitivity analysis. The metabolites are decomposed into dependent (substrate concentrations) and independent (enzyme concentration) variables where enzyme concentrations generally take constant values [46]. In metabolic control theory, the parameters for the sensitivity analysis are the enzyme activities. The sensitivity analysis gives rise to control coefficients, which are global pathway properties quantifying the control of overall metabolic flux by a single enzyme [45]. Enzyme concentrations are assumed to be constant and reaction rates are treated as constant parameters. Finally, in flux-oriented theory, sensitivities are calculated as the ratio of the relative change of the reaction rate (or flux) in response to a small internal or external stimulus. Enzyme concentrations are generally treated as constants in flux-oriented theory.

The assumption of constant enzyme concentration has been questioned for some time, however [47]. Enzymes are not indefinitely stable; they are metabolites like their substrates and products [19]. The synthesis of enzymes is an essential part of metabolism and is catalyzed by other enzymes. This phenomenon is known

as metabolic closure [48]: all catalysts essential for the survival of an organism must be synthesized internally. While the theory of metabolism-replacement has presented an abstract model of metabolic closure [49–51], it has limited practical applicability for the investigation of metabolic regulation [48] in the biomedical sciences. A theory to investigate metabolic regulation in cancer cells that takes into account enzyme production and depletion is critically needed in medicine.

2.2 Modeling Signal Transduction

A number of ODE-based models of signal transduction can be found in the literature [20, 21, 52–54]. In contrast to central carbon metabolism, however, significant information about the structure of signal transduction networks is often not known *a priori*. Alternative methods for modeling signal transduction include Bayesian network analysis, Markov models, and Boolean logic-based models [55].

As previously mentioned, a number of Boolean network models of gene regulation and signal transduction have generated experimentally valid predictions [26, 55–58]. In its simplest form, a Boolean model updates all nodes in a network at the same time, forcing all processes in the network to operate on identical timescales. This assumption results in a deterministic outcome similar to that of cellular automata. Boolean networks can be extended to utilize more biologically realistic variable timescales by performing asynchronous updates where nodes are selected at random and updated instantaneously [26]. Any given Boolean model will have one or more attractors or steady states each associated with a unique set of initial conditions (called its basin of attraction) that converge into that attractor [26, 57]. It is, therefore, possible to study the qualitative dynamical behavior of Boolean networks.

We would like to note that it is essential to carefully characterize the interactions included in any logic-based model. This is because the signaling dynamics of a network can be very different if an *OR* is used when an *AND* is needed. A detailed survey of the literature is required to build a reliable and robust logic-based model. For an example of the level of detail needed to justify each rule in a Boolean model, refer to the appendix in Albert and Othmer [56].

2.3 Linking Metabolic and Signal Transduction Models of Cancer

Metabolism and protein signaling do not operate in isolation. Gene expression and protein signal transduction have important downstream effects on metabolism, especially on metabolic enzyme synthesis. It is also likely that metabolite levels play a role in the regulation of gene transcription and protein translation.

How can we investigate and analyze the complex regulatory relationships between metabolic pathways and protein signaling in cancer? One possibility is to use large ODE-based models of protein signaling and metabolism without any of the simplifying assumptions made in the standard theories of metabolic regulation discussed above. Although this is theoretically possible, it would be a task for Laplace's Demon[1] because it would require a detailed knowledge of every chemical species, every interaction, and all associated rate constants involved in the reactions included in the model.

In practical terms, if the interactions are known, the law of mass action can be applied to derive an ODE system describing the pathways under consideration. Biochemical reaction dynamics are strongly dependent of parameter values. Central carbon metabolism, which is an essential part of tumor metabolism, has fortunately been well studied and characterized in mammalian cells. As a result, while experimentally and/or computationally intensive, methodologies exist for estimating kinetic parameters for metabolic networks [45]. This is less true of protein signaling networks largely due to their extreme complexity. While a tremendous amount of experimental work has identified a large number of protein interactions involved in both normal and malignant protein signaling, the kinetic details of these interactions are generally not known nor easily obtained. How, then, can we build predictive models that link cancer metabolism and protein signaling? One possibility is the use of hybrid models.

2.4 Hybrid Models

Hybrid models link discrete and continuous models across timescales and are widely used in the engineering and computational sciences. In models of tumor growth, cells can be modeled as discrete entities that respond to intracellular and extracellular signals which are modeled continuously [59–63]. For example, Ribba et al. [61] developed a multiscale model that linked a set of discrete models with continuous models of colorectal cancer growth. The model accounted for the cellular, genetic, and environmental factors regulating tumor growth. Key oncogenes involved in colorectal cancer evolution were integrated into a Boolean gene network regulated by a discrete cell cycle model. The response to signals from the intracellular gene network determined whether each cell proliferated or died and, therefore, directly influenced the cellular and the extracellular tissue scales. The spatial distribution of cells was computed using a continuous macroscopic tissue

[1]The idea of a Laplace Demon came from a thought experiment proposed by Pierre-Simon Laplace of a perfect entity who would know the precise location of each atom and of all forces in nature at any given moment. This entity (or demon, as it later came to be called) would have incredible predictive power because it could infer the past and determine the future from any set of initial conditions.

model based on Darcy's law. Finally, the number and spatial configuration of cells were used to activate antigrowth signals, which in turn were input into the Boolean model. This combination of discrete and continuous modeling was used to predict the qualitative effect of therapeutic protocols on colorectal cancer and demonstrated that the efficacy of irradiation protocols depends on the type of anti-growth signals to which tumors are exposed. Thus, a primary conclusion of this work was that the efficacy of irradiation therapy could be improved (without increasing radiation doses) by devising therapeutic schedules that take into account features of tumor growth through cell cycle regulation.

In a recent paper by Singhania et al. [64], a continuous model of the cell cycle was linked to a Boolean gene network model that regulated critical substrates involved in the progression of the cell cycle. By combining a continuous ODE model with a discrete Boolean model, the authors effectively obtained a piecewise ODE model system. In the model, each state was composed of a set of ODEs where specific species or parameters were null (or effectively "off") based on node values in the Boolean network.

In a similar manner, we propose that it is possible to combine ODE-based models of metabolism with discrete signaling models. While discrete and continuous hybrid models have been used in cancer research for more than 10 years, we are not aware of any that have directly linked metabolism and signal transduction. To successfully implement a hybrid model of this type, timescale separation will need to be carefully considered.

Comparing average protein half-life with average turnover in the number of enzyme molecules can provide insight into the separation of timescales needed in such a model. An assay of 100 proteins in living human cancer cells showed protein half-life range between 45 minutes and 22.5 hours [65]. The turnover numbers of most enzymes with their physiological substrates range from 1 to 1×10^5 substrate molecules converted into product molecules per second [66]. Using these numbers, we estimate that enzymes convert between 3.9×10^3 and 1.1×10^{10} substrate molecules into product molecules during their mean lifetime. Thus, due to the large difference in timescales, metabolic enzyme catalyzed reactions can be assumed to effectively operate under steady-state kinetics. If an enzyme concentration decreases, the steady state kinetics will change from a state of high enzyme steady-state kinetics to a low enzyme steady-state kinetics. Changes between these kinetic states will be driven by signal transduction pathways approximated in the discrete Boolean model.

3 Conclusion

Over the last 30 years much of cancer research has shifted to focus on molecular features of cancer and away from cancer metabolism and the Warburg effect. As a result, a wealth of experimental data now exists related to the role of gene and protein expression in cancer. Glucose uptake and metabolism are essential

features of cancer that, in our view, should be included in system level models of intracellular regulation (and dysregulation) in cancer. Developing theories that integrate this wealth of molecular information with experimental evidence related to cancer metabolism is the domain of cancer systems biology.

Of course, it is not practical to create models that exactly replicate the complexity of a tumor cell. Trade-offs, therefore, must be made between the scope and level of detail included in any model of cancer. Continuous ODE models are useful when kinetic information is available. When kinetic information is not available, logic-based Boolean models can be used to understand regulatory dynamics of known interactions from any set of initial conditions. A large number of regulatory interactions have been characterized in human cancers but the kinetic parameters governing the interactions are typically not known. As a result, Boolean models are useful tools for understanding the dynamics of these regulatory networks.

A theory to investigate the regulation of the malignant metabolic phenotype is critically needed. We suggest that hybrid models can be leveraged to integrate discrete Boolean signaling models with continuous metabolic models of cancer. Ideally this theory will also include aspects of existing control theories of metabolism. The ultimate goal of models built based on this theory will be to predict rational therapeutic targets that can be further experimentally validated.

Acknowledgments This work was partially supported by the University of Michigan Center for Computational Medicine & Bioinformatics Pilot Grant 2010 and the National Science Foundation under Grant No. DMS-1135663. MLW acknowledges support from the Rackham Merit Fellowship, NIH T32 CA140044, and the Breast Cancer Research Foundation. SDM acknowledges support from the Burroughs Wellcome Fund, Breast Cancer Research Foundation, the Avon Foundation and NIH CA77612. SS acknowledges support from NIDDK R25 DK088752.

References

1. Warburg O (1956) On the origin of cancer cells. Science 123(3191):309–314
2. DeBerardinis RJ, Sayed N, Ditsworth D, Thompson CB (2008) Brick by brick: metabolism and tumor cell growth. Curr Opin Genet Dev 18(1):54–61
3. Nakajo M, Jinnouchi S, Inoue H, Otsuka M, Matsumoto T, Kukita T, Tanabe H, Tateno R, Nakajo M (2007) FDG PET findings of chronic myeloid leukemia in the chronic phase before and after treatment. Clin Nucl Med 32(10):775–778
4. Hsu PP, Sabatini DM (2008) Cancer cell metabolism: Warburg and beyond. Cell 134(5): 703–707
5. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324(5930):1029–1033
6. Gillies RJ, Robey I, Gatenby RA (2008) Causes and consequences of increased glucose metabolism of cancers. J Nucl Med 49(Suppl 2):24S–42S
7. Hitosugi T, Kang S, Vander Heiden MG, Chung TW, Elf S, Lythgoe K, Dong S, Lonial S, Wang X, Chen GZ, Xie J, Gu TL, Polakiewicz RD, Roesel JL, Boggon TJ, Khuri FR, Gilliland DG, Cantley LC, Kaufman J, Chen J (2009) Tyrosine phosphorylation inhibits PKM2 to promote the Warburg effect and tumor growth. Sci Signal 2(97):ra73

8. Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC (2008) Pyruvate kinase M2 is a phosphotyrosine-binding protein. Nature 452(7184):181–186
9. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. Nat Med 10(8):789–799
10. Vivanco I, Sawyers CL (2002) The phosphatidylinositol 3-kinase AKT pathway in human cancer. Nat Rev Cancer 2(7):489–501
11. Buzzai M, Bauer DE, Jones RG, Deberardinis RJ, Hatzivassiliou G, Elstrom RL, Thompson CB (2005) The glucose dependence of Akt-transformed cells can be reversed by pharmacologic activation of fatty acid beta-oxidation. Oncogene 24(26):4165–4173
12. Elstrom RL, Bauer DE, Buzzai M, Karnauskas R, Harris MH, Plas DR, Zhuang H, Cinalli RM, Alavi A, Rudin CM, Thompson CB (2004) Akt stimulates aerobic glycolysis in cancer cells. Cancer Res 64(11):3892–3899
13. Thompson CB (2009) Metabolic enzymes as oncogenes or tumor suppressors. New Engl J Med 360(8):813–815
14. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC (2008) The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature 452(7184):230–233
15. Zu XL, Guppy M (2004) Cancer metabolism: facts, fantasy, and fiction. Biochem Biophys Res Commun 313(3):459–465
16. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674
17. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. Nat Cell Biol 8(11):1195–1203
18. Turner TE, Schnell S, Burrage K (2004) Stochastic approaches for modelling in vivo reactions. Comput Biol Chem 28(3):165–178
19. Cornish-Bowden A (2004) Fundamentals of enzyme kinetics, 3rd edn. Portland, London
20. Goldbeter A, Koshland DE Jr (1981) An amplified sensitivity arising from covalent modification in biological systems. Proc Natl Acad Sci USA 78(11):6840–6844
21. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7(3):165–176
22. Flach EH, Schnell S (2006) Use and abuse of the quasi-steady-state approximation. Syst Biol 153(4):187–191
23. Segel LA (1984) Modeling dynamic phenomena in molecular and cellular biology. Cambridge University Press, Cambridge
24. Gombert AK, Nielsen J (2000) Mathematical modelling of metabolism. Curr Opin Biotechnol 11(2):180–186
25. Heinrich R, Schuster S (1996) The regulation of cellular systems. New York
26. Albert I, Thakar J, Li S, Zhang R, Albert R (2008) Boolean network simulations for life scientists. Source Code Biol Med 3:16
27. Thomas R, D'Ari R (1990) Biological Feeback. CRC, Boca Raton
28. Goldbeter A, Lefever R (1972) Dissipative structures for an allosteric model. Application to glycolytic oscillations. Biophys J 12(10):1302–1315
29. Sel'kov EE (1968) Self-oscillations in glycolysis. 1. A simple kinetic model. Eur J Biochem 4(1):79–86
30. Heinrich R, Rapoport SM, Rapoport TA (1977) Metabolic regulation and mathematical models. Prog Biophys Mol Biol 32(1):1–82
31. Heinrich R, Schuster S (1996) The regulation of cellular systems. Chapman & Hall, New York
32. Fersht A (1999) Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding. WH Freeman, New York
33. Schnell S, Maini PK (2003) A century of enzyme kinetics: reliability of the K_M and v_{max} estimates. Comm Theor Biol 8(2–3):169–187
34. Cook PF, Cleland WW (2007) Enzyme kinetics and mechanism. Garland Science, London
35. Savageau MA (1969) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. J Theor Biol 25(3):365–369

36. Savageau MA (1969) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. J Theor Biol 25(3):370–379
37. Savageau MA (1970) Biochemical systems analysis. III. Dynamic solutions using a power-law approximation. J Theor Biol 26(2):215–226
38. Kacser H, Burns JA (1973) The control of flux. Symp Soc Exp Biol 27:65–104
39. Heinrich R, Rapoport TA (1974) A linear steady state treatment of enzymatic chains: general properties, control and effector strength. Eur J Biochem 42(1):89–95
40. Heinrich R, Rapoport TA (1974) A linear steady state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector. Eur J Biochem 42(1):97–105
41. Crabtree B, Newsholme EA (1978) Sensitivity of a near-equilibrium reaction in a metabolic pathway to changes in substrate concentration. Eur J Biochem 89(1):19–22
42. Crabtree B, Newsholme EA (1985) A quantitative approach to metabolic control. Curr Top Cell Reg 25:21–76
43. Crabtree B, Newsholme EA (1987) The derivation and interpretation of control coefficients. Biochem J 247(1):113–120
44. Varma A, Morbidelli M, Wu H (1999) Parametric sensitivity in chemical systems. Cambridge University Press, New York
45. Fell D (1997) Understanding the control of metabolism. Frontiers in metabolism. Portland, London
46. Voit EO (2000) Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists. Cambridge University Press, New York
47. Cornish-Bowden A, Cardenas ML (2005) Systems biology may work when we learn to understand the parts in terms of the whole. Biochem Soc Trans 33(3):516–519
48. Piedrafita G, Montero F, Morán F, Cárdenas ML, Cornish-Bowden A (2010) A simple self-maintaining metabolic system: Robustness, autocatalysis, bistability. PLoS Comput Biol 6(8):e1000872
49. Rosen R (1991) Life itself: a comprenhensive inquiry into the nature of origin and fabrication of life. Columbia University Press, New York
50. Cornish-Bowden A, Cárdenas ML, Letelier JC, Soto-Andrade J (2007) Beyond reductionism: metabolic circularity as a guiding vision for a real biology of systems. Proteomics 7(6):839–845
51. Letelier JC, Soto-Andrade J, Guíñez Abarzúa F, Cornish-Bowden A, Luz Cárdenas M (2006) Organizational invariance and metabolic closure: analysis in terms of (M, R) systems. J Theor Biol 238(4):949–961
52. Huang CY, Ferrell JE Jr (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci USA 93(19):10078–10083
53. Ventura AC, Jiang P, Van Wassenhove L, Del Vecchio D, Merajver SD, Ninfa AJ (2010) Signaling properties of a covalent modification cycle are altered by a downstream target. Proc Natl Acad Sci USA 107(22):10032–10037
54. Ventura AC, Sepulchre JA, Merajver SD (2008) A hidden feedback in signaling cascades is revealed. PLoS Comput Biol 4(3):e1000041
55. Aldridge BB, Saez-Rodriguez J, Muhlich JL, Sorger PK, Lauffenburger DA (2009) Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. PLoS Comput Biol 5(4):e1000340
56. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol 223(1):1–18
57. Li S, Assmann SM, Albert R (2006) Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. PLoS Biol 4(10):e312
58. Zhang R, Shah MV, Yang J, Nyland SB, Liu X, Yun JK, Albert R, Loughran TP Jr (2008) Network model of survival signaling in large granular lymphocyte leukemia. Proc Natl Acad Sci USA 105(42):16308–16313

59. Anderson AR, Quaranta V (2008) Integrative mathematical oncology. Nat Rev Cancer 8(3):227–234
60. Alarcon T, Byrne HM, Maini PK (2004) A multiple scale model for tumor growth. Multiscale Model Sim 3(2):440–467
61. Ribba B, Colin T, Schnell S (2006) A multiscale mathematical model of cancer, and its use in analyzing irradiation therapies. Theor Biol Med Model 3:7
62. Ribba B, Saut O, Colin T, Bresch D, Grenier E, Boissel JP (2006) A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents. J Theor Biol 243(4):532–541
63. Frieboes HB, Chaplain MA, Thompson AM, Bearer EL, Lowengrub JS, Cristini V (2011) Physical oncology: a bench-to-bedside quantitative and predictive approach. Cancer Res 71(2):298–302
64. Singhania R, Sramkoski RM, Jacobberger JW, Tyson JJ (2011) A hybrid model of mammalian cell cycle regulation. PLoS Comput Biol 7(2):e1001077
65. Eden E, Geva-Zatorsky N, Issaeva I, Cohen A, Dekel E, Danon T, Cohen L, Mayo A, Alon U (2011) Proteome half-life dynamics in living human cells. Science 331(6018):764–768
66. Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry. 5th edn. WH Freeman, New York

Part II
Cellular Decision Making: Adaptation, Differentiation and Death

Chapter 10
The Cell as a Thermostat: How Much does it Know?

Dennis Bray

Abstract How does bacterial thermotaxis compare to a simple wall thermostat? Elements with similar function can be found in the two, including a temperature-sensing element, an output switch, and an external control. But they differ in their origins. A thermostat is designed and made by humans and embodies their understanding of seasonal fluctuations in temperature and how these affect room comfort. By contrast, the bacterial system is self-contained and assembles according to information in its genome acquired by evolution. This information is far richer than anything carried by a thermostat and closer to the 'knowledge' that higher animals have about the world.

When cells of the common gut bacterium *Escherichia coli* are taken from a growing culture with density less than about 2×10^8 cells per ml and placed in a chamber with temperature ranging from 18°C at one end to 30°C at the other, they swim to the warm end [1,2]. This response, termed thermotaxis, is a specific adaptation quite distinct from the universal effects of temperature on the rates of enzyme reactions and other processes. It is advantageous to the bacteria since it allows them to avoid extremes of hot or cold and seek an optimal temperature for growth. Evidently, these simple cells have the capacity to sense a change in temperature and produce a response – in this case a change in motility.

A physicist friend observing this behaviour might then say that the cell is just behaving like a thermostat on a wall. It is a dumb device that blindly changes state with temperature and thereby turns on or off some other process in the cell... end of story. There is no reason to suppose, he or she might declare, that the cell "knows" about temperature, or that particular temperature states are "better"

D. Bray (✉)
Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
e-mail: db10009@cam.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_10,

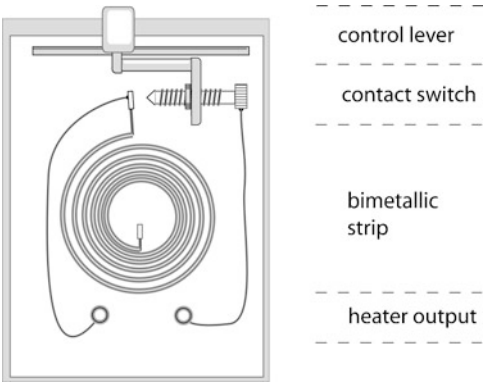


Fig. 10.1 Conventional room thermostat

or "more desirable" than another, or that the cell "expects" certain temperatures to be more likely than others. Indeed, the very use of these terms is egregiously anthropomorphic. They impute to a simple bacterium capacities such as knowledge of the world and of individual requirements that are truly only present in humans.

In this short essay I argue, by contrast, that what the cells are doing in the above experiment is far from dumb. Indeed I believe their behaviour reveals the presence of a large body of specific information relating to the natural variations in temperatures in the world and their potential significance for bacterial survival. This information is far richer and more sophisticated than anything carried by a thermostat and much closer to the "knowledge" that higher animals and humans have about the world, although without the conscious component.

Let me start by considering a typical household thermostat such as might be used to control a gas-fired boiler. The temperature-sensing element of this device is a bimetallic strip wound into a coil, anchored at one end and free to move at the other. Because the two metals making up the coil have different thermal expansions, a change in temperature causes the coil to tighten or loosen and its free end to move. In the configuration shown in Fig. 10.1, a falling temperature drives the free end clockwise until it encounters a contact point at the end of a screw; an electric circuit is thereby completed causing the boiler to operate. The temperature at which the switch operates is regulated by the position of the grub screw set by the manufacturer, and by a control lever operated from the room by the occupier. A typical thermostat also includes a magnet (not shown in the figure) close to the point of contact to ensure good contact and to avoid hunting oscillations of a few degrees.

The device portrayed in Fig. 10.1 is indeed no more than a collection of inert metal parts. It has a function, but surely no one would claim that by itself it knows anything about temperature or the boiler or the room. However the situation changes when we remember where the thermostat came from. For it was made and used by humans and for that reason embodies their understanding of the world. Whoever designed the thermostat knew the range of temperatures over which it would need to operate. He or she would have chosen the materials, the dimensions and other specifications needed to switch on and off at the right temperature and

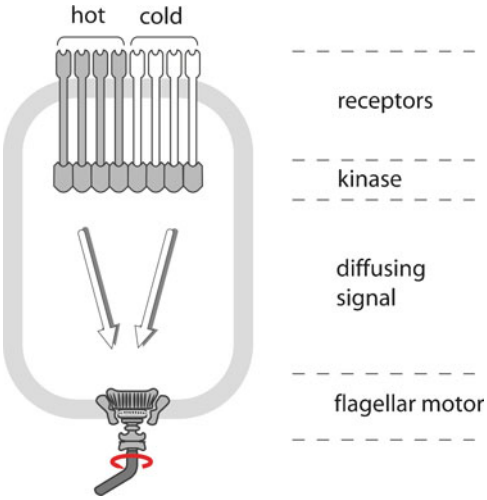


Fig. 10.2 Thermosensitive response of *Escherichia coli*. The cell membrane is represented by a *closed grey line* into which chemoreceptors are inserted like thin probes. A typical cell has several thousand copies of five subtypes, which detect nutrients and poisons in the environment. Depending on the type of receptor and its level of methylation, it can also serve either as heat-avoiding ("hot") or cold-avoiding ("cold"). The layer termed "kinase" contains not only the kinase CheA but also the proteins CheW, CheZ, CheR and CheB. *Two arrows* indicate diffusion of the small protein CheY through the cytoplasm. In its phosphorylated form, CheY binds to the inner face of the flagellar motor causing it to turn clockwise and initiate a tumble. For details of the signal cascade, see [4,5]

with appropriate dynamics for the downstream output device. The designer would know what settings would be likely to be useful in different climates and seasons and allowing for personal preferences, to be adjusted by the manufacturer (grub screw) and the room user (sliding lever). In other words, if we were to wrap together the physical device with the humans that designed, made and used it, then we would indeed have something that carries a form of knowledge about room temperature and its effect on human inhabitants.

Now consider the temperature-sensing system of the bacterium. Despite the fact that it is a million times smaller than the thermostat and made of proteins rather than metal, we can identify parts with similar function (Fig. 10.2). The temperature-sensitive element itself – equivalent to the coiled bimetallic strip – is a cluster of proteins, chemoreceptors, inserted in the bacterial membrane. In common with most protein molecules, these receptors can exist in different shapes, or conformations, each with a distinctive arrangement of atoms. Transitions between protein conformations typically occur in response to changes in the environment such as changes in acidity or the concentrations of small molecules. In the present case, the receptors also change with temperature, in a manner that can be tuned by the environment. Depending on which out of five possible kinds the receptor belongs to and how much it has been modified in the cell by the addition of methyl groups, it can be activated by heat, activated by cold, or be insensitive to temperature.

The effect of these changes – the response of the "device" – is a change in swimming. Swimming entails rotation of flagella on the cell surface and may be either counterclockwise, in which case the cell progresses smoothly in the current direction, or clockwise when it undergoes random changes in direction (termed a tumble) [3]. Tumbles are triggered when the chemoreceptors become active and send a diffusible signal to the motors. Thus, if an individual bacterium moves by chance into a region of excessive temperature, the "heat avoiding" receptors become active and the cell initiates a tumble and changes path. Similarly, if the bacterium wanders into an excessively cold region, "cold avoiding" receptors kick in and the cell will again stall. Only at an optimal intermediate temperature where all receptors are inactive will the cell progress smoothly in a forward direction. For the cells described in the above experiment this range is close to body heat, which is why they swim toward 30°C in the test chamber.

The "switch" of the bacterial device – equivalent to the contact between the coil and grub screw – is the point at which the receptors interface with the downstream signalling proteins. This is provided by a layer of proteins attached to the cytoplasmic tails of the receptors that includes molecules of a kinase – an enzyme that transfers phosphoryl groups from ATP to other proteins. The kinase is turned on whenever its associated receptors are active and then sends a signal to the flagellar motors telling them to turn clockwise and hence generate a tumble. (The signal is carried by another protein, CheY, which receives an active phosphoryl group from the kinase and diffuses through the cytoplasm to the inner face of the motors – see [4, 5]). Thus, if the bacterium experiences a very high or very low temperature, one or other set of receptors will become active and the bacterium will tumble, or stall.

There is even something analogous to the thermostat's control lever. If the experiment mentioned in the first paragraph is performed with cells taken from a dense rather than a sparse culture, then bacteria reverse their preference and swim to the cooler end, set at 18°C [1]. The reason for this curious behaviour is thought to be that in a crowded environment it is actually advantageous for cells to grow more slowly – when food is running short, they are best advised to congregate into stress-resistant colonies and wait until better conditions arrive. The mechanism of inversion depends on modifications in the amounts and efficiency of the receptors mentioned above. As the culture grows more crowded, cells start to make more hot-avoiding receptors. The remaining cold-avoiding receptors also become less effective due to a form of adaptation, caused by the accumulation of the amino acid glycine in the medium. As a consequence of these two changes, the optimum temperature at which the cells swim most efficiently falls and in the above experiment they accumulate at the 18°C end of the chamber.

So like the thermostat, the bacterial cell has (1) a temperature-sensing element; (2) an output switch that is activated only at certain temperatures; and (3) a control by which the critical temperatures can be altered. (There is nothing equivalent to the magnet, so far as I know, but perhaps this is not necessary at the molecular level.) Considered as a simple on–off device, therefore, the cell is indeed comparable to a thermostat on the wall. But when we inquire into its, origins the situation is entirely

different, for the cell was not created by a designer either human or divine, nor is its response continually adjusted by some outside user according to requirements. These functions are supplied not by humans as in the case of the wall thermostat but by the cell itself.

Where does the cell acquire its information? It can only be from the parental cell through the process of cell division. And where did the parental cell and antecedent generations acquire their information? Why, by evolution. Over uncounted millennia the ancestors of present day *E. coli* struggled to survive in a cruel and capricious world. Temperature was a persistent and ever-changing feature of their environment – one that, if correctly interpreted, could provide life-saving clues about how to respond. An increase in ambient temperature, for example, might indicate the presence nearby of a mammal and a source of food; a low temperature might be associated with flowing water and a risk of being swept away. Or, as we have seen, it could signal a desirable place to cool down under crowded conditions. Any bacterium that learned (in an evolutionary sense) to read these signs, especially in conjunction with other indicators, would have an advantage. The first step might have been mutations that by chance installed a sequence of amino acids that produced a particularly large change in structure of a particular protein with temperature. Changes in other proteins would then have followed allowing them to interact and respond to the temperature-sensitive protein. Eventually, a complicated chain of biochemical causation could have been built up by which ambient temperature influenced processes such as motility or metabolism [6].

The nucleotide sequences in *E. coli* DNA that encode the chemoreceptors and downstream proteins, therefore, perform a similar function to the designer and manufacturer of a thermostat. They determine the range of temperatures over which the temperature-sensor needs to operate; they specify the dimensions and other parameters needed for the sensor to switch on and off at the right temperature and with appropriate dynamics for the downstream output device. The DNA carries information relating to the settings a future (bacterial) user will likely need in different climates and seasons and perhaps even allow for individual preferences arising, say, from the nutritional state of the cell or its stage of division. The machinery of protein synthesis and assembly – all of it intrinsic to the cell – uses this information to produce actual functioning molecular parts. No need here for plans produced in an office to be sent to a manufacturing facility for assembly and distribution. It all happens in the same minute volume of cytoplasm, unceasingly and without outside intervention.

A bacterium is, therefore, much more sophisticated than a thermostat. Not only does it possess physical parts that detect and relay changes in temperature but it also carries the information needed to specify and build the device, closely resembling in this respect the role of a human designer. Indeed, since the bug and the human are both products of evolution, one could say they acquired their understanding from the same source, albeit by a very different path.

But what words can we use to describe this cellular information? Contemporary biology embraces reductionism and eschews vitalism. It has been inordinately successful in revealing the structures and functions of biological molecules, often at

an atomic scale. But it has left us with an extreme, almost puritanical rejection of any account of biological processes that goes beyond physics or chemistry. This is why a term such as "knowledge" arouses antipathy when applied to cells or simple organisms – because it is freighted with human connections. Human knowledge is consciously accessible and can be expressed in language and there is no exact equivalent in a bacterium or other biological system. Our physics friend might, therefore, be driven to employ phrases such as "the bacterium has competence to respond to temperature".

But please consider that the dictionary definition of "knowledge" includes the more generic meaning of "specific information about a subject". A library can be said to contain a body of knowledge. The multiple databases accessed by IBM's Watson computer in its recent successful performance in the quiz show Jeopardy! were a source of knowledge. Even human knowledge is not always conscious – the complex sequence of muscle actions a child learns to use when riding a bicycle, for example. Living organisms, from bacteria to humans, carry an enormous legacy of information acquired through evolution. They draw on this information as they grow and interact with the world and it enables them to act in a manner that is beneficial to their eventual survival. A physicist or chemist examining this or that process or set of molecules in isolation usually has little comprehension of its complete function in a living organism. Who does? But considered in context, every protein is enmeshed in a dense thicket of interactions, actual or potential, that have supported survival under the myriad of situations encountered by the organism and its predecessors. When it is challenged, the cell or organism accesses this information and produces an appropriate movement or other response. To me, the most natural way to describe this behaviour is simply to say: yes, the bacterium *knows* about temperature and what it means for its survival.

Acknowledgement I would like to thank Matthew Levin, Ralph Linsker, Jim Shapiro, Kate Storey and Yuhai Tu for insightful comments.

References

1. Salman H, Libchaber A (2007) A concentration-dependent switch in the bacterial response to temperature. Nat Cell Biol 9(9):1098–1100
2. Sourjik V, Wingreen NS (2007) Turning to the cold. Nat Cell Biol 9(9):1029–1031
3. Berg HC (2004) *E. coli* in motion. In: Greenbaum E (ed) Biological and medical physics biomedical engineering. Springer, New York
4. Wadhams GH, Armitage JP (2004) Making sense of it all: bacterial chemotaxis. Nat Rev Mol Cell Biol 5:1024–1037
5. Hazelbauer GL, Falke JJ, Parkinson JS (2008) Bacterial chemoreceptors: high-performance signaling in networked arrays. Trends Biochem Sci 33(1):9–19
6. Tagkopoulos I, Liu Y-C, Tavazoie S (2008) Predictive internal representations underlie anticipatory behavior within microbial genetic networks. Science **320**:1313–1317

Chapter 11
Stem Cell Differentiation as a Renewal-Reward Process: Predictions and Validation in the Colonic Crypt

Kiran Gireesan Vanaja, Andrew P. Feinberg, and Andre Levchenko

Abstract Stem cells serve as persistent reservoirs for replenishment of rapidly renewing tissues, frequently also ensuring that the correct tissue morphology is maintained. This process is inherently stochastic due to the small number and stochastic division patterns within the stem cell compartments, as well as the essentially stochastic differentiation events that follow the initial stem cell expansion. Here we propose a new formalism to describe this process, by employing the approach known in statistics as the renewal-reward process. Using this approximation allows application of the mathematical apparatus developed for renewal-reward processes to the stochastic stem cell biology. We show in the context of colonic crypts that the resulting predictions match the experimental results, while also providing a convenient tool for analysis of normal and abnormal differentiation processes.

1 Introduction

Tissue renewal is one of the most important aspects of systemic homeostasis. Organs like the intestine, blood, skin, etc. undergo various stages of massive renewal throughout the day in a living multicellular organism, such that the cycle of losing cells that perform the function of the organ to various environmental challenges is intricately balanced by the renewal of the organ such that the organ maintains its

K.G. Vanaja • A. Levchenko (✉)
Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
e-mail: kirangv@jhu.edu; alev@bme.jhu.edu

A.P. Feinberg
Department of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA
e-mail: afeinberg@jhu.edu

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_11,

size and functional organization throughout its life [1]. A considerable majority of the organs employ the classic stem—transit amplifying—differentiated cells renewal program wherein a small pool of relatively slowly proliferating stem cells [2–4] ensure genomic integrity and form an indefatigable, replenishable source of more functional differentiated cells. As dictated by the architecture of the organ and the needs of the organ, dividing stem cells can, with defined probabilities, either self-renew or differentiate into transit-amplifying cells, which no longer have the stem cell property. These transit-amplifying cells then rapidly proliferate, make up the core of the structure of the organ and after a certain number of cycles, fully differentiate into the functional cells of the organ. Needless to say, the stem cells, by virtue of their unlimited capacity for cell division, dictate the dynamics of the cell numbers in the crypt. In summary, given the need to maintain genomic integrity and geometrically defined structure of self-renewing organs, stem cells divide infrequently and also employ both symmetric and asymmetric cell divisions to perform the dual nature of maintenance of the stem cell pool and the required number of progeny to differentiate and perform the functional roles of the organ [5,6].

The crypts in the colon are perhaps the most well-characterized and studied examples of epithelial tissue renewal system. Colonic crypts are minute finger like invaginations in the colonic epithelium that are composed of a single layer of cells and are shaped in the form of a test tube [7]. The crypts provide a continuous stream of cells mainly enterocytes and goblet cells to the colonic epithelium and are responsible for replenishing the same as they slough off into the lumen of the colon. The crypt is organized into a small pool of stem cells at the bottom [8,9], a large pool of rapidly proliferating cells along the length of the crypt which constitute the transit-amplifying pool, and a pool of fully differentiated nonproliferating cells along the upper third of the length of the crypt. As the self-renewing stem cells divide, the newly formed daughter cells advance up along the length of the crypt and partially differentiate into the cells of the transit-amplifying compartment. These partially differentiated cells in the transit-amplifying compartment proliferate rapidly and are responsible for supplying differentiated cells to the colonic epithelium at the rate at which they are lost into the lumen of the colon. As the cells in the transit-amplifying compartment rise up the length of the colon, they stop proliferating, acquire very differentiated phenotypes, and become the functional cells of the colonic epithelium [7].

Noise is inherent in all biological systems, most often manifesting itself as stochasticity in measured responses and observed values of system variables. The stochasticity can be a function of many known and unknown processes and serves to impart useful properties to the system like robustness, variability, ability to recover from crippling errors, and such [10]. In terms of the colonic crypt, there is inbuilt stochasticity in almost every aspect of the dynamics and organization of the crypt, i.e., in the density of the number of crypts per unit area of the colonic epithelium, in the number of levels of the cylindrical organization of the crypt, in the total number of cells in each crypt, in the number of stem cells, number of transit-amplifying cells, in the cell cycle time of cells in the different compartments,

effectively from the tissue level organization to the cellular details of cell cycling and proliferation [11].

Many attempts have been made to develop mathematical models to understand the homeostasis and aberrant behavior, in the case of cancer, of the colonic crypts [12–14]. Most of the models that have been recently developed are deterministic in nature in that cell cycle analysis, number of stem cells, etc. are defined by ordinary differential equations [11]. Although they have been used in other population based systems truly stochastic models have not really been used to describe the behavior of tissue renewal or in this case crypt organization and homeostasis. Stochastic models have the advantage of being able to model the colonic crypt in its true form and account for the noise, randomness, and inherent variability that are evident when repeated measurements are made of the crypt parameters. They also have the added advantage that the average behavior of system parameters predicted by the stochastic models will correspond to the values predicted by the deterministic models.

Renewal process is a stochastic random process that is a generalization of the Poisson process [15]. It is a counting process on the space of integers and is defined by the time intervals between the occurrence of events and the probabilities of the counting at each such event occurrence. A renewal-reward process can be defined on the same renewal process with the incremental changes in the renewal process constituting the counting process [15]. As the main events that occur in the stem cell pool are the spontaneous mitosis of the cells into two daughter cells and the probabilistic event of those cells choosing to either stay as stem cells or differentiate [16], it is quite easy to note the similarity between the events in the stem cell pool and the renewal process. Furthermore, the main function of the process, formation of appropriate numbers of differentiated functional cells, can be seen as a reward. We can then attempt to model the number of cells in the stem cell pool as a renewal process and the process of differentiation whereby these cells lose their stem cell properties and become the transit-amplifying cells as a renewal-reward process.

In addition to the simplicity of representation and the inherent stochasticity associated with it, both features being very useful when dealing with stem cell differentiation process, the use of renewal theory enables us to take advantage of the vast number of results and theorems developed independently of description of biological processes. As shown below, by using the fundamental theorem of renewal-reward process, for example, we can quite easily derive an equation for the average rate of cells differentiating and exiting the stem cell pool in terms of other parameters that define the stem cell pool. We then use this result to test the validity of a prediction on the cell division time of the stem cells which to date remains not conclusively proven.

2 Model Description

The model follows the stem-transit amplifying-terminally differentiated program of the crypt as shown in Fig. 11.1. The stem cell pool has no input and is self-contained. Stem cells can divide into two daughter cells and hence generate an

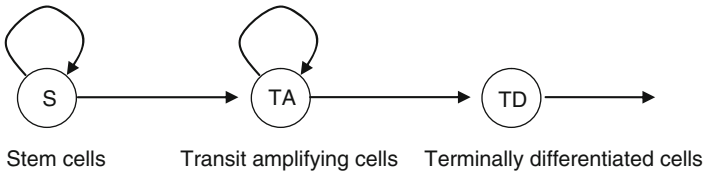


Fig. 11.1 Model description of the stem-transit amplifying-terminally differentiated cells renewal program of the colonic crypt homeostasis

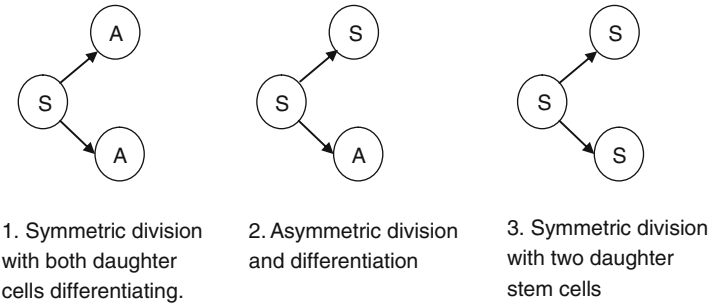


Fig. 11.2 The three forms of cell division in the stem cell pool

event, the dividing cells can either remain as stem cells or differentiate and become transit-amplifying cells. The transit-amplifying cells proliferate rapidly and stay as transit-amplifying cells for integral number of cell cycles and then differentiate into the terminally differentiated cells.

2.1 Stem Cell Compartment

The stem cell pool is composed of cells that proliferate slowly, renew indefinitely and are responsible for the maintenance of the crypt. The stem cells reside at the bottom of the crypt and are relatively few in number. Owing to the inherent stochasticity and the probabilistic nature of cell division and differentiation, the number of stem cells in a crypt at any given time instant of is a random variable with a mean value and variance. Consequently, as a function of time the number of stem cells in a crypt can be described as a random process. Let $N_S(t)$ be that random process. Events in this compartment are generated when a stem cell undergoes mitosis, say at $t = t_a$, and at each cell division one of three things can happen as shown in Fig. 11.2,

1. $N_S(t_a)$ decreases by 1 to $N_S(t_a) = N_S(t) - 1$, where $t < t_a$, a case of symmetric division and when both the daughter cells differentiate and leave the compartment,
2. $N_S(t_a)$ remains the same, $N_S(t_a) = N_S(t)$, where $t < t_a$, a case of asymmetric division and when one of the daughter cell stays in the compartment while the other daughter cell differentiates and leaves the compartment and,
3. $N_S(t_a)$ increases by 1 to $N_S(t_a) = N_S(t) + 1$, where $t < t_a$, a case of symmetric division where both the daughter cells remain as stem cells and stay in the compartment.

If the time sampling is sufficiently fine, we can avoid the situation of two cells dividing at the same time and thus simplify the system specification.

Let $\{t_0, t_1, t_2, \ldots t_n, t_{n+1}, \ldots\}$ be the time instants at which events occur in the stem cell compartment and let $\{S_1, S_2, S_3 \ldots, S_n, S_{n+1}, \ldots\}$ be the corresponding time intervals between successive cell divisions such that $S_1 = t_1 - t_0$ is the time to the first division and S_2 is the time interval between the second and the first cell division and so on. We can safely assume that S_1, S_2, etc. are a sequence of independent, identically distributed random variables such that $0 < E[S_i] < \infty$ be true $\forall\ i$, where $\boldsymbol{E}[.]$ is the expectation operator.

Let $I(t)$ be the indicator function that gives the jump in $N_S(t)$ at each time instant, such that $N_S(t_i + \Delta) - N_S(t_i - \Delta) = I(t_i)$ be the height of the jump at $t = t_i$ for any arbitrarily small Δ. $I(t)$ is the amount by which N_S changes, i.e., by $0, +1$, or -1 corresponding to the three outcomes of cell division. Also, let J_I be such that $J_I = \sum_{n=0}^{I} S_n$, the sum of all time intervals between cell divisions (also called the holding times in random processes parlance). $N_S(t)$ defined as $N_S(t) = \sum_{n=1}^{\infty} I(J_n \leq t)$ is a renewal process. Assuming wide-sense stationarity/time invariance $\bar{N}_S(t) = E[N_S(t)] = \bar{N}_S$ is the average number of stem cells expected to be found in a crypt.

Let us define $\{W_1, W_2, W_3, \ldots W_n, W_{n+1}, \ldots\}$ as the events that lead to stem cells differentiating and leaving the compartment. This happens when (1) when both the daughter stem cells after division differentiate and leave the compartment corresponding to a jump of -1 in $N_S(t)$ and (2) when one daughter stem cell differentiates and leaves the compartment corresponding to a jump of 0 in $N_S(t)$. Since its reasonable to assume that the probabilities of symmetric and asymmetric differentiation in crypt stem cells do not change with time, $\{W_1, W_2, W_3, \ldots. W_n, W_{n+1}, \ldots\}$ is a set of independent and identically distributed random variables. So W_1 and its sample space can be defined as

$$W_1 = \left\{ \begin{array}{cc} +2 & \text{both daughter stem cells leave compartment} \\ +1 & \text{one daughter stem cell leaves the compartment} \\ 0 & \text{none of the daughter cells leave the compartment} \end{array} \right\}$$

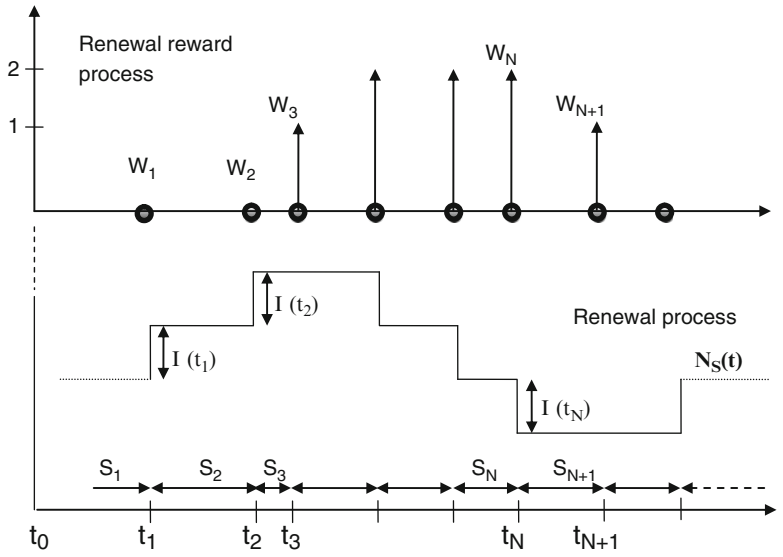


Fig. 11.3 The renewal process and renewal-reward process: every time a stem cell divides, based on either of the three cell division outcomes highlighted in the text $N_S(t)$ changes by the corresponding $l(t)$ which then forms the basis of the renewal-reward process

and the associated probabilities can be written as,

$$p_{W_1} = \left\{ \begin{array}{l} p \text{ for } W_1 = +2 \\ q \text{ for } W_1 = +1 \\ r \text{ for } W_1 = 0 \end{array} \right\}; \quad \text{where } p + q + r = 1$$

Also let $E[W_1] = \sum_W W_1 p w_1$ be such that $0 < E[W_1] < \infty$, then the random variable $Y_t = \sum_{i=1}^{t} W_i$ defined over $\{W_1,\ W_2,\ W_3,\ \ldots.\ W_n,\ W_{n+1},\ \ldots\}$ and $\{S_1,\ S_2,\ S_3,\ \ldots.\ S_n,\ S_{n+1},\ \ldots\}$ is a *renewal-reward process* (Fig. 11.3). The convenience of formulating the cells differentiating and exiting the stem cells compartment as a renewal-reward process is in being able to relate fundamental biological quantities of the crypt dynamics using fundamental renewal-reward theorems.

The important parameters that define the stem cell compartment are the number of stem cells at any point in time, the average cycling time of stem cells, and the probabilities of symmetric and asymmetric division. Directly from the model definition, the number of stem cells in the compartment is given directly by $N_{SC}(t)$, the probabilities of division are given by p, q, and r and as shown below the average cycling time can be expressed in terms of these parameters.

From the fundamental theorem of elementary renewal-reward processes, we have

$$\lim_{t \to \infty} \frac{E[Y(t)]}{t} = \frac{E[W_1]}{E[S_1]}$$

The equation basically relates the time average or rate of exit of stem cells (the left hand side) from the stem cell compartment to the probabilities of the various mitosis outcomes and the average cell cycle time (right hand side of the equation) in the stem cell compartment. $E[W_1]$ can be enumerated as,

$$E[W_1] = p \times 2 + q \times 1 + r \times 0.$$

Assuming that all stem cells in the crypt have an equal chance of dividing, we can approximate $E[S_1]$ as,

$$E[S_1] = \frac{C}{\bar{N}_{SC}},$$

where C is the cell cycle time and $\bar{N}_{\mathrm{SC}}$ is the average number of stem cells.

Use of the above equation makes intuitive sense because if there were N stem cells with a cell cycle time of C hours, we would expect a cell to divide every C/N hours on an average. Thus, we have

$$R_{\mathrm{SC}}^{\mathrm{out}} = \frac{\bar{N}_{\mathrm{SC}}(2p+q)}{C}$$

as the equation for the rate of exit of stem cells from the compartment.

2.2 The Transit-Amplifying Compartment

The cells that comprise the transit-amplifying compartment proliferate prodigiously to make up for the high rate of loss of cells from the terminally differentiated compartment and the very low proliferation rate of the cells in the stem cells compartment. The stem cells exiting from the stem cell compartment enter the transit-amplifying compartment, go through cycles of mitosis, and then leave the compartment. Given the enter, proliferate, and exit nature of this compartment, it can be analyzed purely as a rate amplifier. Assume an epoch of time C_{TA} equal to the average cell cycle time of the transit-amplifying cells. The number of cells that enter during C_{TA} is determined by the rate of exit of stem cells $R_{\mathrm{SC}}^{\mathrm{out}}$ such that $K = R_{\mathrm{SC}}^{\mathrm{out}} \times C_{\mathrm{TA}}$ is the number of cells that enter the transit-amplifying compartment during a time C_{TA}. Every C_{TA} that elapses the K cells double and another K cells enter the compartment. If the first set of K cells that enter stay for $L = l \times C_{\mathrm{TA}}$ hours where l is any integer, then rate of exit of cells from the compartment can be easily given as

$$R_{\mathrm{TA}}^{\mathrm{out}} = \mathbf{2}^l \times R_{\mathrm{SC}}^{\mathrm{out}}.$$

The number of cells found in the TA compartment can also be easily expressed as,

$$N_{\mathrm{TA}} = K \times (2^0 + 2^1 + 2^2 + \cdots + 2^l) = K \times (2^{l+1} - 1) \text{ cells.}$$

3 Prediction of Stem Cell Division Time in Homeostasis

A typical human colonic crypt contains on an average about 82 levels of cells from the bottom to the top of the crypt. The bottom most level contains one cell and levels 2 to 7 contain 6, 12, 18, 24, 30, and 36 cells, respectively. The straight portion of the crypt from levels 8 to 82 contain 42 cells per each level adding to about 3,193 cells per crypt [11]. From stem cell markers staining data, Musashi-1 and Lgr5, it is evident that the stem cells predominantly populate the bottom 3 to 4 levels of the crypt. There are about 19 cells in the bottom 3 levels while there are 37 cells in the bottom 4 levels. For the purposes of an illustrative example, we can take the average of these two numbers, 30 as the number of stem cells in a crypt. Data from S-phase labeling and Ki-56 labeling indicate that levels 4 to 45 which have more than or about the average level of S-phase staining can be considered to be the transit-amplifying compartment comprising of rapidly proliferating cells. From the distribution of the number of cells/level, levels 4 to 45 contain 1,686 cells. The remainder $3{,}193 - (1{,}686 + 30) = 1{,}477$ is the number of cells in the terminally differentiated compartment/crypt. $N_{\mathrm{SC}} = 30$, $N_{\mathrm{TA}} = 1{,}686$, and $N_{\mathrm{TD}} = 1{,}477$, respectively.

The crypt turnover rate is about 5 days and it is expected that 95% of the cells in a crypt are turned over in a 5 day period, thus the rate of loss of cells from the crypt or the terminally differentiated compartment is $R_{\mathrm{Out}} = \frac{0.95\times 3193}{120} = 25.27$ cells per hour. Since there is almost no proliferation in the terminally differentiated compartment, the average rate of cells exiting out of the transit-amplifying compartment is the same, $R_{\mathrm{TA}}^{\mathrm{out}} = R_{\mathrm{Out}}$. From the previous section, we have

$$R_{\mathrm{TA}}^{\mathrm{out}} = \mathbf{2}^{l} \times R_{\mathrm{SC}}^{\mathrm{out}},$$

and so we have $R_{\mathrm{SC}}^{\mathrm{out}} = \mathbf{2}^{-1} \times R_{\mathrm{TA}}^{\mathrm{out}} = \mathbf{2}^{-1} \times 25.27$ cells per hour.

Previous work in estimating critical parameters of the crypt dynamics have resulted in many estimates of parameters. Here we use some of the parameters available in the literature and test the predictions of the model and hence its validity. The number of stem cells in a crypt is put at about 30 per crypt, so $\bar{N}_{\mathrm{SC}} = 30$ cells per crypt. Given the very unlikely chance of symmetric division and the more common occurrence of an asymmetric division with the stem cells, we can use $p = 0.05$, $q = 0.9$, and $r = 0.05$. Since the stem cells proliferate a lot slower than the cells of the transit-amplifying compartment, we test a prediction for the stem cell cycling time $C = 90$ h. With these values

$$R_{\mathrm{SC}}^{\mathrm{out}} = \frac{\bar{N}_{\mathrm{SC}} \times (2p + q)}{C} = \frac{30 \times (0.1 + 0.9)}{9.0} = 0.333 \text{ cells per hour}$$

$$R_{\mathrm{TA}}^{\mathrm{out}} = 2^{l} \times R_{\mathrm{SC}}^{\mathrm{out}}$$

$$l = \log_2\left(\frac{25.27}{0.33}\right) = 6.26.$$

Thus a value of $l = 6.26$ indicates that cells in the transit-amplifying compartment cycle about 6.26 times or undergo about 6.26 successive divisions before they become terminally differentiated. Also from literature, we have the cell cycle times in the transit-amplifying compartment as $C_{\mathrm{TA}} = 29.9$ to 39.9 h. Thus N_{TA} is bound by,

$$N_{\mathrm{TA}} = \frac{K \times (1 - 2^{l+1})}{(1-2)} = \frac{(35 \pm 4.9) \times 0.33 \times (1 - 2^{7.29})}{-1} = 1800 \pm 266 \text{ cells},$$

where $C_{\mathrm{TA}} = 35$ h is the mean value. This average value of $N_{\mathrm{TA}} = 1{,}800$ cells approximately matches 1,686, the number of cells in the transit-amplifying compartment obtained by s-phase labeling of the crypt.

4 Discussion

We have presented here a stochastic random process based model for the homeostasis and organization of the colonic crypt. The pool of stem cells is modeled as a renewal process with every cell division creating the event and the subsequent decision to either remain a stem cell or differentiate creating the counting process. The differentiation and exit of cells from the stem cell pool is modeled as a renewal-reward process and by using the fundamental theorem of the reward process we have been able to quantify the rate of exit of cells from the stem cell pool.

The stem cells have been the most elusive and secretive of all the cells in the colonic epithelium and it is only in the last few years that a definitive marker, Lgr5, has been found that can reliably mark the stem cell pool. Much less is known either about the nature of the symmetric or asymmetric division in the stem cell pool and the probabilities with which they occur. The very infrequent if not dormant cell cycling times of the stem cells have also been a question that has not been resolved adequately due to the extreme difficulty in marking these cells for proliferation markers. Using the equation for the rate of differentiation of the stem cells and numbers obtained from the general knowledge of the crypt dynamics, we tested a value for the cell cycling time of the stem cells. Using this value, we were able to verify, within bounds, the approximate number of cells expected to be found in the transit-amplifying compartment and thus validating the value for the stem cells cycling time.

The advantage of using the renewal and the renewal-reward process is that some of the simple yet powerful theorems and results derived for these stochastic processes can be used to describe the processes and dynamics of the crypt evolution and organization. The stem cell pool and the rate of differentiation of the stem cells determine the existence of the crypt and its size respectively. Apart from the mean rate, the variance of the rate of differentiation of stem cells, which can also be analytically expressed, gives a wealth of information regarding the stochasticity in crypt length and numbers observed. Renewal-reward theory enables

us to analytically express the variance as well and this in turn can lead to the development of other aspects of crypt dynamics, namely feedback regulation, that is not explicitly discussed here.

It can also be easily seen that this model can, in principle, be used to analyze the effects of mutations in the stem cell compartment that can lead to colon associated cancers. For example, the APC mutation leads to an increase in the probability of symmetric cell division where both the daughter cells remain stem cells thereby increasing the number of stem cells. This can then easily lead to a decrease in the differentiation rate and thus to a reduced pool of transit-amplifying cells. Thus the corresponding stochastic models not only enable a more faithful representation of the inherent randomness in crypt biology but allows derivation of expected or average results that are observed in experiments and provide a handle to discuss dynamical variability that might lead to cancer and other abnormalities.

References

1. Frank SA (2007) Dynamics of cancer: incidence, inheritance, and evolution. Princeton University Press, Princeton
2. Bach SP, Renehan AG, Potten CS (2000) Stem cells: the intestinal stem cell as a paradigm. Carcinogenesis 21:469–476
3. Ghazizadeh S, Taichman LB (2001) Multiple classes of stem cells in cutaneous epithelium: a lineage analysis of adult mouse skin. EMBO J 20:1215–1222. doi:10.1093/emboj/20.6.1215
4. Potten CS, Booth C (2002) Keratinocyte stem cells: a commentary. J Invest Dermatol 119:888–899. doi:0020 [pii] 10.1046/j.1523–1747.2002.00020.x
5. Watt FM, Hogan BL (2000) Out of Eden: stem cells and their niches. Science 287:1427–1430. doi:8287 [pii]
6. Morrison SJ, Kimble J (2006) Asymmetric and symmetric stem-cell divisions in development and cancer. Nature 441:1068–1074. doi:nature04956 [pii] 10.1038/nature04956
7. Simons BD, Clevers H (2011) Strategies for homeostatic stem cell self-renewal in adult tissues. Cell 145:851–862. doi:S0092–8674(11)00594–0 [pii] 10.1016/j.cell.2011.05.033
8. Bjerknes M, Cheng H (2002) Multipotential stem cells in adult mouse gastric epithelium. Am J Physiol Gastrointest Liver Physiol 283:G767–777. doi:10.1152/ajpgi.00415.2001
9. Barker N et al (2007) Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449:1003–1007. doi:nature06196 [pii] 10.1038/nature06196
10. Wang Z, Zhang J (2011) Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci USA 108:E67–76. doi:1100059108 [pii] 10.1073/pnas.1100059108
11. Boman BM, Fields JZ, Cavanaugh KL, Guetter A, Runquist OA (2008) How dysregulated colonic crypt dynamics cause stem cell overpopulation and initiate colon cancer. Cancer Res 68:3304–3313. doi:68/9/3304 [pii] 10.1158/0008–5472.CAN-07–2061
12. Gerike TG, Paulus U, Potten CS, Loeffler M (1998) A dynamic model of proliferation and differentiation in the intestinal crypt based on a hypothetical intraepithelial growth factor. Cell Prolif 31:93–110
13. Johnston MD, Edwards CM, Bodmer WF, Maini PK, Chapman SJ (2007) Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. Proc Natl Acad Sci USA 104:4008–4013. doi:0611179104 [pii] 10.1073/pnas.0611179104

14. Buske P et al (2011) A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt. PLoS Comput Biol 7:e1001045. doi:10.1371/journal.pcbi.1001045
15. Cox D (1970) Renewal theory. London: Methuen & Co. ISBN 041220570X
16. Boman BM, Wicha MS, Fields JZ, Runquist OA (2007) Symmetric division of cancer stem cells – a key mechanism in tumor growth that should be targeted in future therapeutic approaches. Clin Pharmacol Ther 81:893–898. doi:6100202 [pii] 10.1038/sj.clpt.6100202

Chapter 12
A Dynamic Physical Model of Cell Migration, Differentiation and Apoptosis in *Caenorhabditis elegans*

Antje Beyer, Ralf Eberhard, Nir Piterman, Michael O. Hengartner, Alex Hajnal, and Jasmin Fisher

Abstract The germ line of the nematode *C. elegans* provides a paradigm to study essential developmental concepts like stem cell differentiation and apoptosis. Here, we have created a computational model encompassing these developmental landmarks and the resulting movement of germ cells along the gonadal tube. We have used a technique based on molecular dynamics (MD) to model the physical movement of cells solely based on the force that arises from dividing cells. This novel way of using MD to drive the model enables calibration of simulation and experimental time. Based on this calibration, the analysis of our model shows that it is in accordance with experimental observations. In addition, the model provides insights into kinetics of molecular pathways within individual cells as well as into physical aspects like the cell density along the germ line and in local neighbourhoods of individual germ cells. In the future, the presented model can be used to test hypotheses

A. Beyer
Department of Genetics, University of Cambridge, Cambridge, UK
e-mail: ab704@cam.ac.uk

R. Eberhard
Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

PhD Program in Molecular Life Sciences, Life Science Zurich Graduate School and MD/PhD Program, University of Zurich, Zurich, Switzerland
e-mail: ralf.eberhard@imls.uzh.ch

N. Piterman
Department of Computer Science, University of Leicester, Leicester, UK
e-mail: nir.piterman@leicester.ac.uk

M.O. Hengartner • A. Hajnal
Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
e-mail: michael.hengartner@mnf.uzh.ch; alex.hajnal@imls.uzh.ch

J. Fisher (✉)
Microsoft Research, Cambridge, UK
e-mail: Jasmin.Fisher@microsoft.com

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_12, © Springer Science+Business Media, LLC 2012

about diverse aspects of development like stem cell division or programmed cell death. An iterative process of evolving this model and experimental testing in the model system *C. elegans* will provide new insights into key developmental aspects.

1 Introduction

Since the early 1970s [1], the nematode *C. elegans* has been a widely studied model in biomedical research (reviewed in e.g. [2–5]). Through the worm's transparent body it is possible to trace any cell by light microscopy or to study gene expression and cellular development *in situ* [6]. The fixed number of cells of the somatic cell lineages have been meticulously described (cf. [7]) and are invaluable for the genetic analysis of regulatory pathways in development (cf. [8]) or in neurobiology. The germ line of *C. elegans* allows for the observation of several essential developmental processes like stem cell proliferation, gametogenesis and programmed cell death, also termed apoptosis. Importantly, these biological processes are spatially well-resolved in this system, where germ cells mature in sequential steps along a tube-shaped gonad. It has, therefore, been extensively used in basic research (reviewed in [9–12]). Other than the highly predictable development of somatic tissues, cellular events in the germ line seem to be very stochastic; consequently the underlying general mechanisms are little understood for some of these processes. This is particularly true for physiological germ cell apoptosis. Programmed cell death is a crucial developmental process that is found in many different species; aberrations in this program have important implications in complex diseases like human cancers [13] or neurodegenerative disorders [14]. It is, therefore, key to gain fundamental understanding of its mechanisms. In this work, we propose a computational model of the germ line that is mainly based on physical properties and which aims to provide more insights into the previously mentioned developmental processes. With our model, we are able to test hypotheses about the causes and mechanisms of programmed cell death, among other developmental processes, and to highlight promising theories to be validated experimentally.

1.1 The C. elegans Germ Line

The reproductive system of *C. elegans* has a symmetric structure with two U-shaped gonads extending from a single vulva, one anteriorly and one posteriorly. Our model considers the development from stem cells to mature oocytes within one gonad (see Fig. 12.1). Although the nuclei and their cytoplasm within the germ line are not completely encapsulated cells and thus are part of a syncytium, they are usually referred to as germ "cells". As the differential interference contrast (DIC) picture and the electron microscopy imaging in Fig. 12.1 indicate, the cytoplasmic membranes are not fully delimiting, leaving a connection of all cells to a common shared cytoplasm in the centre of the gonad tube, called rachis.

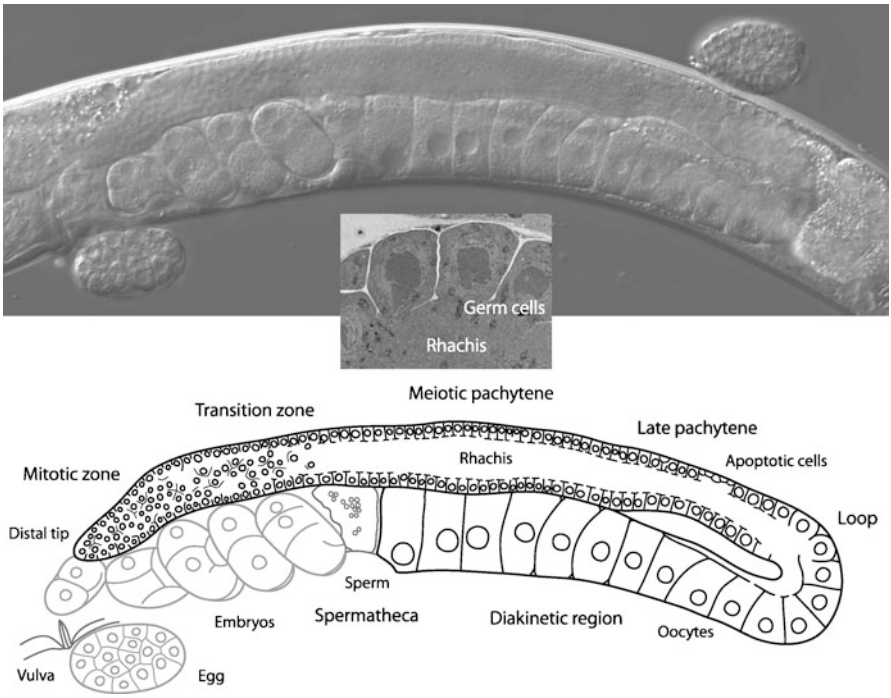


Fig. 12.1 The germ line of *Caenorhabditis elegans* by DIC (*top*), as a schematic (*bottom*), and in a cross-section (transmission electron microscopy, *middle*). The head of the worm is to the *right*, the posterior gonad to the *left* of the picture. Differential interference contrast (DIC) microscopy allows us to observe live animals in any focal plane; here, an adult hermaphrodite is virtually dissected along a plane through the centre of the gonad tube. The germ cells in the meiotic pachytene region form a monolayer around a concentric inner tube, seen as a nuclei-free area in the longitudinal and cross-sections (rachis). The limits of the transition zone and of the late pachytene stage within the meiotic pachytene region are not strictly defined by DIC. The oocytes in the loop have exited pachytene and begin the diakinetic stage of meiosis

The mature hermaphroditic germ line can be divided into functionally different zones with specific developmental properties [15–18]. At the distal most end of the gonadal tube, the mitotic zone is located ("distal" here meaning farthest from the uterus), containing dividing stem cells and representing a stem cell niche. The potential of the mitotic cell pool to divide is maintained by molecular signals – directly via activation of proliferation or, more likely, indirectly via inhibition of differentiation. Delta ligand from extrinsic sources (the distal tip cell) activates the Notch pathway, promoting a high Notch within the germ cells of this region. In the transition zone, where no external Delta ligand is presented, the Notch level gradually decays. When the germ cells are left without Notch, they complete the mitotic cell cycle, enter meiosis and start their differentiation into oocytes [16]. A small transition zone in which mitotic and first meiotic cells are interspersed links to a seemingly well-orchestrated meiotic pachytene region, where chromosomes

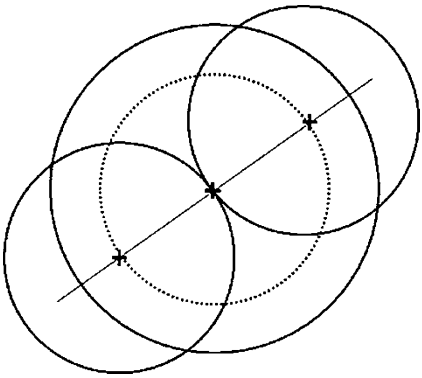


Fig. 12.2 Schematic representation of a dividing cell. The *big circle* represents the parent cell, the two *diametrically positioned circles* are the daughter cells, the *centres* of the three cells are represented by the + signs. The *straight line* signifies the division axis and the *dotted circle* is the imaginary circle with one basic radius around the centre of the parent cell on which the centres of the daughter cells are placed

undergo homologous recombination. At some point within the meiotic zone, the germ cells start growing at a low rate so that they have visibly increased their size by the time they reach the bend of the gonad and exit the pachytene stage of meiosis. In this loop region, the rachis is thinned to an eccentric tube, but still connecting the growing oocytes before they become proper cells with a fully closed membrane. Distal to the loop with the young oocytes, programmed cell death can be observed as part of normal oogenesis [19]. Physiological apoptosis, the fate of about half of all germ cells, is considered to be restricted to this area of the gonad [11]. Ras/MAPK activity is required for pachytene exit [20] and oocyte maturation; its absence also disables apoptosis [19]. For our model, we premise that germ cells start accumulating Ras activity towards the end of the meiotic pachytene region, induced by an external Ras signal. If the Ras level surpasses a certain threshold in a germ cell, it starts to grow to become a fully grown oocyte filling the complete diameter of the tube when it reaches the proximal end of the gonad. We also assume here that the Ras level is decisive for germ cell death: it renders a cell capable for or insensitive to physiological apoptosis.

1.2 Molecular Dynamics Model

Dividing cells in the mitotic zone apply pressure on the surrounding cells as the two daughter cells need more space than the parent cell (cf. Fig. 12.2). This leads to physical movement of the cells away from the pressure centre. We have constructed our model using an algorithm based on the molecular dynamics (MD) modelling framework [21] to capture this movement according to the physical properties of

each individual cell. This makes the movement of cells in our model very realistic so that we get a "virtual germ line". Apart from realistic movement, the MD framework allows for good visualisation and tracing of parameters for single or multiple cells. Originating from theoretical physics and chemistry to investigate the behaviour and properties of various particles like planets or molecules, physical algorithms similar to MD have also been used in a few biological settings [22–27]. In contrast to our specific modelling system, these MD models were applied to simulate the collective behaviour of tissues and aggregations of cells moving along a chemical or nutritional gradient. In these cases, the physical movement through MD is just a side effect of the main movement along the gradient, while in our case the MD-movement is the main component of movement. In fact, it is the only driving source of movement; without the forces derived within the MD approach, the cells in our model would not move at all.

In addition to movement, we have built our MD-model to include developmental processes such as cell growth and division, as well as apoptosis, to make it sufficiently realistic. These processes depend on signals received by the cells according to their location in the tube. We show that our model reproduces cellular behaviour observed in experimental settings [15, 28, 29] very closely. This suggests that our model is a useful tool to gain novel insights into the core developmental processes observed in the *C. elegans* germ line.

2 Model

The main part of our model is the movement algorithm from molecular dynamics. We have used the *velocity Verlet* algorithm [30], which is based on the following basic formula of Newtonian motion:

$$F = ma. \tag{12.1}$$

A Taylor series development of formula (12.1) and some further transformations imply the following steps of the algorithm for each cell. We will elaborate them a little more in the subsequent two paragraphs.

Step 1: $\vec{v}\left(t + \frac{1}{2}\Delta t\right) = \vec{v}(t) + \frac{1}{2}\vec{a}(t)\Delta t$.
Step 2: $\vec{x}\,(t + \Delta t) = \vec{x}(t) + \vec{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t$.
Step 3: Derive $\vec{a}(t + \Delta t)$ from the interaction potential using $\vec{x}(t + \Delta t)$.
Step 4: $\vec{v}(t + \Delta t) = \vec{v}\left(t + \frac{1}{2}\Delta t\right) + \frac{1}{2}\vec{a}(t + \Delta t)\Delta t$.

Here, t represents the time and Δt signifies the timestep of the execution which is usually very small. The variable $\vec{x}(t)$ represents the location at the current time, $\vec{v}(t)$ stands for the velocity at the current time and $\vec{a}(t)$ for the acceleration at the current time. In our model, we replace acceleration with the force and the mass based on formula (12.1), i.e. $a = F/m$. For now, we simplify this further by assuming the mass to be one, which leads to $a = F$.

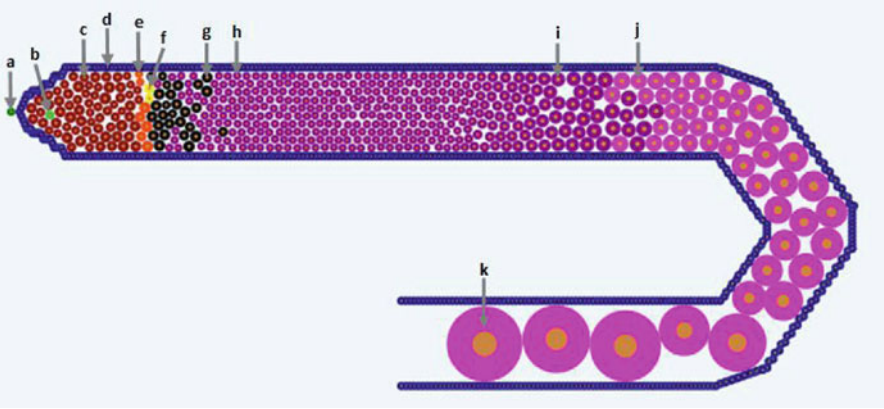


Fig. 12.3 Snapshot of an execution of the germ line model: (*a*) distal tip cell (*dark green*), (*b*) marked cell (*light green*), (*c*) mitotic cell with highest Notch level (*dark red*), (*d*) cells defining the border of the tube (*blue*), (*e*) mitotic cell with Notch level between highest and 0.5 times the highest level (*orange*), (*f*) mitotic cell with Notch level between 0.5 times the highest level and 0 (*yellow*), (*g*) mitotic cell with Notch level equal to 0 (*black*), (*h*) meiotic cell (*purple*), (*i*) meiotic cell that has grown to about twice its original size (*purple*), (*j*) oocyte with Ras level above threshold (pink) and (*k*) fully grown oocyte (*pink*)

The first step of the algorithm computes the velocity at the intermediate time $t + 1/2\Delta t$, which is used in the next step to evaluate the location of the cell at the next time, $t + \Delta t$. Using this new location of the cell, the force working on it at time $t + \Delta t$ is derived from the so-called interaction potential in the third step. The interaction potential is a function describing how a particle, here a cell, interacts with its environment. In our case, the potential acts in such a way that the cell is pushed away from cells that overlap it and there is no force working on the cell if other cells do not overlap it. The force on the cell is stronger the more overlap there is. If more than one cell overlaps the cell, the forces are accumulated. In the last step of the algorithm, this new force is used to calculate the velocity at the next time $t + \Delta t$. This algorithm is computed for each cell and each timestep Δt of the execution for a certain number of timesteps.

Since the algorithm works on a per cell basis, we have constructed the cells in our model as objects. To unite the functional style of the algorithm and the object oriented style of the cells, we have used the F# programming language [31], which incorporates both of these environments in a natural way. Additionally, the language also provides us with a very straightforward visual front end. For simplification, our current model is a two-dimensional representation of the germ line as shown in Fig. 12.3. A movie of an execution of the model can be viewed at [32]. The different colours are used to represent different states or component levels of the cells. To define the general structure of the germ line in our model, we estimated germ cell numbers along the gonadal tube from microscopic pictures of the germ line. This

provided us with estimates of cells across the tube diameter (cells per column in our model) and per developmental zone along the tube. The sizes of the different zones of the germ line were translated into the range of pathway activation in our model.

2.1 Internal Properties of Cells

As previously mentioned, the cells in our model are represented as objects. All cells are described by the same object class; they all have the same general internal properties. The different values of these properties describe the cells' current status within the model. Differences in cell behaviour between cells arise from differences in the specific momentary environment and slight randomisations of certain internal properties. In this section, we will describe the most important of these internal properties and how their values are derived.

2.1.1 Location

Every cell has a specific two-dimensional location assigned to it. The location is updated at each timestep of the execution depending on the velocity Verlet algorithm. The location is in continuous space as opposed to models which work with a grid of possible locations.

2.1.2 Velocity and Force

As previously mentioned, the velocity Verlet algorithm computes the new cell locations based on the velocity and force that are acting on the cell. For this purpose, we have equipped each cell with two-dimensional velocity and force vectors which are changed by the algorithm. The values of these two properties define the degree of change in the location through the velocity Verlet algorithm.

2.1.3 Cell State

Another important property of the cells in our model is the state of a cell. We have defined the states "Mitotic" (c, e, f and g in Fig. 12.3) and "Meiotic" (h, i, j and k in Fig. 12.3) which depend on the location and the cell cycle state of the cells. We also mark dead and fertilised cells in this way to make them countable and to be able to remove them from the model. Furthermore, we defined a state "Stopped" (a and d in Fig. 12.3) that marks all boundary cells of the tube not to move and to form the walls of the germ line. This definition of the germ line walls does not exactly conform to nature, but it is a simplification for computation purposes that is sufficiently realistic.

2.1.4 Ras and Notch Level

The cells also contain variables defining their Ras and Notch levels. These are changed according to the environment at a cell's location along the tube as this informs about the presence of external ligand molecules for the two signalling pathways. Along the x-axis of the coordinate system of our model, the Delta ligand is active in the region from 3.8 (which is the beginning of the tube) to 22. From coordinate 22, the Delta ligand is turned off; this is where the transition zone starts. The external Ras is present from 75 up to the beginning of the bend at 117 when it is turned off. When Delta ligand is present, the Notch level within the cells will jump to its highest value (c in Fig. 12.3). The level decays linearly over time in the absence of ligand (from c to e to f to g, representing no Notch, in Fig. 12.3). The Ras level is accumulated linearly over time as long as the ligand is present.

2.1.5 Size and Growth

The cells are also assigned a size that is updated at each step of the execution. For simplicity, we consider our cells to be round. Hence, the size of a cell is its radius. The size of a cell changes as it progresses along the gonadal tube depending on the cell's status and location. All cells have the same basic size to which they go back after a division. The change of size is defined by growth functions common to all cells. An exception to this is the function defining the growth rate of mitotic cells that is randomised for each cell. This assures that cells which are born at the same time do not necessarily divide synchronously, but at slightly different times. This results in a more realistic timing of the cell divisions.

2.1.6 Analysis Parameters

The previously mentioned cell properties are all included to achieve a behaviour and movement of the cells that is as realistic as possible. We have also included a few parameters for purely analytical purposes.

GFP

All cells contain a variable GFP – named after the visual marker green fluorescent protein in biological experiments – that can be turned on to visually follow this cell and its offspring in the simulation (b in Fig. 12.3). In addition, the data for these cells, i.e. the values of the previously mentioned parameters and the analytical parameters described below, can be read out to be analysed. If desired, this could be adjusted to track only one cell.

Density Factor

The cells also have a density factor associated with them. This density factor f of cell c is computed at each timestep using the following equation:

$$f(c) = \sum_{c_i \in N(c)} \frac{r_i}{6\,(d\,(c_i, c) - r_i)},$$

where $N(c)$ is the set of all cells touching cell c, r_i is the radius of cell c_i and $d(c_i, c)$ is the distance between the cells c_i and c. This factor is set to zero if no cell is touching c and to one if the neighbouring cells are ideally packed, i.e. c is surrounded by six cells of the same radius as c. The density is above one if other cells are overlapping c. This density factor could later be used to test hypotheses about the apoptotic mechanism.

Movement and Division Rates

To compare the movement rate of the cells with experimental findings, we have also introduced a list to represent this rate. A function writes each timestep to this list in which the cell has moved at least one diameter along the x-axis compared to the location at the previously stored timestep. Similarly to the movement rate, we have also defined a list representing the division rate. This list is appended by the current timestep when the cell divides so that each cell carries a history of all previous divisions for its ancestors. This list is passed to both daughter cells upon a division.

2.2 Other Properties of the Model

The cells in our model are not stand alone objects, but they interact with each other and their environment, i.e. the gonadal tube. The interplay between the internal configuration of the cells and their surroundings results in the behaviour that can be observed in the model. For the specification of these interactions, we have defined some general properties of the model which are described in this section.

2.2.1 Cell Growth and Division

Before they divide to form two daughter cells, cells marked as mitotic grow to a size of $\sqrt{2}$ times their basic radius, which corresponds to twice their area. The two daughter cells both have the same basic radius and are placed so that the distance of their centres is two basic radii (cf. Fig. 12.2). In the *C. elegans* germ cells, the orientation of the division axis appears to be random. In our model, the daughter cells are placed symmetrically with regards to a random central division axis in

their parental cell (straight line in Fig. 12.2). Their centres will thus be positioned diametrically on an imaginary circle around the centre of the parent cell with the basic radius (dotted circle in Fig. 12.2). The daughter cells inherit the Notch level and GFP status from their parent. At the same time, the division rate list in the daughter cells is updated and a cell is added to a counter of the total number of cells.

2.2.2 Growth of Early Meiotic Cells and Oocytes

Meiotic cells stay at the basic radius until they reach a certain point in the tube from which they start growing slowly, i.e. their radius is increased, until they have reached about double their area right before the loop. When the Ras level within these cells has reached a certain threshold, they exit the pachytene stage of meiosis and proceed into diakinesis to become mature oocytes (j and k in Fig. 12.3). Initially, the radii of the oocytes grow about ten times faster than the ones of late pachytene cells and, as they reach the end of the loop, this growth rate is increased by another 10-fold. The oocytes stop growing when their diameter is the width of the tube (k in Fig. 12.3). All cell growth in our model is defined by an increase in radius and happens at linear rates.

2.2.3 Death

In vivo, programmed cell death is normally confined to the late meiotic pachytene region before the loop where some cells become apoptotic. Accordingly, we defined a death zone in our model. In the worm, the death zone seems to be dependent on the location of the oocytes. Hence, we have defined a death zone of fixed size which ends at the x-axis location of the eighth oocyte and begins at a fixed distance distal to it. At the moment, random cells within this death zone will be eliminated as soon as the amount of cells surpasses a certain number. As a constraint to random cell death, we defined an artificial Ras threshold, above which cells become insensitive to apoptosis; only cells below this level are selectable for cell death. This is a somewhat naïve approach to the induction of apoptosis; the present model is just preliminary in this respect. Once a cell has died, the timestep number of its death is recorded for analysis purposes and the cell disappears from the model.

2.2.4 Fertilisation

For simplicity, we currently define a cell as being fertilised when it reaches the end of the tube. As with the dead cells, the timestep numbers when fertilisations happen are recorded and the cells are removed from the model.

3 Results

3.1 Calibration

To calibrate the model, we extracted parameter values (shown in the first column of Table 12.1) from the literature and from our own videos of the germ line to compare with our model. Our observations have shown that, on average, it takes the cells 90 minutes to advance along the x-axis by the distance of their own diameter. Literature [28, 29] suggests that one mitotic cell cycle in the germ line is 16 to 24 hours. This calculates to an average time of 20 hours between divisions. For the death rate, our observations suggest that on average two cells die per hour per gonad arm, a value observed independently by many groups since the first characterisation of physiological germ cell apoptosis in *C. elegans* [19]. The literature and our own observations further suggest that the average egg laying rate is about four eggs per hour per animal [15]. We have to be careful to translate this for our model since we only consider one gonad, whereas the egg laying rate accounts collectively for both the anterior and posterior gonad. Consequently, we consider a fertilisation rate of two oocytes per hour per gonad arm as an approximation of the average over time.

As we will delineate in the following sections, our model produces values which are in correspondence with these expected experimental values.

3.1.1 Movement Rate

For the calculation of the movement rate, each cell is equipped with a list containing every timestep at which the cell has moved a distance forward equal to its current diameter. To evaluate the time that each of these steps has taken, we calculated the difference between every two succeeding timesteps. Figure 12.4 shows a histogram of these differences for all of the cells within the model. The distribution looks basically like a Gaussian. The average movement rate in our model is 5,250 steps for one diameter. Since we have observed the movement rate to be one diameter per 90 minutes *in vivo*, we can derive here that 90 minutes is equivalent to approximately 5,250 steps in our model. As a consequence, we can assume that one hour in real time is approximately 3,500 steps in our model.

Table 12.1 Table showing different properties extracted from experiments and literature (expected) and the according values derived from the present model (model) with the translation of these values into real time in the fourth column

Rate	Expected	Model	
Movement rate	1 row/90 min	1 row/5,250 steps	1 h = 3,500 steps
Time between divisions	20 h	$40k$ to $90k$ steps	19.19 h
Death rate	2 cells per h	1.57 cells/3,500 steps	2 cells per h
Fertilisation rate	2 cells per h	4.27 cells/3,500 steps	4 cells per h

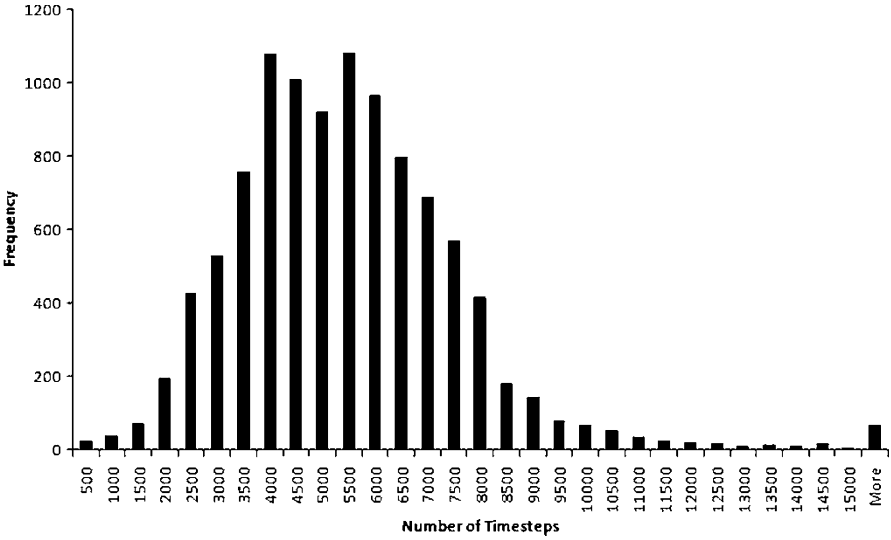


Fig. 12.4 Histogram showing a distribution of the number of timesteps needed for the advancement of each cell by one diameter

3.1.2 Time Between Divisions

Each cell carries a list containing all timesteps in which a division has occurred for this cell. By subtracting succeeding values, we have calculated the number of timesteps between two divisions. Figure 12.5 shows a histogram of these numbers for all of the cells in the model. The values for the time between divisions range from 40,000 to 90,000 timesteps. Using our estimation for real time from above, this is a range of 11 to 26 hours. The average of all times between divisions for all cells in our model is 67,158 timesteps which evaluates to about 19 hours.

3.1.3 Death Rate

For each dead cell, we have recorded the timestep number at which it died. To get an estimate of the death rate, we looked at bins of 3,500 timesteps, which is approximately one hour in real time (cf. Section 3.1.1). Figure 12.6 shows a histogram of the number of cell deaths occurring in these bins during an execution. Averaged over all bins, this computes to 1.57 deaths per bin or, in real time, 1.57 deaths per hour (continuous horizontal line). The figure shows that our reference indicated by the dashed horizontal line is only slightly higher than the average resulting in the model.

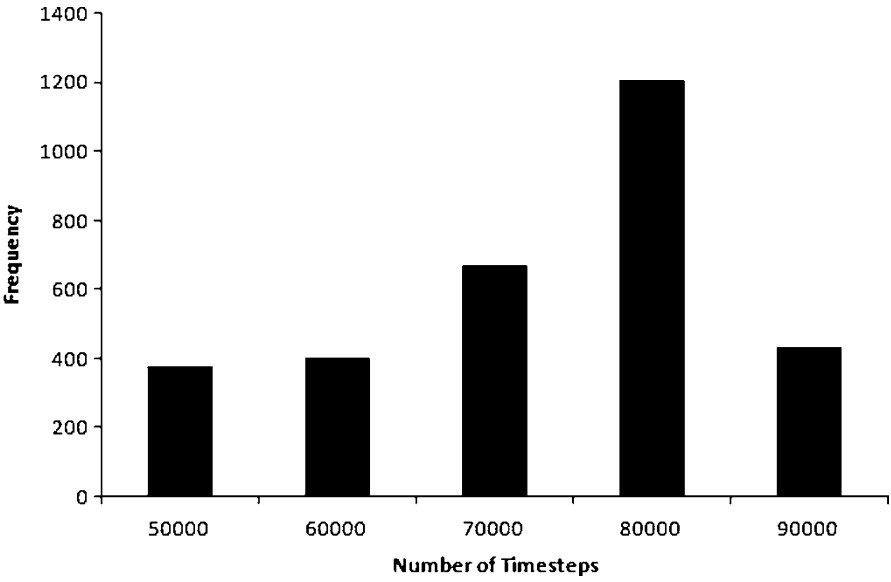


Fig. 12.5 Histogram showing the distribution of timesteps between divisions for all cells

3.1.4 Fertilisation Rate

We have also recorded the timestep at which fertilisation occurred for each fertilised cell. In a similar fashion to the death rate, we have sectioned this data in bins of 3,500 timesteps to get an estimate of the fertilisation rate. Figure 12.7 shows a histogram of the number of fertilisations per 3,500 timesteps. The figure also shows the average number of fertilisations per bin (or hour) which is 4.27 cells (continuous horizontal line). The literature reference, indicated by the dashed horizontal line in the figure, is lower than our model average.

3.1.5 Summary

Table 12.1 summarises the findings of our model calibration. Relative to the movement rate observed in our model, the time between divisions and the death rate in our model are very much in accordance with the values observed *in vivo*. Solely the fertilisation rate differs between our model and the experimental observations. These parameters are the major developmental components in our model. As a consequence, we can consider this model as, for our purposes, a very good approximation to the real organ since it reproduces the major developmental features very accurately.

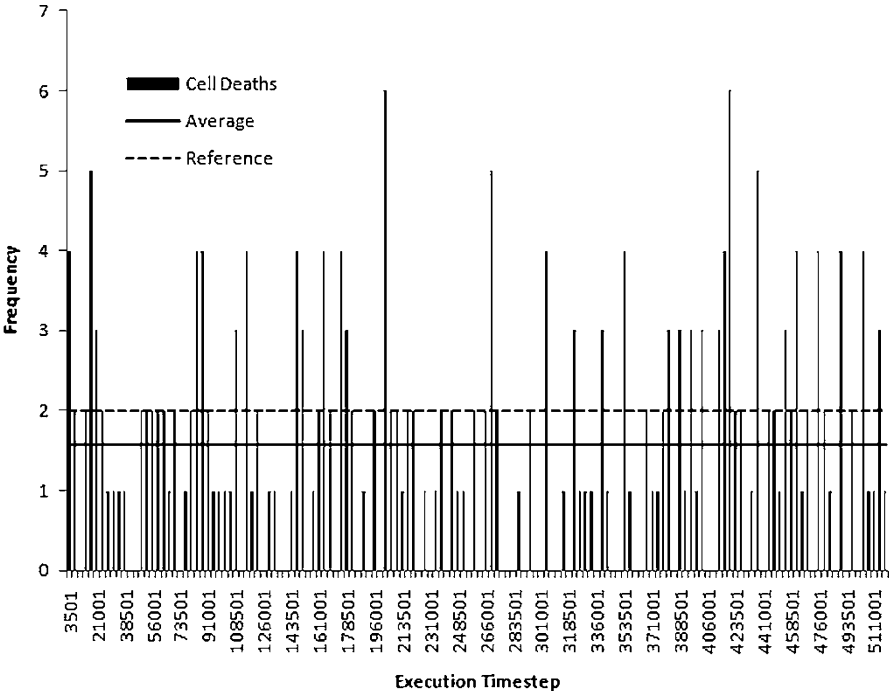


Fig. 12.6 Histogram of the number of cell deaths over bins of 3,500 execution timesteps. The *continuous horizontal line* indicates the average number of cell deaths per 3,500 timesteps, the *dashed line* shows the experimental reference

3.2 Further Results

3.2.1 Cell Numbers

Figure 12.8 shows the development of cell numbers in the different developmental stages and of the total number of cells over execution time. One can see that our model is in a steady state in terms of cell numbers and that none of these values significantly fluctuates. This is in accordance with *in vivo* observations of the germ line of an adult wild-type animal [15, 28].

3.2.2 Density

Figure 12.9 shows the average density in bins of about 1.5 basic cell radii from the distal end of the germ line up to the beginning of the loop. The average is taken of the density factors of all cells whose centre is within the respective bin. As defined earlier, a density factor of 1 represents an ideal packaging around an individual cell, while a factor of 0 indicates a cell that does not have neighbours that touch it.

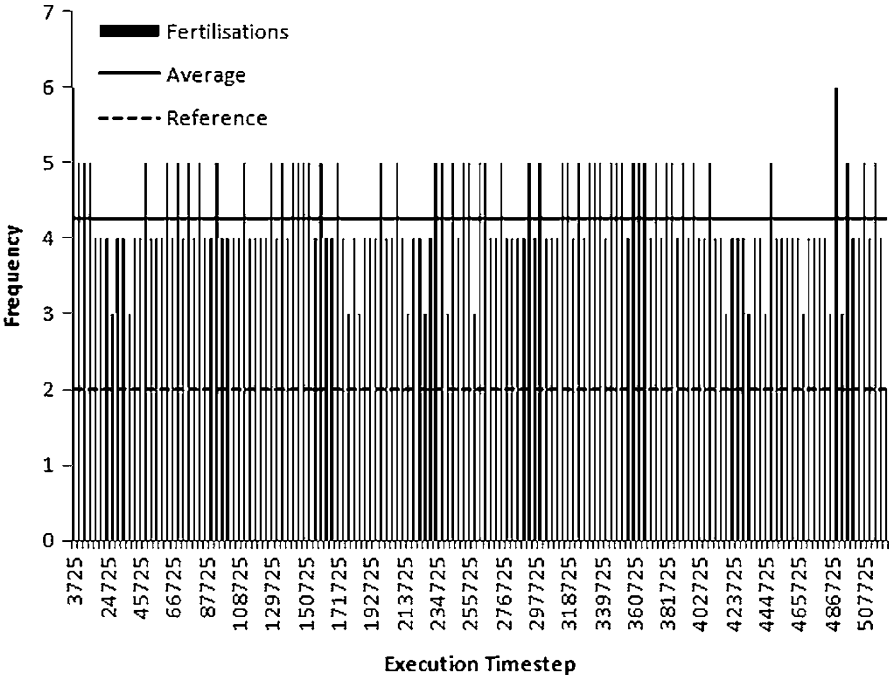


Fig. 12.7 Histogram of the number of fertilised cells over bins of 3,500 execution timesteps. The *continuous horizontal line* indicates the average number of fertilised cells per 3,500 timesteps, the *dashed line* shows the literature reference

A value above 1 hence indicates overcrowding through overlap of neighbouring cells. Figure 12.9 shows that the average density factor is relatively low. The figure also shows that the size changes of germ cells are especially important in terms of the density factor. In the region between 33 and 75, where there should be only few dividing cells – apart from a few outliers – and where the cells do not grow, the density is lowest and it does not fluctuate as much as in the other regions. The changes in density by growth are especially apparent in the region from 75 up to the end of the plot. Apart from a dip between 105 and 109, the density constantly increases, which is also true for the size of the cells in this region. The dip in this increase could be due to space that is freed up by dead cells since this is the region of the tube where the death zone is located.

3.2.3 Following Marked Cells

As mentioned in the Model section, we are able to label cells in our model and follow them on their course through the germ line. Here, we have followed one cell and its 12 offspring to get more information about what happens to individual cells in our model.

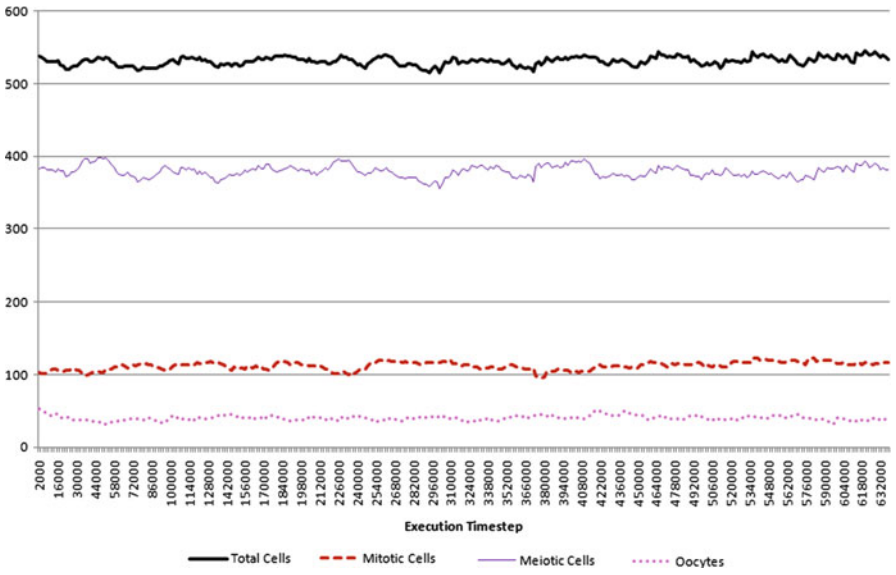


Fig. 12.8 Graph showing the development of cell numbers over execution time; the *thick continuous line* represents the total number of cells, the thin continuous line stands for the number of meiotic cells, the *dashed line* is for the mitotic cells and the *dotted line* for the oocytes

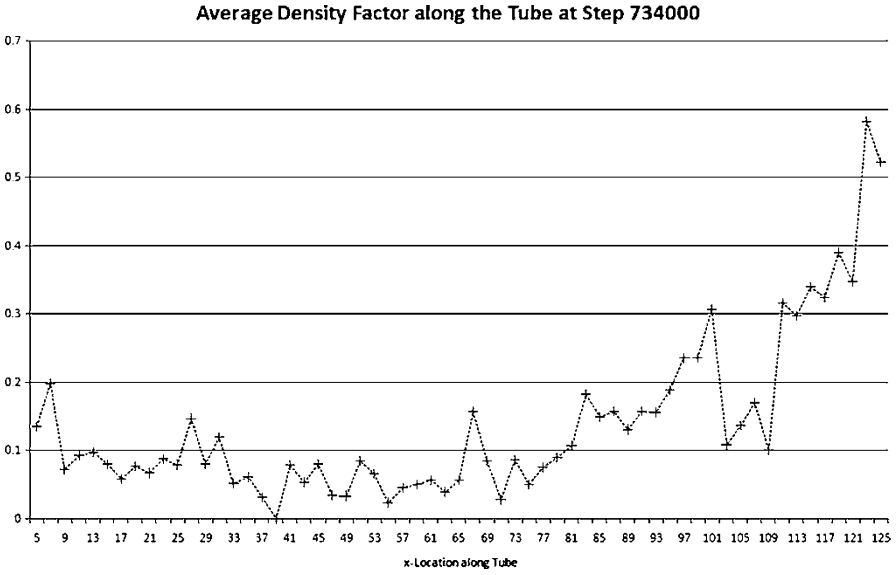


Fig. 12.9 Graph showing the average density along the germ line from the distal end up to the bend in bins of approximately 1.5 basic cell radii at execution step 734,000

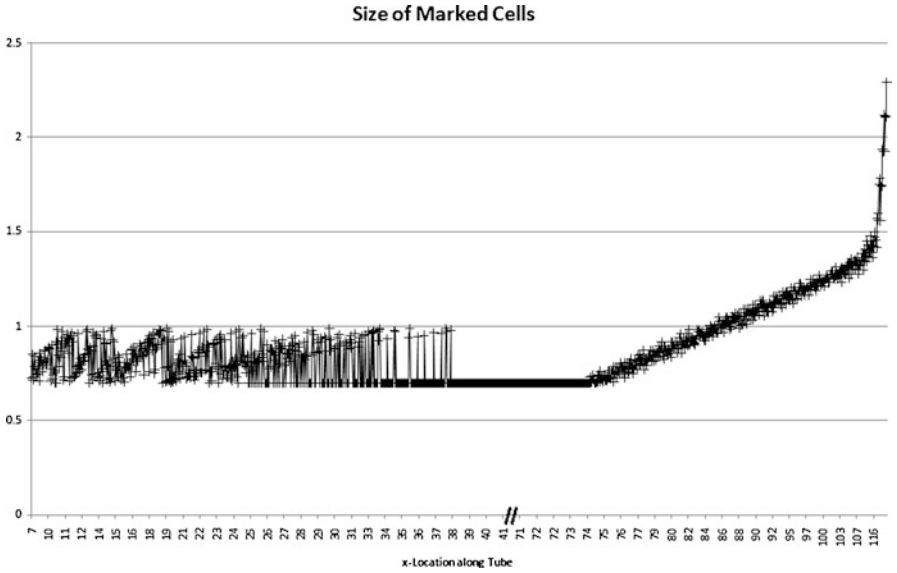


Fig. 12.10 Graph showing the sizes of the marked cells according to their location in the germ line from the distal end up to the bend during the whole execution

Figure 12.10 depicts the development of size of these individually marked cells from the distal end up to the beginning of the loop. For illustration, we have excluded values ranging between 42 and 70 on the x-axis since the values remain constant. The figure shows that the size of cells in the mitotic zone and the transition zone ranges between 0.7 and a value just below 1. This is due to the fact that the basic cell size is 0.7 and dividing cells grow from this size up to $\sqrt{2} \cdot 0.7$ which is just below 1. From about 30 up to 37 there does not seem to be a smooth change of size but rather an oscillation between maximal and basic size. This indicates that cells of big size are pushed from the mitotic zone into this area of the germ line to divide there. They do not grow anymore since they enter meiosis after the division. This area represents the transition zone. The figure also shows that between about 37 and 75 there is no growth before a constant linear growth of the cells sets in. This growth changes to a much steeper one at the end which corresponds to the region of the tube where the first cells reach Ras levels above the threshold causing them to grow faster. This figure also underlines our assumptions in the previous paragraph that growth and division of cells are very important in the development of the density factor.

Figure 12.11 shows the development of the density factor for the marked cells over the distance from the distal end of the gonad to the beginning of the loop. As in the averaged case, the density factor is generally fairly low. Since each mark in the graph stands for the density factor of one single cell, it is expected that the factor scatters more widely here. While some cells have a density factor of 0, others have

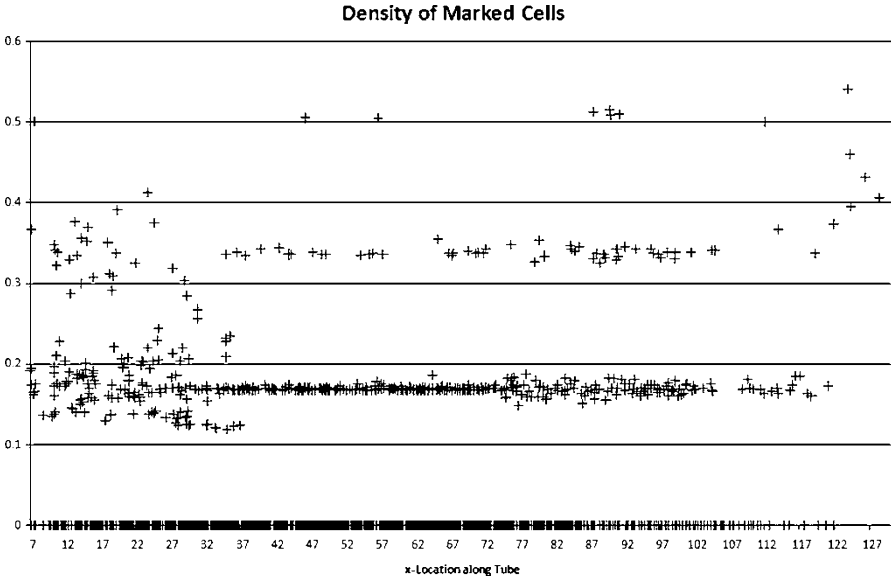


Fig. 12.11 Graph showing the density factors of the marked cells according to their location in the germ line from the distal end up to the bend during the whole execution

one of more than 0.5. While in Fig. 12.9, the average density is never above 0.2 up to about 95 on the x-axis, individual cells in Fig. 12.11 show several values above 0.2 and even some that are above 0.5 in the same region. Similar to Fig. 12.9, Fig. 12.11 also shows that in the region without cell growth and with only some division, i.e. between about 35 and 75 (cf. Fig. 12.10), the density stays relatively constant and there are more cells with very low density factors. In fact, Fig. 12.11 shows that from about 80, where growth is continuously strong, less and less of the marked cells retain a density factor of 0.

Figure 12.12 shows the dynamic behaviour of Notch and Ras within the marked cells according to their location between the distal end and the beginning of the loop in response to external signals. Similar to Fig. 12.10, we have excluded the range between 28 and 72 on the x-axis since the component levels do not change in this area. In the most distal part of the tube, the external Delta ligand of Notch signalling is present. Figure 12.12 shows that, in response to this, all cells reach the maximal Notch level, which again decays in the absence of Delta. The figure further shows that some cells reach the lowest Notch level farther down the tube than others. Since, in each cell, the Notch level linearly decays over time at the same rate, this indicates a difference in speed of the marked cells. Figure 12.12 also shows that in the presence of external Ras inducing signal, generally the Ras level within the cells constantly increases as they progress along the tube. There are a few cells before the area where the external Ras signal is present which nonetheless have an internal Ras level above 0. These are probably cells that were pushed back out of the Ras zone

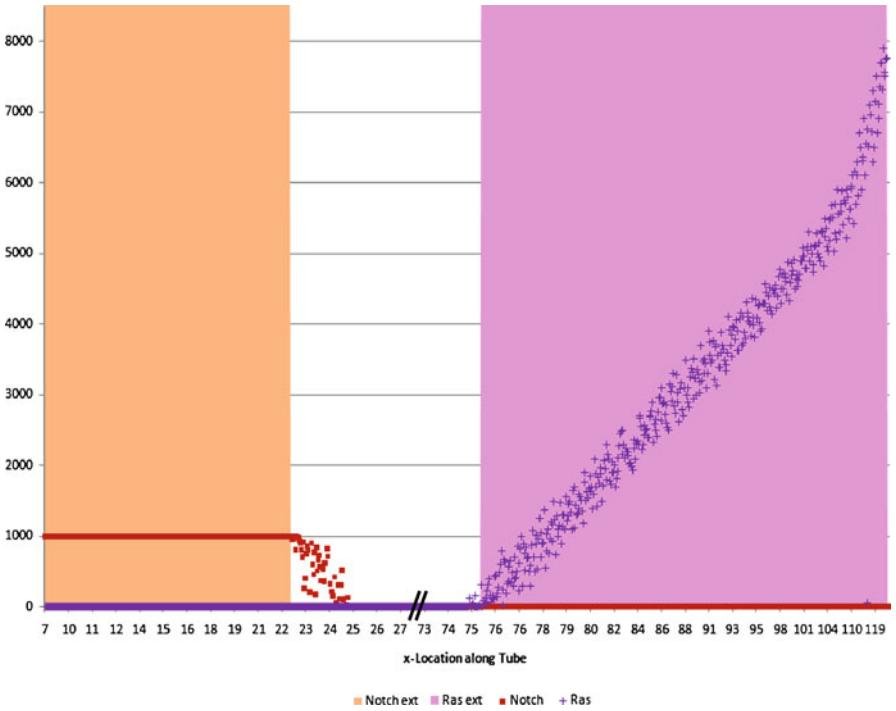


Fig. 12.12 Graph showing the Notch and Ras levels within the marked cells along with the status of external Delta and external Ras according to location in the germ line from the distal end up to the bend during the whole execution. The *shaded area* on the left indicates the region of the germ line where the Delta ligand is present (corresponding to the mitotic region), the *shaded area* on the right signifies the same for the external Ras (corresponding to the region before the gonad bend). The *squares* represent the cell internal Notch levels, the *crosses* represent the internal Ras levels

after having been in it for a short while. Similar to the speed of the cells relating to the Notch decay, Fig. 12.12 shows that there seems to be one cell at location 114 which still has a very low Ras level. Presumably, this cell was pushed much faster to this location than the other ones and has hence not had enough time to accumulate as much Ras. Apart from this single case, the Ras level does not seem to fluctuate significantly, which indicates a mainly constant forward speed of the cells.

4 Discussion

In the present work, we have developed and calibrated a physical computational model of the *C. elegans* germ line. The model highlights a novel way of using MD as a central engine of computational models simulating physical cell movement. The usage of the accurate physical features allows our model to be calibrated in a way that matches simulation and experimental time. The stress on physical features allows us to combine important developmental aspects such as cell division and

growth in a realistic way. In addition, key molecular pathways, differentiation and apoptosis are added, leading to a realistic model that can be used to an extensive analysis of our concept of the underlying system.

The usefulness of physical and molecular dynamics-like approaches to cell movement has been demonstrated in different settings [22, 24, 25, 27], mostly involving cell populations of tissues. In these models, especially the ones modelling more or less constant tissue-cell populations without cell divisions [24, 25], the physical movement caused by overlap of cells is only a by-product of an active movement along a chemical or nutritional gradient. In models including cell division [22, 27], this aspect of movement is more prominent but still limited due to nutrient dependence or contact inhibition. In our model, however, cells move purely due to physical compression by other cells arising from cell divisions and growth of cells. Our model visually represents this compression by overlap of the circles. To our knowledge, the present model is the first to simulate cell movement purely based on these physical aspects without any active migration. The lack of active movement in our model is in accordance with the experimental observations in the biological system [28, 29].

While still a simplified representation of the germ line, our model produces cell movement that looks very natural, and we were able to calibrate the model in accordance to different experimental observations found in the literature [15, 19, 28] as well as from our own experiments. We have used the cell movement rate derived from our own observations and [28, 29] to incorporate a measure of real time in our model. The comparison allowed us to identify the approximate number of timesteps in our model that represents an hour in real time. Based on this value, the division rate from our model is very much in accordance with the one observed in the literature [28]. The same is true for the death rates that we observed in the model and experimentally. Only the fertilisation rate differed slightly from the value in the literature [15]. Furthermore, the value might change after including the rachis in a revised version of our model, which will change the cell movement close to the end of the germ line. In addition to calibrating, our model allows us to observe properties of single marked cells like density factor, size and kinetics at any timepoint. These aspects can be used in the future to test or evolve new hypotheses like density dependent apoptosis or time-dependent developmental progression of individual cells. Furthermore, these parameters allow feeding the model by results from wet lab experimentation and vice versa. Conclusively, our physical MD modelling approach can be very useful in this setting and extensions to the present model promise to advance our understanding of fundamental biological processes.

5 Future Prospects

We will try to overcome certain limitations of our current simplified model. Including the rachis into our model is very important as this might improve the values of the death and fertilisation rates and will render the model as a whole

visually more realistic. Further, extending the abstract "Notch" and "Ras" signalling pathways by their individual regulatory components could help us to gain a more detailed understanding of the biological system.

The successful calibration of our model renders it amenable to testing hypotheses about the germ line system and about specific aspects of development. Our major goal will be to test different possible mechanisms of apoptosis. It is known that about 50% of all potential oocytes die; one likely explanation is the need for nurse cells that synthesise enough cytoplasm for the rapidly growing oocytes [15, 19]. So far it is unclear how the decision is taken when and which of the cells should die. A probable explanation is that there is a balance between the cytoplasm being produced and the number of cells allowed to progress [19]. Still, apoptosis could be random or clearly determined by aspects like Ras activity level, developmental timing or cell size. Another very interesting aspect of the germ line is the question of symmetric versus asymmetric division of stem cells. It has not been fully resolved whether the post-larval germinal stem cells divide symmetrically or asymmetrically, i.e. if both daughter cells retain stem cell properties and can divide further or if only one of the daughter cells remains in the mitotic zone whereas the sister progresses toward maturation into a gamete, respectively. The prevailing model assumes that the division happens symmetrically [33]. Still, some of the cells might remain in the mitotic zone if they are not pushed by other dividing cells [28]. We will focus our future modelling work on these two and other sparsely understood aspects of the *C. elegans* germ line to gain further understanding and generate input for wet lab experiments that can validate the model's predictions.

Acknowledgements We are grateful to James Margetson for contributing an initial environment program in F# on which this model is based and for his support on the usage of F#. Antje Beyer is grateful to Adrian Hemmen for introducing her to the MD framework and for critical and helpful discussions of the model. This work was supported in part by the European Union grant FP7 PANACEA 222936 (Jasmin Fisher, Michael O. Hengartner and Alex Hajnal) and the Swiss National Science Foundation (Michael O. Hengartner). Antje Beyer is funded by Microsoft Research through its PhD Scholarship Programme.

References

1. Brenner S (1974) The genetics of *Caenorhabditis elegans*. Genetics 77(1):71–94
2. Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH (2005) Genomics in *C. elegans*: so many genes, such a little worm. Genome Res 15(12):1651–1660
3. Potts MB, Cameron S (2010) Cell lineage and cell death: *Caenorhabditis elegans* and cancer research. Nat Rev Cancer 11:50–58
4. Joshi PM, Riddle MR, Djabrayan NJV, Rothman JH (2010) *Caenorhabditis elegans* as a model for stem cell biology. Dev Dynam 239(5):1539–1554
5. Riddle DL, Blumenthal T, Meyer BJ, Priess JR (1997) *C. elegans* II. CSHL, Cold Spring Harbor, New York
6. Corsi AK (2006) A biochemist's guide to *Caenorhabditis elegans*. Anal Biochem 359(1):1–17
7. Sulston J (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. Dev Biol 56(1):110–156

8. Kipreos ET (2005) *C. elegans* cell cycles: invariance and stem cell divisions. Nature reviews. Mol Cell Biol 6(10):766–776
9. Hubbard EJA, Greenstein D (2005) Introduction to the germ line. WormBook: the online review of *C. elegans* biology, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.18.1, http://www.wormbook.org
10. Kimble J, Crittenden SL (2005) Germline proliferation and its control. WormBook: the online review of *C. elegans* biology, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.13.1, http://www.wormbook.org
11. Gartner A, Boag PR, Blackwell TK (2008) Germline survival and apoptosis. WormBook: the online review of *C. elegans* biology, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.145.1, http://www.wormbook.org
12. Korta DZ, Hubbard EJA (2010) Soma-germline interactions that influence germline proliferation in *Caenorhabditis elegans*. Dev Dynam 239(5):1449–1459
13. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell144(5):646–674
14. Krantic S, Mechawar N, Reix S, Quirion R (2005) Molecular basis of programmed cell death involved in neurodegeneration. Trends Neurosci 28(12):670–676
15. Hirsh D, Oppenheim D, Klass M (1976) Development of the reproductive system of *Caenorhabditis elegans*. Dev Biol 49(1):200–219
16. Hansen D, Hubbard EJA, Schedl T (2004) Multi-pathway control of the proliferation versus meiotic development decision in the *Caenorhabditis elegans* germline. Dev Biol 268(2):342–357
17. Kimble J, White J (1981) On the control of germ cell development in *Caenorhabditis elegans* 1. Dev Biol 81(2):208–219
18. Waters KA, Reinke V (2011) Extrinsic and intrinsic control of germ cell proliferation in *Caenorhabditis elegans*. Mol Rep Dev 78(3):151–160
19. Gumienny TL, Lambie E, Hartwieg E, Horvitz HR, Hengartner MO (1999) Genetic control of programmed cell death in the *Caenorhabditis elegans* hermaphrodite germline. Development 126(5):1011–1022
20. Church DL, Guan KL, Lambie EJ (1995) Three genes of the MAP kinase cascade, mek-2, mpk-1/sur-1 and let-60 ras, are required for meiotic cell cycle progression in *Caenorhabditis elegans*. Development 121(8):2525–2535
21. Alder BJ, Wainwright TE (1959) Studies in molecular dynamics. I. General method. J Chem Phys 31(2):459
22. Drasdo D, Kree R, McCaskill J (1995) Monte Carlo approach to tissue-cell populations. Phys Rev E 52(6):6635–6657
23. Beyer T, Meyer-Hermann M (2007) Modeling emergent tissue organization involving high-speed migrating cells in a flow equilibrium. Phys Rev E 76(2):27
24. Beyer T, Meyer-Hermann M (2009) Multiscale modeling of cell mechanics and tissue organization. IEEE Eng Med Biol Mag 28(2):38–45
25. Palsson E (2001) A three-dimensional model of cell movement in multicellular systems. Future Gener Comput Syst 17(7):835–852
26. Palsson E, Othmer HG (2000) A model for individual and collective cell movement in *Dictyostelium discoideum*. Proc Natl Acad Sci USA 97(19):10448–10453
27. Schaller G, Meyer-Hermann M (2005) Multicellular tumor spheroid in an off-lattice Voronoi–Delaunay cell model. Phys Rev E 71(5):1–16
28. Crittenden SL, Leonhard KA, Byrd DT, Kimble J (2006) Cellular analyses of the mitotic region in the *Caenorhabditis elegans* adult germ line. Mol Biol Cell 17(7):3051–3061
29. Maciejowski J, Ugel N, Mishra B, Isopi M, Hubbard EJA (2006) Quantitative analysis of germline mitosis in adult *C. elegans*. Dev Biol 292(1):142–151
30. Swope WCA (1982) Computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. J Chem Phys 76(1):637

31. Syme D, Granicz A, Cisternino A (2007) Expert F#. Apress, Berkeley, California
32. Accompanying movie [Internet] (2011) [cited 2011 Mar 27]. http://www.cs.le.ac.uk/people/npiterman/publications/2011/BEPHHF/index.html
33. Cinquin O, Crittenden SL, Morgan DE, Kimble J (2010) Progression from a stem cell-like state to early differentiation in the *C. elegans* germ line. Proc Natl Acad Sci USA 107(5):2048–2053

Chapter 13
A Modular Model of the Apoptosis Machinery

E.O. Kutumova, I.N. Kiselev, R.N. Sharipov, I.N. Lavrik, and Fedor A. Kolpakov

Abstract Using a modular principle of computer hardware as a metaphor, we defined and implemented in the BioUML platform a module concept for biological pathways. BioUML provides a user interface to create modular models and convert them automatically into plain models for further simulations. Using this approach, we created the apoptosis model including 13 modules: death stimuli (*TRAIL*, *CD95L*, and *TNF-α*)-induced activation of *caspase-8*; survival stimuli (*p53*, *EGF*, and *NF-κB*) regulation; the mitochondria level; *cytochrome C*- and *Smac*-induced activation of *caspase-3*; direct activation of effector caspases by *caspase-8* and *−12*; *PARP* and apoptosis execution phase modules. Each module is based on earlier published models and extended by data from the Reactome and TRANSPATH databases. The model ability to simulate the apoptosis-related processes was checked; the modules were validated using experimental data. Availability: http://www.biouml.org/apoptosis.shtml.

E.O. Kutumova (✉) • I.N. Kiselev • F.A. Kolpakov
Institute of Systems Biology, Ltd, Novosibirsk, Russia

Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia
e-mail: helenka@biouml.org; axec@biouml.org; fedor@biouml.org

R.N. Sharipov
Institute of Systems Biology, Ltd, Novosibirsk, Russia

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: shrus79@biouml.org

I.N. Lavrik
German Cancer Research Center (DKFZ), Heidelberg, Germany
e-mail: i.lavrik@dkfz-heidelberg.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_13,

1 Introduction

Apoptosis is a highly regulated and evolutionary conserved pathway of cell death that plays a critical role in development and maintenance of tissue homeostasis. Over the recent years, molecular biologists have significantly enriched the formal description of the pro- and anti-apoptotic machineries, and amassed a range of mathematical models [1–10]. However, these models mainly describe different segments of implicated pathways, and a comprehensive model of the apoptosis regulation still does not exist. The motivation for creating a more extensive model is that there exists experimental information on the interactions of the subsystems – interactions which the submodels cannot account for [11, 12]. At the same time, the analysis of the large model is difficult. Therefore, we need to modularize the model for better understanding of how the parts of the model interact with each other. In our case, modularization has a similar meaning as a model aggregation in [11].

For creation of the modular model of the pro- and anti-apoptotic machineries, we used BioUML (http://www.biouml.org) – an open source integrated Java platform for systems biology. It spans the comprehensive range of capabilities including access to databases with experimental data, tools for formalized description of biological systems structure and functioning, as well as tools for their visualization, simulation, parameters fitting, and analyses. Plug-in based architecture (Eclipse runtime from IBM is used) allows addition of new functionality using plug-ins.

In this work, we extended the BioUML platform for support of modular models by:

- A new diagram type – composite diagrams using modules as components.
- A new convertor for transformation of the modular models into the plain models for simulation.

2 Modularity

We define a "module" as a group of reactions with a specified set of input, output, and contact ports (Fig. 13.1a), like an electronic board has a set of inputs, outputs, and contacts (Fig. 13.1b).

The input and output ports [11] must be defined so that the modules can be linked together unambiguously. Additionally, we define the contact ports linking common parameters of the model.

If two different modules, for example, a module resulting in the activation of *procaspase-8* by *CD95L* (Fig. 13.1b) and the module describing the direct activation of executioner caspases by *caspase-8* (Fig. 13.2), contain the same species *caspase-8*, then in the modular model they will be renamed, like *A_caspase-8* and *B_caspase-8,* respectively.

Furthermore, if you want to connect these modules, and to merge the renamed molecules declaring that *caspase-8* in the *A* module is the same as *caspase-8* in the *B* module, you can do this via the contact ports interaction. You can also

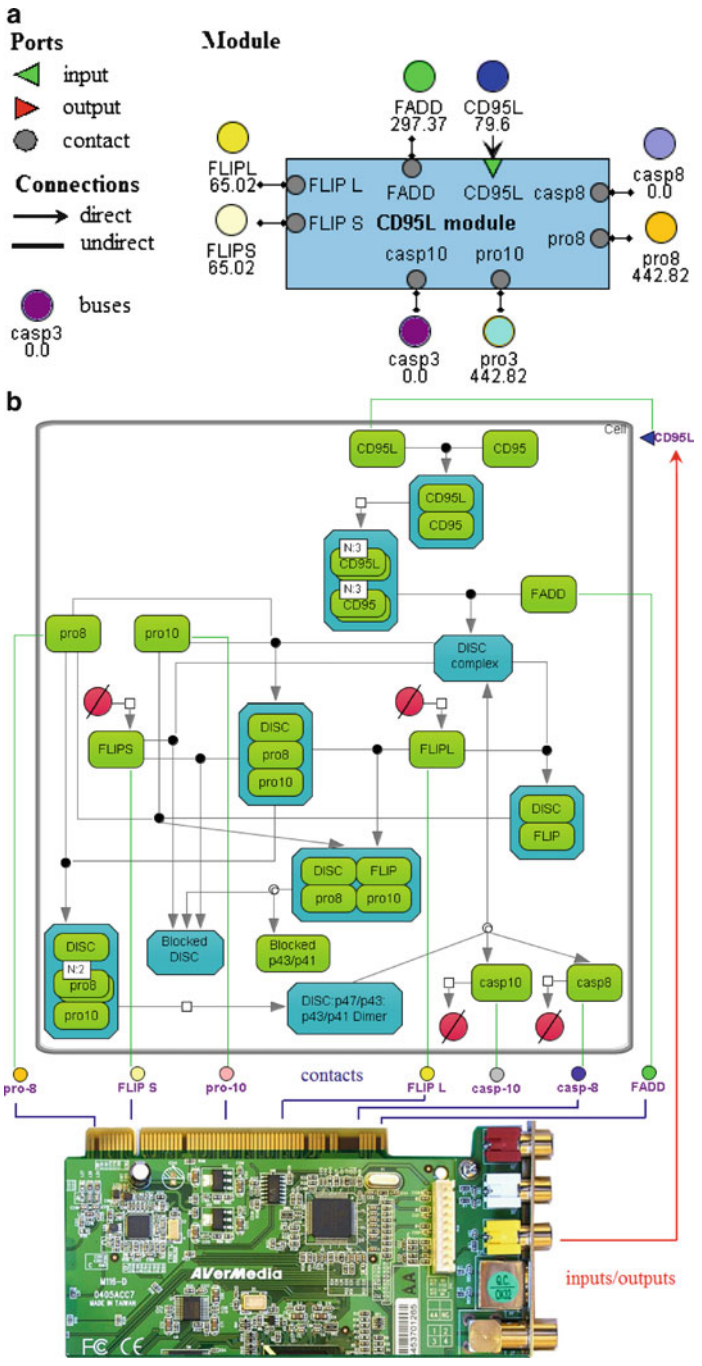


Fig. 13.1 Explanation of the module conception on the base of the *CD95L* module interface. (**a**) The graphic notation for the module depiction in BioUML. (**b**) The analogy of the port usage in the modular modeling and computer hardware

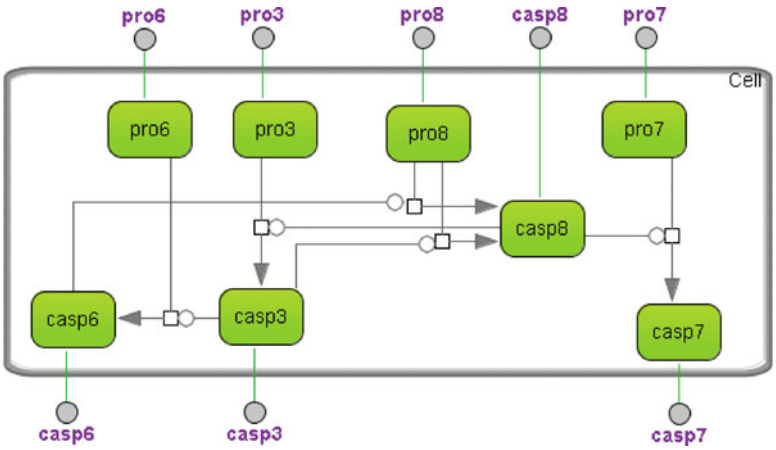


Fig. 13.2 The module of the direct executioner caspases activation, based on the model of Bentele et al. [1]

assign some species as inputs (e.g., *CD95L*) and some as outputs, and specify the mathematical functions in a way that takes into account the concentrations at which the species will pass from one module to another.

3 Generation of Plain Models

In order to obtain a plain, appropriate for simulation model from the modular one, we implemented a flattening algorithm in the BioUML platform. The input for the algorithm is a composite diagram which consists of subdiagrams and the connections between them (Fig. 13.3a). Connections can be of two types:

1. A directed connection between two parameters means that one parameter should be completely replaced, in a simple case, by a parameter from another module, or by an arbitrary function which depends on the parameters from another module.
2. An undirected connection means that two parameters could be changed by several modules simultaneously.

For visual simplicity, we have added buses to the composite diagram type which can be used as transitional nodes for connections (two buses corresponding to one variable may be located far apart in the diagram, and not be connected).

Submodels may directly define their input, output, and contact parameters with corresponding ports. Directed connections may be established between output port and input port, and undirected between two contact ports. This approach corresponds to the model aggregation concept from [11], and is used when modules are initially designed to be submodels, so they can just be connected appropriately in the composite model. However, for greater flexibility, it is allowed to establish

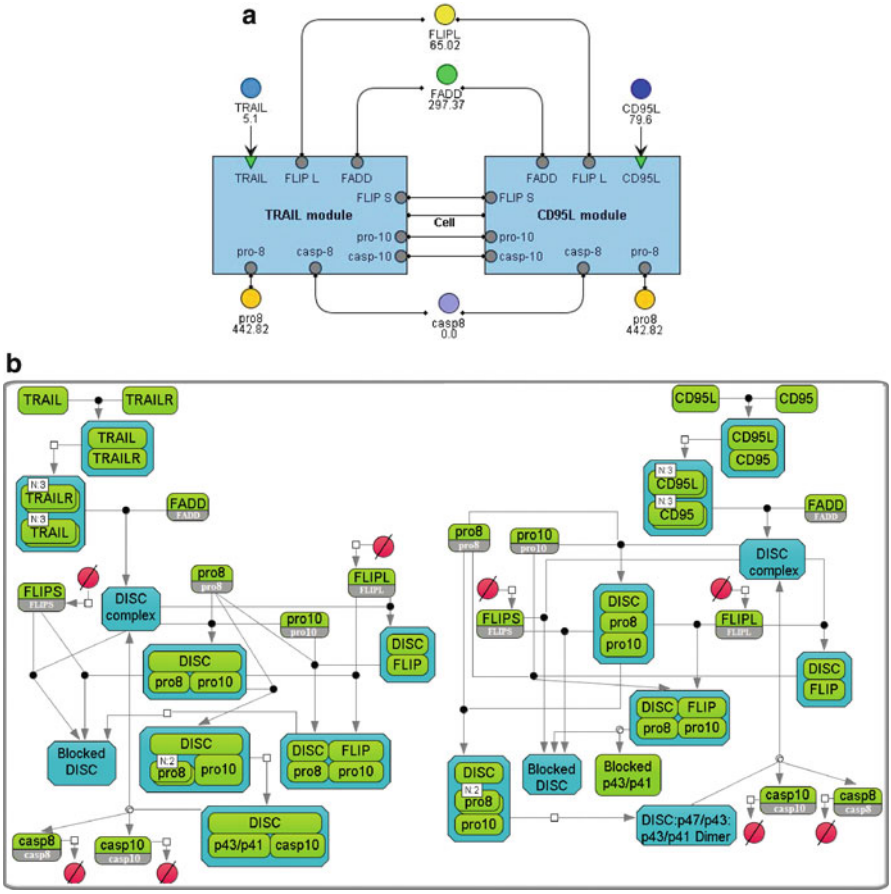


Fig. 13.3 The example of transformation of the modular model into the plain model describing interactions between *TRAIL* and *CD95* pathways in BioUML: (**a**) the modular model; (**b**) the resulting plain model

connections between any parameters of the modules. One composite diagram may contain both connections through ports and connections directly to inner variables and parameters of the modules. This approach corresponds to the composition concept as it is defined in [12].

The output of the algorithm is a plain model (Fig. 13.3b) of the same type as the submodels (in our case a SBML diagram that can be simulated in BioUML using ODE solvers). The result of the algorithm is flattening of the modules [12]. Thus, we have combined three approaches: aggregation, composition, and flattening.

The algorithm consists of three steps considered below.

Step 1 Generate unique names for all parameters and variables whose names are repeated in several modules of the model.

Step 2 Generate substitution rules for species and parameters of the model. For each directed connection we have a mapping like $p \leftarrow f(p_1, \ldots, p_n)$. For each p_i, we may also have an input connection.

Therefore, we may have consequent connections:

$$\begin{gathered} p_1 \leftarrow f_1\left(p_{11}, \ldots, p_{1m}\right) \\ \ldots \\ p_n \leftarrow f_n\left(p_{n1}, \ldots, p_{nm}\right). \end{gathered}$$

This situation is resolved by the function substitution:

$$p \leftarrow f\left(f_1\left(p_{11}, \ldots, p_{1m}\right), \ldots, f_n\left(p_{n1}, \ldots, p_{nm}\right)\right) \leftarrow F\left(q_1, \ldots, q_k\right).$$

Here, $q_1, \ldots, q_k$ have no input directed connections. For each undirected connection we have a mapping $(p_1 \leftrightarrow p_2)$, therefore we may have:

$$p_1 \leftrightarrow p_2 \leftrightarrow \ldots \leftrightarrow p_n.$$

From those variables, we choose the main one (if one of p_i is a species, then the algorithm will choose it as the main variable and create a node for it in the diagram) and transform the mapping to

$$p \rightarrow \{p_1, \ldots, p_n\}.$$

Finally, we have a set of substitution rules. They can be of two types:

1. $p \leftarrow F(p_1, \ldots, p_n)$
2. $p \rightarrow \{p_1, \ldots, p_n\}$

Let U be the set of variables which are on the left side of the second type rule (the main variables), $U(p)$ – variables that will be substituted by p according to this rule, and D – variables that will be substituted according to the first type rule. At first, we should make sure that there can be only one substitution for each variable

$$D \cap \cup U(p) = \varnothing.$$

For that purpose, we consider the situation where one parameter is an input for some directed connection, and has an undirected connection

$$p_1 \rightarrow \{p_2\}, \quad p_2 \leftarrow F\left(p_3\right)$$

as illegal, and do not allow it on the level of composite diagram building. The situation where one parameter has two directed connections as input is also illegal:

$$F\left(p_1\right) \rightarrow p_2 \leftarrow G\left(p_3\right).$$

The situation

$$p_1 \to \{p_2\}, \quad F(p_2) \to p_3$$

is legal, and for the correct function of the algorithm this should be transformed to

$$p_1 \to \{p_2\}, \quad F(p_1) \to p_3.$$

Step 3 Iterating through the subdiagrams for each element, we create a copy of each element that will be added to a plain (not composite) diagram. All iterative methods are implemented with the same interface. Method *read* (*oldCompartment*, *subDiagram*, and *newCompartment*) gets element x from *oldCompartment*, creates a copy, and puts it to *newCompartment* in the plain diagram. The first level compartment is the diagram. The algorithm is recursive, if x is the compartment itself, then we apply this method to it: *read* (x, *subdiagrams*, and *copy*(x)).

There are three different reading methods which are executed in a strict order.

1. Read species and compartments
   Add all species and compartments from all subdiagrams to the plain diagram. Let x be our diagram element containing some variable p. The following rules are used.
   a. If $\exists q : \ p \in U(q)$, then this species will be added to the diagram with the attribute "clone" and replaced variable (q instead of p). If x is a compartment, we do not add it to the diagram.
   b. If $p \in U$, then we add it to the plain diagram and reattach all edges from all species $U(p)$. If x is the compartment, we read all content of the connected compartments from $U(p)$ and copy it to the copy of x in the plain diagram.
   c. If $p \in D$, then we copy x to the plain diagram, set p to the boundary conditions, and add a new equation which defines the value of p with the formula from the directed connection.
   d. If x is a compartment, we recursively apply the method to x.
2. Read equations, and other nodes
   a. If a node represents a rate equation for parameter $p(\mathrm{d}p/\mathrm{d}t = f)$ and $\exists q : \ p \in U(q)$, then we ignore that equation.
   b. If $p \in U$, then we merge all rate equations for the connected variables with the current equation:

$$\mathrm{d}p/\mathrm{d}t = f + \sum f_i, \quad \text{where} \quad \mathrm{d}p_i/\mathrm{d}t = f_i \quad \text{and} \quad p_i \in U(p).$$

   c. If a node represents an equation which defines in any way (differential, scalar, algebraic, event, etc.) the parameter p, and $p \in D$, then we do not add this equation to the plain diagram.
   d. If an equation formula contains a variable $p \in D$, then all inclusions of p should be replaced by the corresponding formula from the directed connection. If an equation formula contains a variable p and $\exists\, q : \ p \in U(q)$, then all inclusions of p should be replaced by q.

Of course, we may achieve an inconsistent algebraic system in situations when two parameters are connected and both have algebraic or scalar rules, for example:

$$F(p) = 0, \quad p \in U(q), \quad G(q) = 0.$$

However, in such situations, we have no reasons to choose one of the rules and remove the others, so this task is beyond the current algorithm.

3. *Read edges* After all nodes are added, we have a correspondence of the old node from subdiagram to the new node. This mapping helps us to add edges. For example, if one compartment is to be substituted by another, we do not add it to the plain diagram. Instead, we map it to the copy of the substituted compartment in the plain diagram, so that all edges are automatically reattached.

4 The Modular Model of Apoptosis

We combined the individual models and obtained a comprehensive map of the pro- and anti-apoptotic pathways for in silico experiments. Then we extended this map using information from the Reactome [13] and TRANSPATH [14] databases, and divided it into the separate modules available at http://www.biouml.org/apoptosis.shtml. Figure 13.4 shows the resulting workflow of the modeling procedure. The ability of the modular model to simulate the apoptosis-related processes was checked successfully.

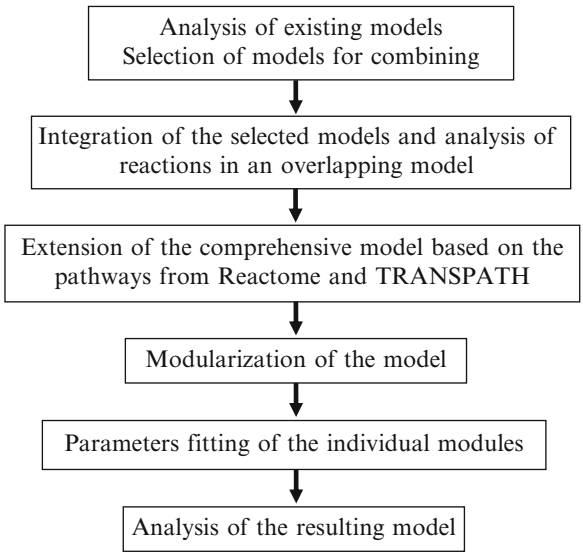


Fig. 13.4 Workflow of the modeling procedure

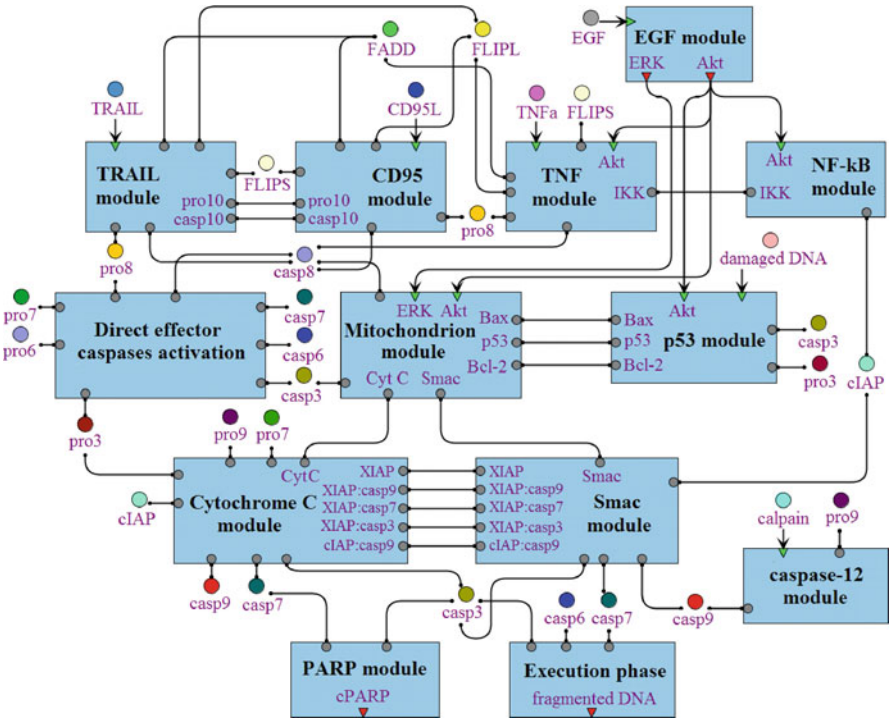


Fig. 13.5 The overview of the modular model of apoptosis

The modular model (Fig. 13.5) includes 13 functional modules located in five different compartments (nucleus, cytoplasm, mitochondria, extracellular space, and endosomal volume). It comprises 279 species (proteins, their complexes, modifications such as different forms of the same molecule, and transformations, for example, phosphorylation) and 372 reactions applying mass action as well as Michaelis–Menten kinetics with 459 parameters.

For estimation of the model parameters, we used experimental data obtained from literature for human cell lines, and represented time courses expressed as relative values of protein concentrations. Table 13.1 contains the data used for estimation of the death stimuli pathway (*CD95L, TRAIL*, and *TNF-α*) modules. In order to estimate the parameters of these modules, we performed a multi-experiment fitting.

We developed an optimization plug-in for BioUML to solve the non-linear optimization problems regarding biochemical pathways by minimization of the distance between model simulation results and experimental data that are time courses or steady states expressed as exact or relative values of substance concentrations. The plug-in includes the range of the optimization methods attempting to minimize the distance using mean, mean square, or standard deviation weight criterions.

Figure 13.6 shows an example of the parameters fitting performed for the *CD95L* module based on the experimental datasets [1,9,15,16] obtained for different human cell lines.

Table 13.1 The experimental data used for the *CD95L*, *TRAIL*, and *TNF-α* modules fitting

References	Cell line	Modules
Bentele et al., 2004 [1]	SKW 6.4	*CD95L*
Hua et al., 2005 [9]	Jurkat	
Neumann et al., 2010 [15]	HeLa	
Scaffidi et al., 1998 [16]	CEM	
Vilimanovich et al., 2008 [17]	LN-71, U343MG	*TRAIL*
Farfan et al., 2004 [18]	Jurkat	
Janes et al., 2006 [19]	HT29	*TNF*

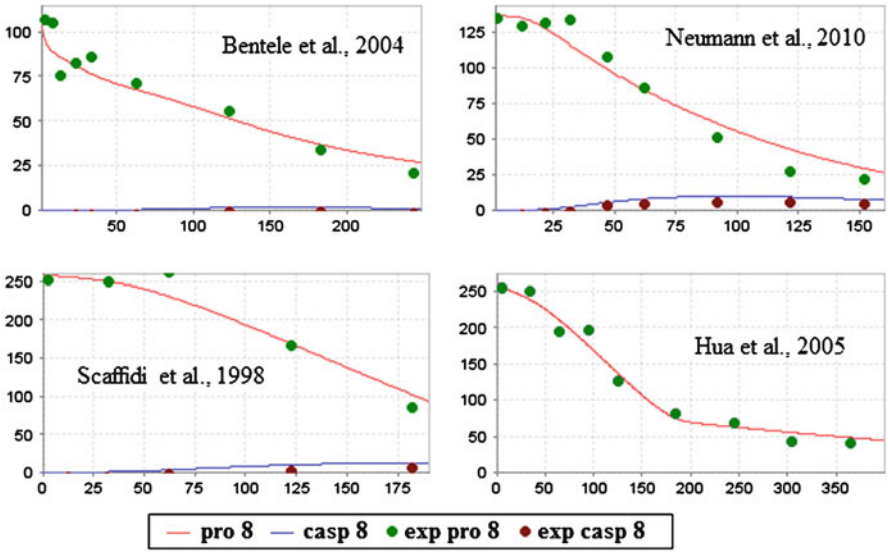


Fig. 13.6 Results of the *CD95L* module simultaneous fitting to the experimental datasets obtained for different human cell lines

5 Conclusion

We developed the modular model of apoptosis based on the existing mathematical models, as well as on information from the Reactome and TRANSPATH databases. The models were reconstructed using the SBML and SBGN standards, and converted into the modules with specified ports and connections. Modularization allows for analyzing the individual parts of the whole model and relations among the various modules.

This work was supported by the European Committee grants No 037590 "Net2Drug" and No 202272 "LipidomicNet."

The modular model is available at http://www.biouml.org/apoptosis.shtml.

References

1. Bentele M, Lavrik I, Ulrich M, Stößer S, Heermann DW, Kalthoff H, Krammer PH, Eils R (2004) Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. J Cell Biol 166(6):839–851
2. Hoffmann A, Levchenko A, Scott ML, Baltimore D (2002) The IkB–NF-kB signaling module: temporal control and selective gene activation. Science 298:1241–1245
3. Hamada H, Tashima Y, Kisaka Y, Iwamoto K, Hanai T, Eguchi Y, Okamoto M (2008) Sophisticated framework between cell cycle arrest and apoptosis induction based on p53 dynamics. PLoS One 4(3):e4795:1–7
4. Bagci EZ, Vodovotz Y, Billiar TR, Ermentrout GB, Bahar I (2006) Bistability in apoptosis: roles of Bax, Bcl-2 and mitochondrial permeability transition pores. Biophys J 90:1546–1559
5. Legewie S, Bluthgen N, Herzel H (2006) Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. PLoS Comput Biol 2(9):e120:1061–1073
6. Rangamani P, Sirovich L (2007) Survival and apoptotic pathways initiated by TNF-alpha: modeling and predictions. Biotechnol Bioeng 97(5):1216–1229
7. Cho K-H, Shin S-Y, Lee H-W, Wolkenhauer O (2003) Investigations into the analysis and modeling of the TNFα-mediated NF-κ B-signaling pathway. Genome Res 13:2413–2422
8. Albeck JG, Burke JM, Spencer SL, Lauffenburger DA, Sorger PK (2008) Modeling a snap-action, variable-delay switch controlling extrinsic cell death. PLoS Biol 6(12):e299
9. Hua F, Cornejo MG, Cardone MH, Stokes CL, Lauffenburger DA (2005) Effects of Bcl-2 levels on Fas signaling-induced caspase-3 activation: molecular genetic tests of computational model predictions. J Immunol 175:985–995
10. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol 20:370–375
11. Randhawa R, Shaffer CA, Tyson JJ (2009) Model aggregation: a building-block approach to creating large macromolecular regulatory networks. Bioinformatics 25(24):3289–3295
12. Randhawa R, Shaffer CA, Tyson JJ (2010) Model composition for macromolecular regulatory networks. IEEE/ACM Trans Comput Biol Bioinform 7(2):278–287
13. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33:D428–D432
14. Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. Nucleic Acids Res 31(1):97–100
15. Neumann L, Pforr C, Beaudouin J, Pappa A, Fricker N, Krammer PH, Lavrik IN, Eils R (2010) Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. Mol Syst Biol 6:352
16. Scaffidi C, Fulda S, Srinivasan A, Friesen C, Li F, Tomaselli KJ, Debatin K-M, Krammer PH, Peter ME (1998) Two CD95 (APO-1/Fas) signaling pathways. EMBO J 17(6):1675–1687
17. Vilimanovich U, Bumbasirevic V (2008) TRAIL induces proliferation of human glioma cells by c-FLIPL-mediated activation of ERK1/2. Cell Mol Life Sci 65:814–826
18. Farfan A, Yeager T, Moravec R, Niles A (2004) Multiplexing homogeneous cell-based assays. Cell Notes 10:15–18
19. Janes KA, Gaudet S, Albeck JG, Nielsen UB, Lauffenburger DA, Sorger PK (2006) The response of human epithelial cells to TNF involves an inducible autocrine cascade. Cell 124:1225–1239

Chapter 14
An Ensemble Approach for Inferring Semi-quantitative Regulatory Dynamics for the Differentiation of Mouse Embryonic Stem Cells Using Prior Knowledge

Dominik Lutter*, Philipp Bruns*, and Fabian J. Theis

Abstract The process of differentiation of embryonic stem cells (ESCs) is currently becoming the focus of many systems biologists not only due to mechanistic interest but also since it is expected to play an increasingly important role in regenerative medicine, in particular with the advert to induced pluripotent stem cells. These ESCs give rise to the formation of the three germ layers and therefore to the formation of all tissues and organs. Here, we present a computational method for inferring regulatory interactions between the genes involved in ESC differentiation based on time resolved microarray profiles. Fully quantitative methods are commonly unavailable on such large-scale data; on the other hand, purely qualitative methods may fail to capture some of the more detailed regulations. Our method combines the beneficial aspects of qualitative and quantitative (ODE-based) modeling approaches searching for quantitative interaction coefficients in a discrete and qualitative state space. We further optimize on an ensemble of networks to detect essential properties and compare networks with respect to robustness. Applied to a toy model our method is able to reconstruct the original network and outperforms an entire discrete boolean approach. In particular, we show that

*The authors Dominik Lutter and Philipp Bruns contributed equally to this work.

D. Lutter (✉) • F.J. Theis
Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum München, Munich, Germany
e-mail: dominik.lutter@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de

P. Bruns
Institute of Bioinformatics and Systems Biology, CMB, Helmholtz Zentrum München, Munich, Germany

Department of Surgery, Technische Universität München, Munich, Germany
e-mail: philipp.bruns@helmholtz-muenchen.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_14, © Springer Science+Business Media, LLC 2012

including prior knowledge leads to more accurate results. Applied to data from differentiating mouse ESCs reveals new regulatory interactions, in particular we confirm the activation of Foxh1 through Oct4, mediating Nodal signaling.

1 Introduction

Systems biology as a new field in biological research has developed and explored a diversity of methods and tools to investigate regulatory models of genes and their products such as proteins and RNAs [1–6]. In the majority of cases these models form gene regulatory networks (GRNs), that can be represented as simple node–edge graphs, where the nodes stand for the genes or proteins and the edges represent their interactions. But in many cases the available knowledge about these interactions is poor leading to either incomplete or imprecise and thus hardly interpretable models. Thus, inferring missing network edges from data allows to predict novel biological interactions. In practice, inferring network interactions from biological data faces several drawbacks like experimental and biological noise, overfitting, indeterminacies, and infinite solution spaces.

However, boolean modeling is considered as a highly abstract form of modeling and has been successfully applied to biological systems [7–9]. The benefit of boolean modeling is its simplicity and thus the relatively low number of unknown parameters. Furthermore boolean approaches have also been used to model and analyze time dependent genetic dynamics [10]. In contrast to continuous ODE based models where dynamics can be modeled within infinite time steps, boolean modeling is limited to discrete and qualitative states. By now, several approaches have been developed to combine these two methods and benefit from boolean simplicity and deal with continuous dynamics [11, 12].

Here we present an intermediate approach working with continuous regulation coefficients, continuous expression values, and discrete time steps [13, 14]. With this modification we are now able to add RNA concentrations to our model but keep ON/OFF states for communication between nodes. The benefit of this method is that we do not need to define any boolean regulation functions ab initio for each node but still deal with the convenient boolean modeling framework. This method is closely related to piecewise linear methods that were also applied for the analysis of GRNs [15, 16]. In contrast to ODE systems based on mass action or Michaelis–Menten kinetics our approach is less precise in dynamics, but allows for a much more flexible search of new possible interactions.

The model itself is now defined by a number of genes that interact among themselves by activating or repressing their expressions. Known interactions between the genes were used to define prior information. Already known interactions can now be used to initialize parameters in order to reduce the set of computationally feasible solutions to a set of potentially biologically reasonable solutions. The strength of the interactions is modeled by the edge-weights that form our parameters. This allows us to adopt the beneficial aspect of relative interaction weightings without introducing too many additional parameters. The implemented algorithm is available upon request to the authors.

Embryonic stem cells (ESCs) appear to play a major role in future medicine since their ability of unlimited self renewal and the potential of forming any differentiated cell type [17]. Thus, these pluripotent cells will form a basis for many new therapies for diseases like cancer, diabetes, neurodegenerative diseases, and many more. Modern high throughput techniques allow to measure the dynamics of gene expression during differentiation, but the regulatory mechanisms driving these dynamics are widely unknown. Inferring GRNs from expression data is therefore a promising but challenging task, that helps to understand the molecular mechanisms driving the differentiation of the cell [18]. Moreover, the self maintenance of stem cells is so far only partly understood and several models have been developed and analyzed [19, 20]. In contrast, the early events that determine cell fate and the genetic machinery that drives segregation are widely unclear [21].

Although several key genes involved in murine lineage segregation are well known, like *Foxa2* and *Sox17* specifying endoderm [22], their functions during differentiation are barely known. Taking into account transcription factors like Sox2, Oct4, and Nanog, that form the core of ESC pluripotency [23, 24], the aim is to create a regulatory model that helps to understand and predict cellular events when ESCs leave pluripotency and differentiate.

In this work we investigate ESC differentiation by inferring a gene regulatory model from gene expression data with use of prior information. Using a toy model we can show that our approach outperforms discrete boolean modeling and improves with the use of prior knowledge towards producing more robust results. Through not only optimizing for one solution, but an ensemble of networks we are able to select for networks with specific attributes like robustness against experimental and biological noise and to look for common and therefore essential properties. We need to point out that information gained through text-mining approaches might be incorrect or useless in this context. For instance, direct effects of protein–protein interactions were not covered by our data since we are only working with microarray data which is based on mRNA levels. However, we show that our approach is able to correctly reproduce a toy model and gains accuracy from the use of prior data. Applied to experimental data, we predict new interactions, whereby the activation of *Foxh1* by Oct4 could be confirmed from published data. Furthermore, we find by adding noise and doing cross validation, that robust networks appear to be remarkably sparse, thus managing to reproduce the data with fewer interactions.

2 Methods

2.1 Mouse Gene Expression Data

For gene expression analysis we used the following dataset available at GEO *www.ncbi.nlm.nih.gov/geo/* [25]: Time course of ESC differentiation into embryoid bodies (EB) (GSE3749, GSE3231 GSE2972). CEL files were analyzed using

Bioconductors simpleaffy package for R [26] and expression values were calculated using the rma algorithm. All gene names and gene symbols to each probe were retrieved from Bioconductors moe430a.db package. The dataset consisted of three mouse ESC lines V6.5, R1, and J1. RNA was measured at 11 time points from $t = 0\,\text{h}$ until $t = 14\,\text{d}$. From each time point and each cell line three technical replicates were measured.

2.2 *Semi-quantitative Modeling Based on External and Internal State Vectors*

In this work, a GRN is described as a *directed*, *weighted graph*, where V denotes the set of genes (vertices) and E denotes the interactions with relative coefficients. For a given set of n genes, V is a vector of n vertices and E forms a $n \times n$ matrix consisting of the edge weight coefficients. Each entry $e_{i,j}$, with $i, j = 1, \ldots, n$, denotes the weight of the edge from gene i to gene j, where positive weights stand for activation and negative for inhibition, respectively.

In contrast to ODE systems, where the system can be described with an infinite number of possible state vectors [12, 27], in this approach a state vector is split up into an external and an internal state vector. The internal state vector $\mathbf{c}_t$ corresponds to the continuous expression levels of all genes at time t for $t = 1, \ldots, m$ where m is the number of all measured time points, thus forming the columns of an expression matrix C with $\mathbf{c}_i$ the time dependent expression profiles for each gene i. For each internal state vector a corresponding external state exists as a boolean vector $\hat{\mathbf{c}}_t$ referring to an either active or an inactive binary state. Each element of $\hat{c}_{i,t}$, is then assigned to a binary value using a threshold ε on the present expression value $c_{i,t}$. After scaling the gene expression data between 0 and 1, we set $\varepsilon = 0.5$. The update rule for each gene i now follows:

$$c_{i,t+\Delta t} := \begin{cases} c_{i,t} + \dfrac{1 - c_{i,t}}{\tau_i} & \text{if} \quad b_i > \theta \\ c_{i,t} + \dfrac{-c_{i,t}}{\tau_i} & \text{otherwise} \end{cases}, \tag{14.1}$$

where $\mathbf{b}$ is the activation vector formed by $\mathbf{b} := E^\top \hat{\mathbf{c}}$, where $E^\top$ is the transposed matrix of E. Thus, b_i is the sum of all relative interaction coefficients of all active regulators of a gene i. The expression of gene i is now increased if the activation value given with b_i is above a predefined threshold θ, or decreased otherwise. To give all interaction coefficients the same weight we set $\theta = 0$. Hence, activation or inhibition of a gene i depends on the current state of its regulators, given by $\hat{\mathbf{c}}$ and the relative weights given by the ith column of the matrix E. τ_i denotes a gene specific time constant defining the speed of ascent and descent of the concentrations over time. For all non-boolean simulations the parameter τ is included in the optimization

process. Setting τ for all genes to 1 and allowing only for discrete weights in E, (14.1) leads to the following generic boolean update rule:

$$c_{i,t+\Delta t} := \begin{cases} 1 \text{ if } \left| \left\{ j : (c_j = 1 \wedge e_{ji} = 1) \right\} \right| - \left| \left\{ j : (c_j = 1 \wedge e_{ji} = -1) \right\} \right| > 0 \\ 0 \text{ otherwise} \end{cases} . \tag{14.2}$$

2.3 *Inference of an Ensemble of Networks Using Optimization via a Genetic Algorithm*

Beginning from each measured state vector $\mathbf{c}_t$, the model is simulated for N update steps, thus generating N internal simulated state vectors $\tilde{\mathbf{c}}_{t,1} \ldots \tilde{\mathbf{c}}_{t,N}$. To evaluate the model we determined the simulated state $\tilde{\mathbf{c}}_{t,\min}$ with a minimum average distance $d_t = 1/n \sum_{i=1}^{n} |c_{i,t+1} - \tilde{c}_{i,t,N}|$ to the following internal state vector $\mathbf{c}_{t+1}$. After a full simulation the column-vectors $\tilde{\mathbf{c}}_{t,\min}$ form a simulated state matrix $\tilde{C}$, where $\tilde{\mathbf{c}}_i$ is a simulated expression profile for a gene i.

The overall error of a given model was measured as the difference $\text{err} = \mu_\Delta - \mu_r$, where μ_Δ is given by the average pairwise differences $\mu_\Delta := 1/(m-1) \sum_{t=1}^{m-1} d_t$ and $\mu_r = \sum_{i=1}^{n} \text{corr}(\mathbf{c}_i, \tilde{\mathbf{c}}_i)$ by the sum of all pearson correlation coefficients between simulated and measured gene expression profiles.

We fitted our model to the data using a generic genetic algorithm (GA) implemented within the MathWorks global optimization toolbox. The individuals of the initial populations were initialized randomly using a normal distribution, $\mathcal{N}(0, 0.2)$ for the edges. Prior knowledge about interactions with a known effect is included by addition of $+0.5$ (activation) or -0.5 (inhibition) to the edges. Knowledge about interactions with an unknown effect is included by addition of further random noise ($\mathcal{N}(0, 0.3)$).

After each simulation step a new generation was created by the GA. To keep connectivity low, between creating a new generation and the following simulation step, all edges with an absolute weight below 0.2 were set to zero. Edge weights were also constrained between $+1$ and -1. However, since our method allows for any edge between the model genes, it may also include indirect edges after optimization. In our case we can distinguish between two types of indirect edges: Redundant indirect edges, an edge between two vertices v_1 and v_3 where the true interaction is mediated through v_2, and hidden indirect edges, where the mediating gene is not included in the model. The latter type arises from an incomplete model setup and can mainly be avoided by a careful gene selection, but still may occur due to missing biological knowledge. One way to avoid the first type is to reduce network density. Therefore, we additionally penalize redundant edges by adding the network density to the optimization error. In order to minimize density only on consistent networks the density term was added only after the error "err" converged to a generic GA stopping criterion.

3 Results and Discussion

3.1 *Optimization on a Toy Model Results in Few Consistent Network Solutions*

To test our approach we generated a toy model shown in Fig. 14.1 consisting of a feedback loop (vertices v_1, v_3, and v_4), an extended feedforward loop (vertices v_1, v_2, v_3, and v_5) and a linear activation motif (vertices v_1, v_2, and v_5). With this model we generated artificial expression data that were used as training set (see Fig. 14.1b) starting with an arbitrarily chosen initial expression vector $c_0 = (1, 0, 0, 0, 0)^\top$ and a unified production and decay rate based on $\tau_i = 2$ for all 5 genes.

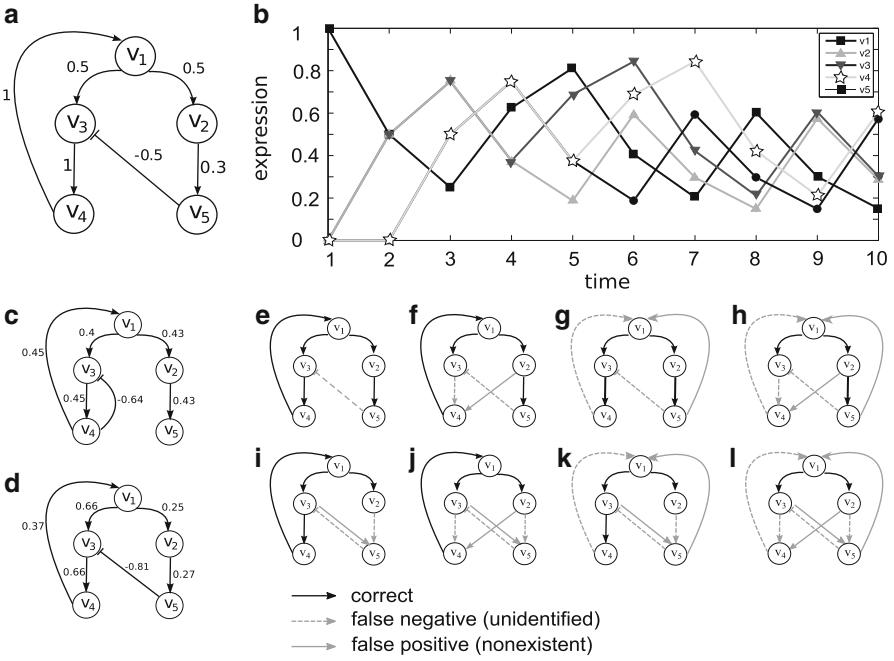


Fig. 14.1 (**a**) Our toy model consisting of five nodes and six weighted and directed edges. (**b**) Expression data generated with the toy model using an arbitrary initial vector (see text) and a time constant $\tau = 2$ for ten update steps. (**c**–**d**) The two consensus networks generated from the two clusters obtained from the 20 best fitting solutions. Mean edge weights per cluster were indicated. (**e**–**l**) All equally rated solutions obtained by the discrete optimization. Black solid edges indicate for correctly identified original edges, grey dashed edges indicate for false negative (not identified but present in toy model) and grey solid edges indicate for false positive edges (present in optimized but not in original model)

Table 14.1 The table shows the mean weights of the concerned edges resulting from optimization with and without prior knowledge. Furthermore, it shows the percentage of the cases in which the type of interaction was identified correctly

Edge	Prior	Real weight	Mean weight without prior (Correct interaction found %)	Mean weight with prior (Correct interaction found %)
e_{41}	−1	1	0.43 (100)	0.51 (100)
e_{13}	−0.5	0.5	0.24 (100)	0.30 (100)
e_{53}	−0.5	−0.5	−0.19 (40)	−0.60 (95)

We performed 20 independent optimization runs with 200 randomly sampled parameters each. To test for consistency we clustered the 20 best fitting (with minimum error) networks according to their edge weights using hierarchical clustering (data not shown). The networks split up into two clusters. The mean edge weights of the clustered networks were then used to generate two consensus networks. The resulting network representations of both clusters are shown in Fig. 14.1c and d. One network equals the original concerning the topology, whereas the other only differs in one edge (e_{53} replaced by e_{43}). The initial extended feedforward loop is now replaced by a feedback loop, but still maintains the delayed inhibition of v_3. Interestingly, the increased weight of the inhibiting edge in both consensus networks is relatively balanced by decreased weights within the edges e_{34} and e_{41}. In all cases, the unified time constant τ of the models was between 1.95 and 2.05.

3.2 Including Prior Knowledge to the Toy Model Increases Accuracy

To test our toy model towards its sensitivity to incorrect prior information, we constructed a prior knowledge matrix comprising two false and one correct edge.

Again, after 20 optimization runs, the results were analyzed. The resulting networks were widely similar to the previous results except for the edges included in the prior. Table 14.1 shows how the average weights of the edges changed by the inclusion of prior knowledge. As one can see in the table, wrong edges in the prior did not affect the results negatively, whereas the inclusion of the edge e_{53} led to improved results: The percentage of runs resulting in the correct cluster (with e_{53} instead of e_{43}) increased from 40% to 95%. This shows clearly that correct prior knowledge is useful to "guide" the GA to local optima and, on the other hand, that wrong prior knowledge does not affect the quality of our results strongly. Furthermore, for multiple solutions with equal scores, prior knowledge supports the GA in choosing the correct one. However, we should keep in mind that the test was performed using a toy model.

3.3 *In Comparison to a Discrete Boolean Modeling Approach, Semi-quantitative Modeling Reveals Fewer and More Accurate Results*

We further tested in our toy model approach, whether a discrete boolean approach will generate comparable qualitative results. Therefore, we set the possible edge weights in the networks to $\{-1, 0, +1\}$, i.e., $E \in \{-1, 0, +1\}^{n \times n}$ and the concentrations of simulated data was set to discrete values in $\{0, 1\}$. To simulate with a generic boolean function we set $\tau = 1$ for all genes.

After 20 optimization runs, eight different solutions of the same quality were obtained (see Fig. 14.1e–l). These networks all produce the same sequence of states (simulated expression profiles) and all solutions share the same number of edges. In all cases, the inhibiting edge e_{53} remains undetected. Compared to our previous results this shows that in several cases gene regulatory interactions might not be detected by a generic boolean approach. Furthermore, the ability of covering the original dynamics using continuous models allows here for a more accurate reconstruction: Due to the continuous decrease of concentration over time, the inhibitory effect of v_5 on v_3 can be recognized, since the concentration of v_3 decreases slower than the concentration of v_2. As a consequence, the concentrations of v_4 and v_5 will also differ. Thus, it could be recognized that v_4 is an activator of v_1 whereas it is also inhibited directly or indirectly by v_5.

To substantiate the claim that our method outperforms boolean approaches, as observed and explained for the toy model chosen above, we also applied both methods to three different artificial datasets generated with the *GeneNetWeaver 3* simulator (http://sourceforge.net/projects/gnw/) used for the DREAM challenges [28]. We chose three in silico networks, each comprising ten genes. For each network, we used a unified randomly generated prior. Using both methods, we performed 100 optimization runs for each network. We analyzed the results by selecting all edges occurring in more than 30% of the obtained network ensembles. Results were summarized in Table 14.2. Although networks generated with the boolean method appear to be more sensitive, semi-quantitative network reconstruction performed better according to specificity and accuracy. In case of the boolean method, the higher sensitivity is based on a remarkably higher network density. This finding also confirms that compared to boolean modeling our semi-quantitative approach predicts fewer but more reliable interactions.

3.4 *Inferring a Model for Gastrulation from Gene Expression Data Reveals New Regulatory Interactions*

Next, we applied our method to the differentiation dataset. Initially we defined the set of genes that form our model from literature. Additionally to the known pluripoteny genes, we mainly selected genes known to be involved in gastrulation

Table 14.2 Results of the comparison based on the GeneNetWeaver generated data. Table lists the original network edge number and the density, true positives (TP), false positives (FP), false classified (FC), and false negatives (FN). Based on these numbers we computed a recovery ratio, the sensitivity, specificity, and the accuracy for both methods and all networks

		Semi-quantitative modeling									Boolean modeling								
Net	#Edges	Density	TP	FP	FC	FN	$\frac{TP}{FP+FC}$	Sens.	Spec.	Acc.	Density	TP	FP	FC	FN	$\frac{TP}{FP+FC}$	Sens.	Spec.	Acc.
No. 1	15	0.27	6	20	1	8	0.29	0.43	0.76	0.71	0.69	11	56	2	2	0.19	0.85	0.34	0.4
No. 2	16	0.26	7	19	0	9	0.37	0.44	0.77	0.72	0.74	9	59	6	1	0.14	0.9	0.3	0.34
No. 3	11	0.2	5	19	1	5	0.25	0.5	0.79	0.75	0.34	7	27	0	4	0.26	0.63	0.69	0.69

[29, 30]. Genes that were not expressed in the dataset were excluded from further modeling. Prior knowledge was extracted from literature using the text mining tool Bibliosphere from the Genomatix software suite (*www.genomatix.de*). For each of the three cell lines we performed an independent optimization approach with a population size of 20,000 individuals each. We tried different population sizes but found that bigger populations do not change results significantly. The resulting populations were used to generate consensus networks. From each cell line specific network ensemble only edges with absolute weights $e_{i,j} > 0.2$, occurring in more than 30% of the networks were used. The results were displayed as an adjacency matrix in Fig. 14.2a–c, where each row i denotes the effect that gene i has on its target genes (columns).

When comparing the three networks, all three networks could confirm at least half of the 21 known interactions (see Fig. 14.2d), with a total overlap of 6 edges. From the 46 potential edges we found an overlap of 3 edges. Here, one has to keep in mind that we initialized potential edges in both directions since, from the prior, we only included an undirected interaction. We found no overlap concerning the rejected known edges. Regarding new edges – edges not present in the prior – we could identify one edge common to all three consensus graphs with respect to weight and direction: Oct4 activates the expression of the transcription factor gene *Foxh1*, which mediates Nodal signaling during anterior–posterior patterning and node formation in the early mouse embryo [31]. Furthermore, it could be experimentally confirmed that Oct4 binds to the *Foxh1* promoter [32]. We assume that the reason for the deviation of the fitted networks mainly arise from biological divergence, e.g., all three cell lines differentiate with varying speed. A further reason is that with the proceeding differentiation process the heterogeneity of cell types increases. Since bulk RNA was measured this heterogeneity is reflected in the expression profiles.

3.5 *Among the Ensemble of Gastrulation Networks Sparse Networks Appear to be More Robust Against Noise*

GRNs are commonly assumed to be robust against noise, since robustness is beneficial regarding evolvability and selection [33]. We use this fundamental property to analyze our solutions for biological relevance. Therefore we add random noise to the edge weights of the different networks, simulate it, and then compare the resulting expression data with the original expression data.

We performed ten iterations on all networks optimized with all three cell lines separately. In each iteration, we added noise to all edge weights and performed the simulation. The noise was generated using $\mathcal{N}\left(0, \frac{k}{10}\right)$ $(k \in \{1, 2, \ldots, 10\})$. This was performed 200 times for each network in each iteration. Finally, we computed the mean value for each noise level for all three cell lines and identified the best scoring network, respectively (see Fig. 14.3a). Interestingly, the networks trained

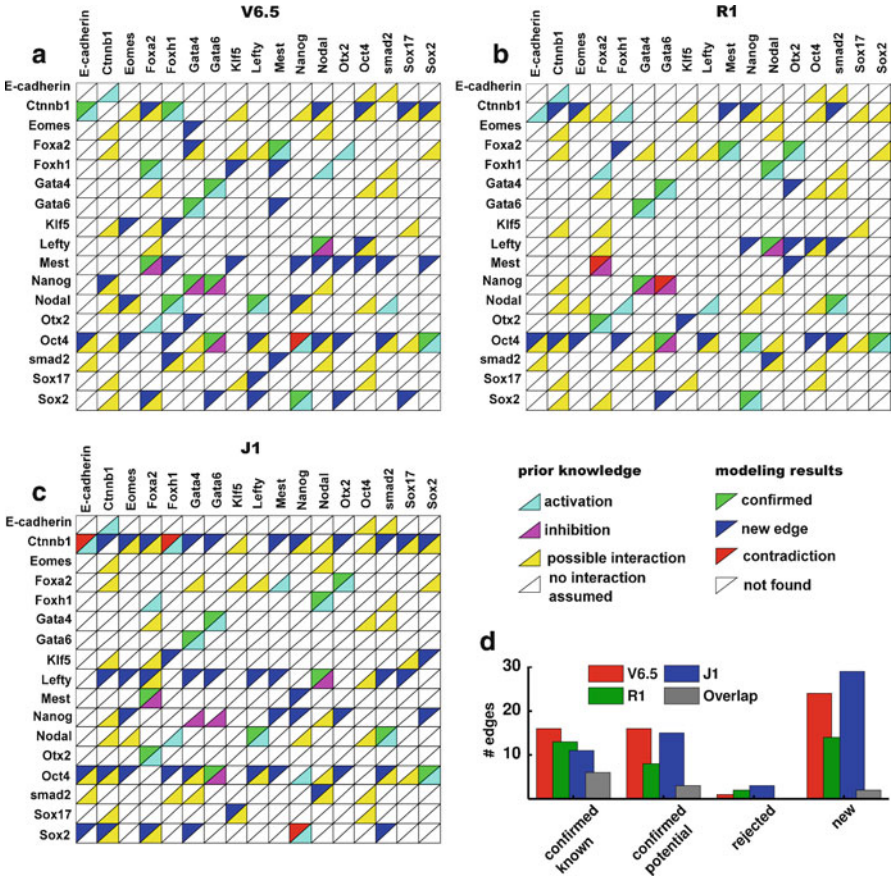


Fig. 14.2 Comparison of consensus networks. (**a**–**c**) Adjacency matrices of the three consensus networks, one for each cell line. Each row *i* denotes the effect that gene *i* has on its target genes (columns). The triangles denote for prior knowledge (lower, right) and for the modeling results (upper, left). (**d**) The barplot summarizes the number of edges found or rejected in all three consensus networks and the overlap. In particular the number of confirmed known prior edges, the number of confirmed potential edges, the number of rejected known edges and the number of new identified edges

with the cell line R1, that were overall more sparse compared to the others (see Fig. 14.2b), perform most robust. This finding, that sparseness is associated with robustness, agrees with the work of Leclerc that shows that biological networks are parsimonious [34]. The network that performs best when adding noise is shown in Fig. 14.3c. Its connectivity of 34 is significantly lower when compared to the mean of all trained networks of 64 and a standard deviation of 9.9. Within the network two known edges could be confirmed, 12 known edges are missing and two were rejected (opposite direction). Six possible and 19 new edges are included.

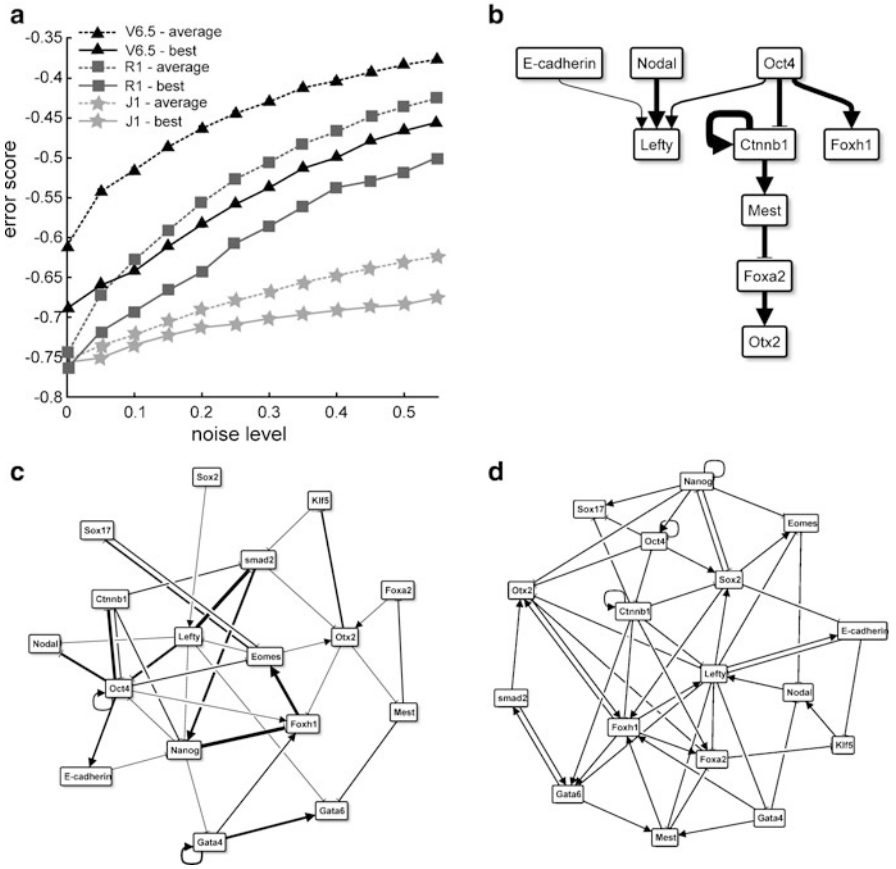


Fig. 14.3 Robustness analysis. To test for robustness, we added increasing noise to each optimized network and computed the error score combined of correlation and pairwise deviation (see text). (**a**) For each cell line we show the mean error of all networks and the best network, respectively. (**b**) Network consisting of all common edges of the three most robust networks against noise. The thickness of edges is relative to the mean edge weight of the three best networks. (**c**) Most robust network against increasing noise. Edge thickness refers to edge weights. (**d**) Most stable network after cross validation

We then compared the three best networks and selected all common edges. The resulting graph is shown in Fig. 14.3b. The network shows a clear hierarchical structure with Oct4, Nodal, and E-cadherin on top. Again these networks confirm the activation of *Foxh1* by Oct4.

To test for biological consistency we performed a cross validation analysis. From all cell line specific networks the five best performing were validated against the two other cell lines, respectively. The network with the best cross-validation score is shown in Fig. 14.3c and was originally trained on the J1 cell line. Interestingly, its connectivity of 51 is also relatively low when compared to the average of 64.

This again agrees with our previous finding that sparse networks perform more robustly. Taken together the network confirms three, rejects four, and lacks 14 of the known interactions. It further includes 8 possible interactions and 36 new edges.

This strong deviation of the robust networks probably arises from the large heterogeneity in the data. Since the cell lines differentiate in a diverse manner, robustness of the networks reduces to the minimal accordance in the expression profiles of the genes. Furthermore, the lack of most interactions of the core pluripotency network can be explained by the fact that the networks were trained on differentiating data, where these interactions were mostly displaced by mechanisms driving the differentiation. Taken together, heterogeneity here leads to sparseness that in turn helps to identify core regulatory interactions.

Acknowledgements We kindly thank Dominik Wittmann and Sabine Hug for proofreading the manuscript and many helpful comments. This research was partially supported by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (project CoReNe) and by the European Union within the ERC grant LatentCauses (grant agreement number 259294).

References

1. Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, Haus UU, Weismantel R, Gilles ED, Klamt S, Schraven B (2007) A logical model provides insights into t cell receptor signaling. PLoS Comput Biol 3:e163
2. Busch H, Camacho-Trullio D, Rogon Z, Breuhahn K, Angel P, Eils R, Szabowski A (2008) Gene network dynamics controlling keratinocyte migration. Mol Syst Biol 4:199
3. Lu R, Markowetz F, Unwin RD, Leek JT, Airoldi EM, MacArthur BD, Lachmann A, Rozov R, Ma'ayan A, Boyer LA, Troyanskaya OG, Whetton AD, Lemischka IR (2009) Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature 462:358–362
4. Lutter D, Langmann T, Ugocsai P, Moehle C, Seibold E, Splettstoesser WD, Gruber P, Lang EW, Schmitz G (2009) Analyzing time-dependent microarray data using independent component analysis derived expression modes from human macrophages infected with *F. tularensis* var. *holartica*. J Biomed Inform 42:605–611
5. Glauche I, Herberg M, Roeder I (2010) Nanog variability and pluripotency regulation of embryonic stem cells – Insights from a mathematical model analysis. PLoS One 5:e11238
6. Cara AD, Garg A, Micheli GD, Xenarios I, Mendoza L (2007) Dynamic simulation of regulatory networks using squad. BMC Bioinformatics 8:462
7. Glass L, Kauffman S (1973) The logical analysis of continuous, non-linear biochemical control networks. J Theor Biol 39:103–129
8. Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, Klamt S, Sorger PK (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Mol Syst Biol 5:331
9. Davidich MI, Bornholdt S (2008) Boolean network model predicts cell cycle sequence of fission yeast. PLoS One 3:e1672
10. Fauré A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics 22:e124–e131
11. Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ (2009) Transforming boolean models to continuous models: methodology and application to t-cell receptor signaling. BMC Syst Biol 3:98

12. Krumsiek J, Pölsterl S, Wittmann DM, Theis FJ (2010) Odefy–from discrete to continuous models. BMC Bioinformatics 11:233
13. Bornholdt S (2008) Boolean network models of cellular regulation: prospects and limitations. J R Soc Interface 5 (Suppl 1):S85–S94
14. Glass L (1975) Classification of biological networks by their qualitative dynamics. J Theor Biol 54:85–107
15. Casey R, de Jong H, Gouz JL (2006) Piecewise-linear models of genetic regulatory networks: equilibria and their stability. J Math Biol 52:27–56
16. de Jong H, Page M (2008) Search for steady states of piecewise-linear differential equation models of genetic regulatory networks. IEEE/ACM Trans Comput Biol Bioinform 5:208–222
17. Tam PPL, Loebel DAF (2007) Gene function in mouse embryogenesis: get set for gastrulation. Nat Rev Genet 8:368–381
18. Costa IG, Roepcke S, Hafemeister C, Schliep A (2008) Inferring differentiation pathways from gene expression. Bioinformatics 24:i156–i164
19. Niwa H (2007) How is pluripotency determined and maintained? Development 134:635–646
20. Chickarmane V, Peterson C (2008) A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. PLoS One 3:e3478
21. Burtscher I, Lickert H (2009) Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. Development 136:1029–1038
22. Tamplin OJ, Kinzel D, Cox BJ, Bell CE, Rossant J, Lickert H (2008) Microarray analysis of foxa2 mutant mouse embryos reveals novel gene expression and inductive roles for the gastrula organizer and its derivatives. BMC Genomics 9:511
23. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov AA, Ko MSH, Niwa H (2007) Pluripotency governed by sox2 via regulation of oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9:625–635
24. Chambers I, Tomlinson SR (2009) The transcriptional foundation of pluripotency. Development 136:2311–2322
25. Sene KH, Porter CJ, Palidwor G, Perez-Iratxeta C, Muro EM, Campbell PA, Rudnicki MA, Andrade-Navarro MA (2007) Gene function in early mouse embryonic stem cell differentiation. BMC Genomics 8:85
26. Wilson CL, Miller CJ (2005) Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. Bioinformatics 21:3683–3685
27. Aij T, Lhdesmki H (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics 25:2937–2944
28. Greenfield A, Madar A, Ostrer H, Bonneau R (2010) Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. PLoS One 5:e13397
29. Tam PPL, Loebel DAF, Tanaka SS (2006) Building the mouse gastrula: signals, asymmetry and lineages. Curr Opin Genet Dev 16:419–425
30. Pfister S, Steiner KA, Tam PPL (2007) Gene expression pattern and progression of embryogenesis in the immediate post-implantation period of mouse development. Gene Expr Patterns 7:558–573
31. Yamamoto M, Meno C, Sakai Y, Shiratori H, Mochida K, Ikawa Y, Saijoh Y, Hamada H (2001) The transcription factor foxh1 (fast) mediates nodal signaling during anterior–posterior patterning and node formation in the mouse. Genes Dev 15:1242–1256
32. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CWH, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH (2006) The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38:431–440
33. Kitano H (2004) Biological robustness. Nat Rev Genet 5:826–837
34. Leclerc RD (2008) Survival of the sparsest: robust gene networks are parsimonious. Mol Syst Biol 4:213

Chapter 15
Cell Death and Life in Cancer: Mathematical Modeling of Cell Fate Decisions

Andrei Zinovyev, Simon Fourquet, Laurent Tournier, Laurence Calzone, and Emmanuel Barillot

Abstract Tumor development is characterized by a compromised balance between cell life and death decision mechanisms, which are tightly regulated in normal cells. Understanding this process provides insights for developing new treatments for fighting with cancer. We present a study of a mathematical model describing cellular choice between survival and two alternative cell death modalities: apoptosis and necrosis. The model is implemented in discrete modeling formalism and allows to predict probabilities of having a particular cellular phenotype in response to engagement of cell death receptors. Using an original parameter sensitivity analysis developed for discrete dynamic systems, we determine variables that appear to be critical in the cellular fate decision and discuss how they are exploited by existing cancer therapies.

1 Introduction

Evading various programmed cell death modalities is considered as one of the major hallmarks of cancer cells [1]. A better understanding of the pro-death or pro-survival roles of the genes associated with various cancers, and their interactions with other pathways would set a ground for reestablishing a lost death phenotype and identifying potential drug targets.

A. Zinovyev (✉) • S. Fourquet • L. Calzone • E. Barillot
U900 INSERM/Institut Curie/Ecole de Mines, Institut Curie, 26 rue d'Ulm, Paris 75005, France
e-mail: andrei.zinovyev@curie.fr; simon.fourquet@curie.fr; laurence.calzone@curie.fr; emmanuel.barillot@curie.fr

L. Tournier
INRA, Unit MIG (Mathématiques, Informatique et Génome), Domaine Vilvert, Jouy en Josas 78350, France
e-mail: laurent.tournier@jouy.inra.fr

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_15,

Recent progress in studying the mechanisms of cell life/death decisions revealed its astounding complexity. Among many, one can mention three difficulties on the way to characterize, describe, and create strict mathematical descriptions of these mechanisms.

First, the signaling network allowing a cell to react to an external stress (such as damage of DNA, nutrient and oxygen deprivation, toxic environment) is assembled from highly redundant pathways which are able to compensate each other in one way or another. For example, there exist at least seven distinct and parallel survival pathways associated with action of AKT protein [2]. Disruption of one of these pathways in a potential cell death-inducing cancer therapy can be in principle compensated by the others. Thus, understanding and modeling the survival response in its full complexity is a daunting task.

Second, cellular death is an extremely complex phenotype that cannot merely be described as a simple disaggregation of cellular components driven by purely thermodynamical laws. Several distinct modes of cell death were identified in the last decade [3], such as, necrosis, apoptosis, and autophagy. Importantly, all these cell death modalities are controlled by cellular biochemical mechanisms, activated in response to diverse types of stress: roughly speaking, a cell is usually preprogrammed to die in a certain manner, sending appropriate signals to its surroundings so as to limit tissue toxicity and allow recycling of its components. *Necrosis* is a type of cell death usually associated with a lack of important cellular resource such as ATP, which makes functioning of many biochemical pathways impossible. This is why it was long thought of as an uncontrolled and purely thermodynamics-driven degradation of cellular structures. However, recent research showed that necrosis can be triggered by specific signals through the activation of tightly regulated pathways, and can even proceed without ATP depletion [3]. By contrast, *apoptosis* as a form of cellular suicide was, from the very beginning, described as a mode of cell death requiring energy for the permeabilization of mitochondrial membranes and cleavage of intracellular structures. *Autophagy* remains a relatively poorly understood cell death mechanism, which seems to serve both as a survival or a death modality. Upon certain stress conditions, and until this stress is relieved, cellular components such as damaged proteins or organelles are digested and recycled into reusable metabolites, and metabolism is reoriented so as to spare vital functions. Long lasting, non-relievable stress was described as triggering autophagic cell death, through unaffordable cellular self-digestion. However, no experimental evidence ever unambiguously demonstrated that such cell death is directly executed by autophagy in vivo, but this is seen in the special case of the involution of *Drosophila melanogaster* salivary glands [3].

The third difficulty can be attributed not directly to the complexity of the biochemical mechanisms but rather to our capabilities of apprehending the design principles used by biological evolution. Inspired by engineering practices, we tend to investigate complex systems by splitting them into relatively independent modules and associating well-characterized non-overlapping functions to each molecular detail. Applying such reductionist approaches to biology comes with

a caveat. Most cellular molecular machineries cannot be naturally dissected or associated with well-defined functions, and sets of overlapping functions can be distributed among groups of molecular players.

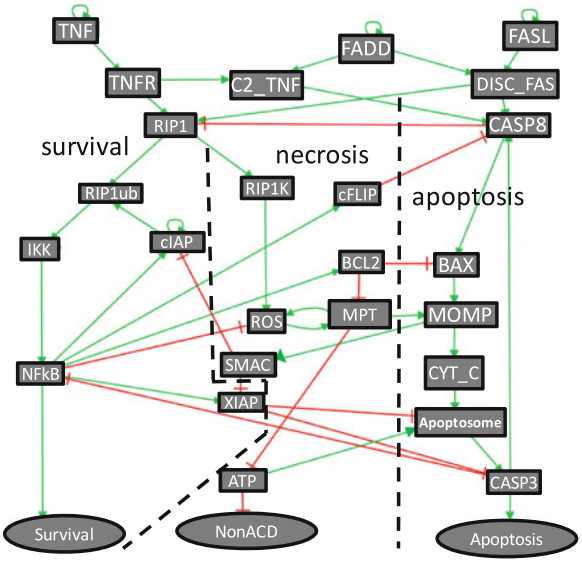
Not having the ambition to deal with the whole complexity of cell fate decisions in vivo, we decided to concentrate on modeling the outcome of a classical and rather well-defined experiment of inducing cell death: adding to a cell culture specific ligands (Tumor Necrosis Factor, TNF, or other members of its family such as FASL). These so-called death ligands can engage death receptors and trigger apoptosis or necrosis, or activate pro-survival mechanisms [5]. The net outcome of such experiments depends on many circumstances: cell type, dose of the ligand, duration of the treatment, specific mutations in cell genomes, etc. Moreover, it is believed that the outcome can have intrinsic stochastic nature governed by cellular decision making mechanisms and intrinsic molecular noise [6]. Trying to characterize the biochemical response of a cell to this relatively simple kind of perturbation allows to understand certain cell fate decision mechanisms.

In this paper, we briefly describe and carefully analyze a mathematical model of cell fate decision between survival and two alternative modes of cell death: apoptosis and necrosis. The model was created and introduced in [4]. Here we propose the principles for wiring and parametrizing a biological diagram that describes this cellular switch. In addition to [4], here, by applying a novel sensitivity analysis specifically developed for discrete modeling, we identify fragile sites of the cell fate decision mechanism. In conclusion, we compare our analysis with our current knowledge of cellular decision making fragilities utilized by cancer and cancer therapies.

2 Mathematical Model of Cell Fate Decision

In [4] we summarized the current knowledge on the interactions between cell fate decision mechanisms in a simplistic wiring diagram (see Fig. 15.1) where a node represents either a protein (TNF, FADD, FASL, TNFR, CASP8, cFLIP, BCL2, BAX, IKK, NFκB, CYT_C, SMAC, XIAP, CASP3), a state of protein (RIP1ub, RIP1K), a small molecule (ROS, ATP), a molecular complex (Apoptosome, C2_TNF, DISC_FAS), a group of molecular entities sharing the same function (BAX can thus represent either of BAX and BAK, cIAP either cIAP1 or cIAP2, and BCL2 any of the BH1–4 BCL2 family members, etc.) a molecular process (Mitochondria permeabilization transition, MPT, Mitochondrial outer membrane permeabilization, MOMP) or a phenotype (Survival, Apoptosis, Non-apoptotic cell death, and Non-ACD). Each directed and signed edge represents an influence of one molecular entity on another, either positive (arrowed edge) or negative (headed edge).

The phenotype nodes on the diagram are simple interpretations of the following molecular conditions: (a) activated NFκB is read as survival state; (b) lack of ATP



DISC_TNF' = TNFR AND FADD
RIP1' = (TNFR OR DISC_FAS) AND (NOT CASP8)
CASP8' = (DISC_TNF OR DISC_FAS OR CASP3) AND (NOT cFlip)
RIPub' = RIP1 AND cIAP
cIAP' = (NFkB OR cIAP) AND (NOT SMAC)
BAX' = CASP8 AND (NOT BCL2)
ROS' = (RIP1K OR MPT) AND (NOT NFkB)
MPT' = ROS AND (NOT BCL2)
MOMP' = BAX OR MPT
NFkB' = IKK AND (NOT CASP3)
XIAP' = NFkB AND (NOT SMAC)
Apoptosome' = CYT_C AND ATP AND (NOT XIAP)
CASP3' = Apoptosome AND (NOT XIAP)

TNF' = TNF
FADD' = FADD
FAS' = FAS
TNFR' = TNF
RIP1K' = RIP1
cFlip' = NFkB
IKK' = RIP1ub
BCL2' = NFkB
SMAC' = MOMP
CYT_C' = MOMP
DISC_FAS' = FAS AND FADD
ATP' = NOT MPT

Fig. 15.1 Biological diagram of molecular interactions involved in cell fate decisions derived from the biological literature. The diagram is roughly divided by dashed lines into three modules corresponding to three submechanisms of cell fate decisions. Notations: (1) Proteins: TNF, FADD, FASL, TNFR, CASP8, cIAP, cFLIP, BCL2, BAX, IKK, NFκB, CYT_C, SMAC, XIAP, CASP3; (2) States of proteins: RIP1ub (ubiquitinated form of RIP1), RIP1K (kinase function of RIP1); (3) Small molecules: ATP, ROS (Reactive oxygen species); (4) Molecular complexes: Apoptosome, C2_TNF, DISC_FAS; (5) Molecular processes: MPT (Mitochondria permeabilization transition), MOMP (Mitochondrial outer membrane permeabilization); (6) Phenotypes: Survival, Apoptosis, and Non-ACD (Non-apoptotic cell death). Below the table of logical rules defining the discrete mathematical model is provided

is read as non-apoptotic cell death state; (c) activated CASP3 is read as apoptotic cell death. Absence of any of such conditions is interpreted as a "naive" cell state, corresponding to the fourth cellular phenotype.

After extensive examination of the biological literature we converted the diagram into a logical mathematical model of cell fate decisions triggered by activation of cell death receptors. The wiring diagram and the logical rules defining the model are shown in Fig. 15.1.

By applying a technique adapted to discrete formalism [7], we reduced this model to a 11-dimensional network, thus enabling a complete analysis of the asynchronous dynamics (see [4] for details). This analysis identified 27 stable logical states and no cyclic attractors. Moreover, it showed that the distribution of the stable logical states in the discrete 22-dimensional space of internal model variables (without considering input and output variables) forms four compact clusters, each corresponding to a particular cellular phenotype. Three of these clusters can be attributed to a particular cell fate (survival, apoptosis, necrosis) while the forth represents a "naive" survival state, where no death receptors are induced.

3 Computing Phenotype Probabilities

As we have already mentioned, the cellular fate decision machinery is characterized by stochastic response, i.e., given a stimuli, the cell can reach several final states, corresponding to different phenotypes, with different probabilities. The role of mathematical modeling in this case could predict these probabilities as absolute values that can be matched to an experiment, or at least can predict the relative changes of the probabilities after introducing some perturbations to the system.

We have implemented this idea for the mathematical model of cell fate decisions described above in the following manner.

In order to describe our results, let us introduce the notion of asynchronous state transition graph. On this graph, each node represents a state of the system which in this case can be encoded by a n-dimensional vector of 0s and 1s (n being the dimension of the system). A directed edge exists between two states x and y if there exists an index $i \in \{1, \ldots, n\}$ such that $y_i = f_i(x) \neq x_i$ and $y_j = x_j$ for $j \neq i$ (here, f_i denotes the logical rule of variable x_i, see Fig. 15.1 for a complete list of the model logical rules). In principle, the state transition graph could be defined independently and without the biological diagram, however, this would require a tremendous amount of empirical knowledge about the set of all permissible transitions between the cell states which is not available. Hence, the biological diagram with associated logical rules is used as a compact representation and a tool to generate the state transition graph. Detailed instructions on this procedure can be found in [8, 9].

The set of all possible states provides a discrete phase space of the system. The state transition graph contains all possible ways of the systems dynamics (trajectories). In other words, it is the *multidimensional epigenetic landscape* of the cell fate decision system. Note that the state transition graph is assumed to be rather sparse compared to the fully connected graph where any two state transitions would be possible. Hence, on this landscape, one can determine bifurcating states, points of no return, etc.

The state transition graph allows to address the following question: *Starting from a distinguished state of a cell, what is the probability to arrive to each of the stable states*? In biological terms: *Which proportions of a population of resting cells exposed to death ligand will eventually display each of the different phenotypes – cell fate*?

To answer the question, we converted the state transition graph into a Markov process of random walk on a graph, following the method described in [9]. To do that, we associated to each transition between two states a probability (called transition probability). By applying classical algorithms to the transition probability matrix (strongly connected decomposition and topological sort), we obtained an *absorbing discrete Markov chain*, and then analyzed it with classical techniques [10].

One of the critical points in such type of analysis lies in the choice of the transition probabilities. Once again, defining these probabilities directly from some empirical observations is impossible at present time. Hence, these probabilities should be derived from the logical model with the use of some additional assumptions.

The simplest assumption is to consider all transitions firing from a given state as *equiprobable*. Biological interpretation of such an assumption is not simple. In a way, we consider a "generic" cell in which all possible system trajectories take place with equal probabilities (without dominance, i.e., any preferable route). One can argue that in any particular concrete cell, this would not be true anymore and that the generic cell is not representative of anything real observed in any biological experiment. Having in mind this difficulty, we avoid direct interpretation of absolute values of probabilities, concentrating rather on relative changes of them in response to some system modifications such as removing a node or fixing a node's activity. It happens that such a "generic" cell model is already capable of reproducing a number of known experimental facts.

When the state transition graph is parametrized by transition probabilities, one can use standard techniques to compute the probability of hitting a given stable state, considering that a random walk starts from a given initial state. Then this probability is associated with a probability of observing a particular phenotype in given experimental conditions. For doing this, it is convenient to define a unique initial state, which we choose to represent the "physiological state", the one representing un-induced cells growing in a plate. Model in Fig. 15.1 is the state in which all elements are inactive except ATP, FADD, and cIAP. This is a stable state, which looses its stability when TNF variable is changed from 0 to 1 and the dynamical system starts to evolve in time.

Using this approach, we performed a series of in silico experiments in which the probability of arriving to stable states was computed for the initial ("wild-type") model, or for a series of modified ("mutant") model. Typical model modifications consisted in fixing some nodes' activities to 0 or to 1. For our cell fate decision model, the results are provided in Fig. 15.2. In [4] this table was systematically compared with the experimental data of the cell death phenotype modifications observed in various mutant experimental systems, including cell cultures and mice. The model was able to qualitatively recapitulate all of them and to suggest some new yet unexplored experimentally mutant phenotypes. The most interesting in this setting would be to consider synthetic interactions between individual mutants, when several nodes on the diagram are affected by a mutation simultaneously.

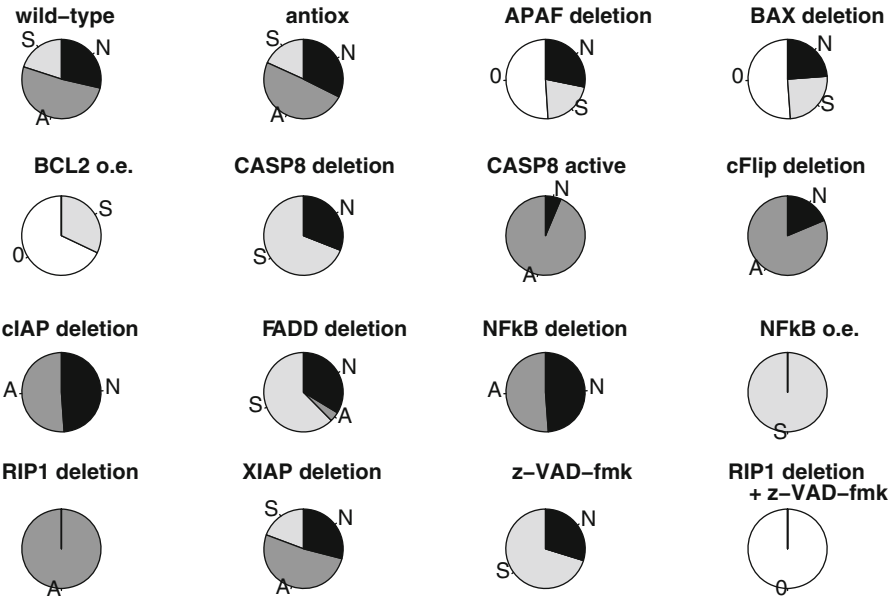


Fig. 15.2 Changes in the phenotype probabilities from the random walk on the state transition graph, starting from the initial physiological state. Various "mutant" modifications of the dynamical system are tested here. Here "A" denotes Apoptosis, "N" denotes Necrosis, and "S" denotes Survival, "0" denotes Naive state. "O.e." stands for overexpression of a protein, "antiox" corresponds to blunting the capacity of NFκB to prevent ROS formation, "z-VAD fmk" simulates the effect of caspase inhibitor z-VAD-fmk

4 Identification of Fragile Points of the Cell Fate Decision Machinery

Changing distribution of transition probabilities on the asynchronous state transition graph can drastically change the probabilistic outcome of a computational experiment. At the same time, the probabilities for a random walk to converge to some attractor depend also on the structure of the state transition graph which is determined solely from the discrete model. In order to understand what are the critical determinants of a cellular choice, we applied a novel strategy of discrete model analysis consisting in parametrizing the state transition graph by changing relative importance of certain variables. In a certain sense, this strategy corresponds to a sensitivity analysis, commonly applied for continuous models based on an ordinary differential equations and chemical kinetics approach [11].

First of all, we postulate that our "reference" parametrization corresponds to the equal probabilities of any possible transition from a state. As mentioned earlier, this corresponds to a "generic" cell model, where the relative speeds of all biochemical processes are assumed equal. Mathematically, considering the

dynamics as a Markov process, all transitions from a given state x to any of its asynchronous successor are assigned equal probabilities (if x has r successors, these probabilities are equal to $1/r$). We will modify this default parametrization by systematically changing relative speeds of certain elements. This will lead to some reparametrization of the state transition graph and consequent changes in the probabilities to reach attractors.

The key idea of priority classes [12, 13] consists in grouping variables of a discrete model into classes according to the speeds of the underlying processes governing their turnover rates. For instance, in the case of genetic regulatory networks, a natural grouping consists in putting de novo protein synthesis (transcription + translation) in a slow transition class in comparison with other processes such as post-translational protein modifications (phosphorylation, ubiquitination, etc.) or complex formation. Following this idea, we can regroup nodes into priority classes to which some priority ratios w are assigned. As it is said differently, each variable x_i is assigned to a priority value w_i. For a given node, a value $w_i > 1$ corresponds to a higher than default priority, and a value $w_i < 1$ to a lower than default priority. The ratio w_i can be interpreted as a global turnover rate of the component represented by this node: those that are produced (activated) and degraded (deactivated) fast will have a large w_i.

Consider a state x, with r asynchronous successors. By definition, between x and each of its successors, one and only one variable can be updated. Let y denote one of the successors of x, and i be the index of the corresponding updated variable. With the uniform assumption described before, the probability of the transition $(x \to y)$ is independent of i and is equal to $1/r$. With priority classes, this probability is now weighted by w_i, making the transition more probable if component i belongs to a "fast" class (w_i greater than one) and less probable if it belongs to a "slow" class (w_i less than one). Obviously, for computing the actual transition probabilities $p_{x\to y}$, a normalization should be applied so that:

$$\sum_{y \text{ succ. of } x} p_{x\to y} = 1.$$

Once the new values of the transition probabilities have been computed, the same treatments as before can be applied, leading to new values for the probabilities to reach the different phenotypes, starting from a given initial condition.

This general method may be applied in two different ways. First, one may use it to compute more realistic probabilities, that could be compared to actual experimental results (the probability to reach an attractor being compared with the proportion of cells exhibiting the corresponding phenotype). However, such calculations would need a complete classification of the relative speeds of all biochemical mechanisms involved in the model. Given the number and heterogeneity of these mechanisms, it is still difficult to obtain such classification. Instead, we used the method as a sensitivity analysis tool, in order to detect which variables are more critical than others in the decision-making process. Using the reduced model evoked earlier

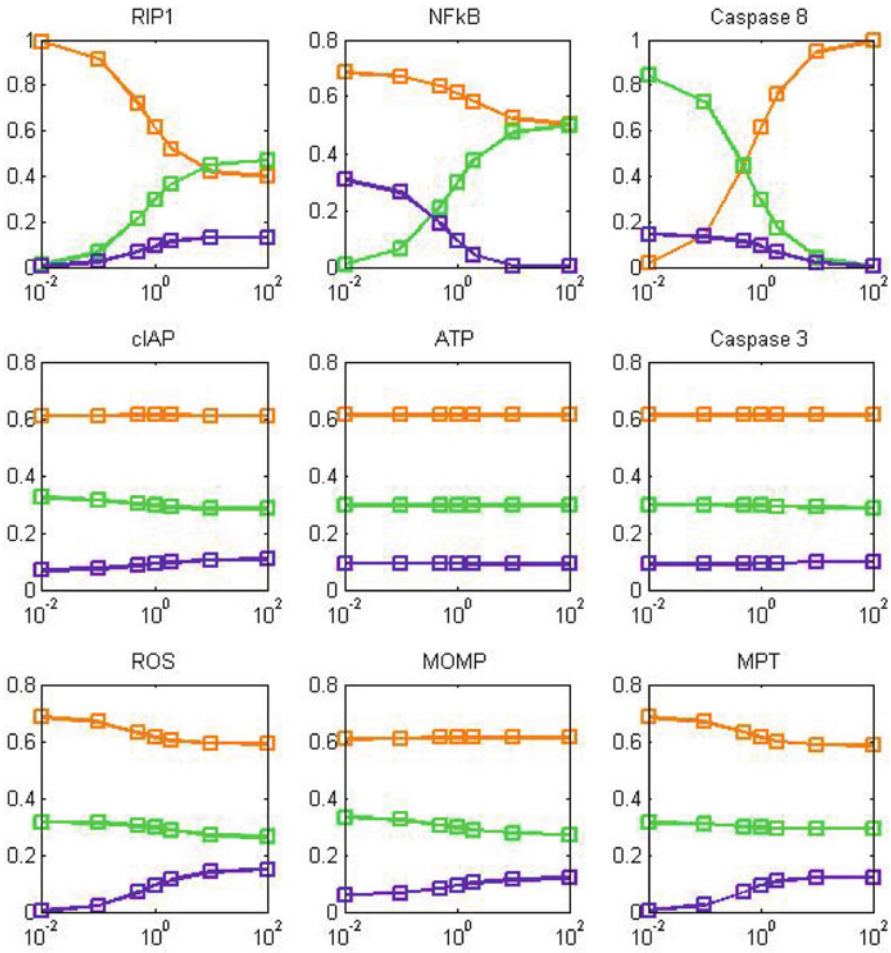


Fig. 15.3 Testing the effect of varying node turnovers on the resulting phenotypic probabilities. The absciss on the graphs shows the value of w priority value, where $w = 1$ corresponds to the probabilities computed for the default wild-type model (see Fig. 15.2). The colors are those adopted in [4]: orange corresponds to apoptosis, purple to necrosis, and green to survival

(see [4]), we considered each variable independently, and successively boosted it or slowed it down by some multiplicative factor. More precisely, to detect the sensitivity of the network with respect to the turnover of variable x_i, we performed the calculations for different values of w_i, the other weights w_j being kept at one (the reference value). By comparing the probabilities to reach the three phenotypes – survival, apoptosis, and necrosis – with those of the initial model, one can detect whether the system's response is sensitive or not to the turnover rate of variable x_i. We performed such experiments for the nine inner variables of the reduced model. Figure 15.3 presents the results we obtained.

The plots reveal several interesting properties. First, the most sensitive components, which correspond to the curves with the highest amplitude, are RIP1, NFkB, and CASP8. This reinforces the idea that these three components play a crucial role in the decision process. This seems reasonable, especially for RIP1 and CASP8, as they occupy an upstream position in the regulatory graph. Interestingly, CASP3 turnover does not seem to be so important, although CASP3 is a marker of apoptosis. This confirms that even though CASP3 is essential for the existence of apoptosis in the model (its removal completely suppress apoptotic outcome, see Fig. 15.2), its turnover rate does not appear to be important in the dynamics of the decision process (once it goes from 0 to 1, most of the decision has already been made). Remarkably, the turnovers of MOMP and MPT, both contributing to the permeabilization of mitochondrial membrane, have different effects: MOMP seems to affect mainly the decision between survival and necrosis, while MPT plays a role in the switch between apoptosis and necrosis.

The sensitivity analysis, that is, presented here is an extension of the results proposed in [4]. In contrast with the all-or-none perturbations evoked in the previous part (where a node is fixed to 0 or 1), here we consider finer perturbations by modifying the turnover rates of the model's variables. A next step would be to consider the relative strengths of the model's interactions, instead of the model's variables. Such an approach is currently investigated.

5 Comparison with the Fragilities Exploited by Cancer and Its Treatment

Deregulations of the signaling pathways studied here can lead to drastic and serious consequences. Hanahan and Weinberg proposed that escape of apoptosis, together with other alterations of cellular physiology, represents a necessary event in cancer promotion and progression [1]. As a result, somatic mutations leading to impaired apoptosis are expected to be associated with cancer. In the cell fate model presented here, most nodes can be classified as pro-apoptotic or anti-apoptotic according to the results of "mutant" model simulations, which are correlated with experimental results found in the literature. Genes classified as pro-apoptotic in our model include caspases-8 and -3, APAF1 as part of the apoptosome complex, cytochrome c (Cyt_c), BAX, and SMAC. Anti-apoptotic genes encompass BCL2, cIAP1/2, XIAP, cFLIP, and different genes involved in the NFkB pathway, including NFKB1, RELA, IKBKG, and IKBKB (not explicit in the model). Genetic alterations leading to loss of activity of pro-apoptotic genes or to increased activity of anti-apoptotic genes have been associated with various cancers. Thus, we can cross-list the alterations of these genes deduced from the model with what is reported in the literature and verify their role and implications in cancer.

For instance, concerning pro-apoptotic genes, frameshift mutations in the ORF of the BAX gene are reported in $>50\%$ of colorectal tumors of the micro-satellite mutator phenotype [14]. Expression of CASP8 is reduced in $\approx 24\%$ of tumors from

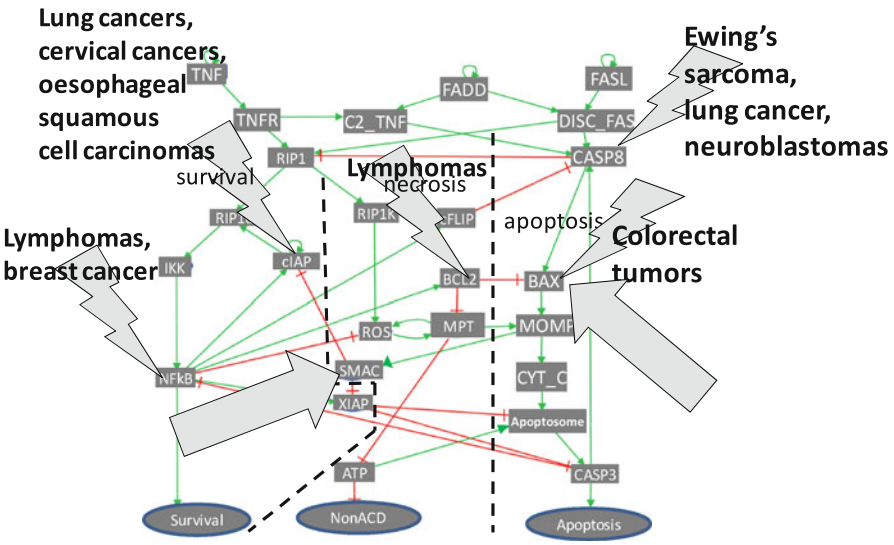


Fig. 15.4 Cell fate decision fragilities identified in various cancers. Flash arrows, hitting from left to right, represent overexpression or amplification, those hitting from right to left show deletion and down-regulation. Rectangular arrows point to components targeted by cancer treatment strategies (SMAC and BAX mimetics)

patients with Ewing's sarcoma [15]. Caspase-8 was suggested in several studies to function as a tumor suppressor in neuroblastomas [16] and in lung cancer [17] (see Fig. 15.4).

On the other hand, constitutive activation of anti-apoptotic genes is often observed in cancer cells. The most striking example is the over-expression of the BCL2 oncogene in almost all follicular lymphomas, which can result from a t(14;18) translocation that positions BCL2 in close proximity to enhancer elements of the immunoglobulin heavy-chain locus [18]. As for the survival pathway, elevated NFkB activity, resulting from different genetic alterations or expression of the v-rel viral NFkB isoform, is detected in multiple cancers, including lymphomas and breast cancers [19]. An amplification of the genomic region 11q22 that spans over the cIAP1 and cIAP2 genes is associated with lung cancers [20], cervical cancer resistance to radiotherapy [21], and oesophageal squamous cell carcinomas [22] (see Fig. 15.4).

Some of the components of the cell fate decision machinery are considered currently for the use in cancer treatment in preclinical or clinical trials. To give some examples, SMAC mimetics directly target dysregulated, neoplastic cells that overexpress IAPs or underexpress SMAC [23]. BCL-2 inhibitors, most notably BAX mimetics, are currently passing clinical trials (for example, see [24]).

In our sensitivity analysis, the variables NFkB and CASP8 appear among the most "vulnerable" components of the cell fate decision machinery, which could explain why the gene products they represent are fragile points used by cancer.

BLC-2 does not show up as a sensitive node in the model. However, its direct target, MPT is a fragile site, accordingly to our analysis. Also analysis of our model shows that RIP1 is a powerful and sensitive switch able to reverse phenotype probabilities. Until so far we are not aware about possible targeting of RIP1 functions in cancer treatment, which can be explained by still relatively poor characterization of its substrates and difficulties connected with targeting specific RIP1 activities.

6 Conclusion

Mathematical models provide a way to test biological hypotheses in silico. They recapitulate consistent heterogeneous published results and assemble disseminated information into a coherent picture using an appropriate mathematical formalism (discrete, continuous, stochastic, hybrid, etc.), depending on the questions and the available data. Then, modeling consists of constantly challenging the obtained model with available published data or experimental results (mutants or drug treatments, in our case). After several refinement rounds, a model becomes particularly useful when it can provide counter-intuitive insights or suggest novel promising experiments.

Here, we have conceived a mathematical model of cell fate decision, based on a logical formalization of well-characterized molecular interactions. Former mathematical models only considered two cellular fates, apoptosis and cell survival [25]. In contrast, we include a non-apoptotic modality of cell death, mainly necrosis, involving RIP1, ROS, and mitochondria functions.

By analyzing properties of the state asynchronous transition graphs associated with the discrete model, we implemented a procedure to simulate the process of stochastic cellular decision making in response to activation of death receptors. These simulations were able to predict relative changes for probabilities of cellular phenotypes in response to some system perturbations such as a knock-out of a gene or treatment with a drug. These predictions happened to be fully compatible with published data from mouse experiments, and provided new predictions to be tested.

Moreover, on this model we have tested a novel strategy of discrete model analysis, consisting in finding fragile or most sensitive places of the cell fate decision machinery. Changing the cellular parameters determining choices made at these fragile sites affect the probabilities for a cell to reach a particular cellular phenotype. We found out that this type of analysis can explain some of the common fragilities associated with tumorigenesis and also with currently employed cancer treatment strategies.

Acknowledgements We would like to acknowledge support by the APO-SYS EU FP7 project. A. Zinovyev, S. Fourquet, L. Calzone and E. Barillot are members of the team “Systems Biology of Cancer”, Equipe labellisee par la Ligue Nationale Contre le Cancer. L. Tournier is member of the Systems Biology team in the laboratory MIG of INRA (French Institute for Agronomical Research). The study was also funded by the Projet Incitatif Collaboratif “Bioinformatics and Biostatistics of Cancer” at Institut Curie.

References

1. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell;144(5): 646–674.
2. McCormick F (2004) Cancer: survival pathways meet their end. Nature 428(6980):267–269.
3. Kroemer G., et al. (2008) Classification of cell death: recommendations of the Nomenclature Committee on Cell Death 2009. Cell Death Differ 16(1):3–11.
4. Calzone L, Tournier L, Fourquet S, Thieffry D, Zhivotovsky B, Barillot E, Zinovyev A. (2010) Mathematical modelling of cell-fate decision in response to death receptor engagement. PLoS Comput Biol 6(3):e1000702.
5. Van Herreweghe F, Festjens N, Declercq W, Vandenabeele P (2010) Tumor necrosis factor-mediated cell death: to break or to burst, that's the question. Cell Mol Life Sci 67(10): 1567–1579.
6. Balazsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. Cell 144(6):910–925.
7. Naldi A, Remy E, Thieffry D, Chaouiya C (2009) A reduction method for logical regulatory graphs preserving essential dynamical properties. Lecture Notes in Computer Science 5688:266–280.
8. Chaouiya C, de Jong H, Thieffry D. (2006) Dynamical modeling of biological regulatory networks. Biosystems 84(2):77–80.
9. Tournier L. and Chaves M. (2009) Uncovering operational interactions in genetic networks using asynchronous boolean dynamics. J Theor Biol 260(2):196–209.
10. Feller W (1968) An introduction to probability theory and its applications, vol. 1 Wiley, New York.
11. Turanyi, T (1990). Sensitivity analysis of complex kinetic systems. Tools and applications. J Math Chem 5:203–248.
12. Fauré A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics 22(14):e124–e131.
13. Naldi A, Berenguier D, Faure A, Lopez F, Thieffry D, Chaouiya C. (2009) Logical modelling of regulatory networks with GINsim 2.3. Biosystems 97(2):134–139.
14. Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, et al. (1997) Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science 275: 967–969.
15. Lissat A, Vraetz T, Tsokos M, Klein R, Braun M, et al. (2007) Interferon-gamma sensitizes resistant Ewing's sarcoma cells to tumor necrosis factor apoptosis-inducing ligand-induced apoptosis by up-regulation of caspase-8 without altering chemosensitivity. Am J Pathol 170:1917–1930.
16. Teitz T, Lahti JM, Kidd VJ (2001) Aggressive childhood neuroblastomas do not express caspase-8: an important component of programmed cell death. J Mol Med 79:428–436.
17. Shivapurkar N, Toyooka S, Eby MT, Huang CX, Sathyanarayana UG, et al. (2002) Differential inactivation of caspase-8 in lung cancers. Cancer Biol Ther 1:65–69.
18. Croce CM (2008) Oncogenes and cancer. N Engl J Med 358:502–511.
19. Karin M, Cao Y, Greten FR, Li ZW (2002) NF-kappaB in cancer: from innocent bystander to major culprit. Nat Rev Cancer 2:301–310.
20. Dai Z, Zhu WG, Morrison CD, Brena RM, Smiraglia DJ, et al. (2003) A comprehensive search for DNA amplification in lung cancer identifies inhibitors of apoptosis cIAP1 and cIAP2 as candidate oncogenes. Hum Mol Genet 12:791–801.
21. Imoto I, Tsuda H, Hirasawa A, Miura M, Sakamoto M, et al. (2002) Expression of cIAP1, a target for 11q22 amplification, correlates with resistance of cervical cancers to radiotherapy. Cancer Res 62:4860–4866.
22. Imoto I, Yang ZQ, Pimkhaokham A, Tsuda H, Shimada Y, et al. (2001) Identification of cIAP1 as a candidate target gene within an amplicon at 11q22 in esophageal squamous cell carcinomas. Cancer Res 61:6629–6634.

23. Chen DJ, Huerta S (2009) Smac mimetics as new cancer therapeutics. Anticancer Drugs 20(8): 646–658.
24. Ready N, Karaseva NA, Orlov SV, Luft AV, Popovych O, Holmlund JT, Wood BA, Leopold L (2011) Double-blind, placebo-controlled, randomized phase 2 study of the proapoptotic agent AT-101 plus docetaxel, in second-line non-small cell lung cancer. J Thorac Oncol 6(4): 781–785.
25. Lavrik IN (2010) Systems biology of apoptosis signaling networks. Curr Opin Biotechnol 21(4):551–555.

Chapter 16
Theoretical Aspects of Cellular Decision-Making and Information-Processing

Tetsuya J. Kobayashi and Atsushi Kamimura

Abstract Microscopic biological processes have extraordinary complexity and variety at the sub-cellular, intra-cellular, and multi-cellular levels. In dealing with such complex phenomena, conceptual and theoretical frameworks are crucial, which enable us to understand seemingly different intra- and inter-cellular phenomena from unified viewpoints. Decision-making is one such concept that has attracted much attention recently. Since a number of cellular behavior can be regarded as processes to make specific actions in response to external stimuli, decision-making can cover and has been used to explain a broad range of different cellular phenomena [Balázsi et al. (Cell 144(6):910, 2011), Zeng et al. (Cell 141(4):682, 2010)]. Decision-making is also closely related to cellular information-processing because appropriate decisions cannot be made without exploiting the information that the external stimuli contain. Efficiency of information transduction and processing by intra-cellular networks determines the amount of information obtained, which in turn limits the efficiency of subsequent decision-making. Furthermore, information-processing itself can serve as another concept that is crucial for understanding of other biological processes than decision-making. In this work, we review recent theoretical developments on cellular decision-making and information-processing by focusing on the relation between these two concepts.

1 Introduction

A traditional example of cellular decision-making is regulation of the Lac operon, in which a cell controls lactose uptake and metabolism by switching the expression of the operon in response to environmental cues [1, 2]. This binary switch, together

T.J. Kobayashi (✉) • A. Kamimura
Institute of Industrial Science, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
e-mail: tetsuya@mail.crmind.net; kamimura@sat.t.u-tokyo.ac.jp

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_16,

with the lytic–lysogenic switch of the λ-phage, have served as prototypes for cellular decision-making for decades [3], and established the conceptual and technical basis for understanding other decision-making phenomena such as metabolic switches, differentiation, and apoptosis [4–6].

This fundamental and classical problem of binary cellular decision-making has attracted renewed attention recently because the stochastic nature of cellular decision-making has been observed directly by single-cell-imaging technology. Balaban et al. have revealed that the stochastic switching of two phenotypes is a mechanism of bacterial persistence, by directly observing that genetically identical cells have at least two phenotypes with different antibiotic resistances and replication rates [7]. Süel et al. also showed that genetically identical *Bacillus subtilis* cells stochastically select replication, competence, and sporulation when they are exposed to a starving environment [8]. These examples have led us to realize that cellular decision-making has both deterministic and stochastic aspects so that randomness is exploited and to some extent controlled [9, 10]. Even though stochasticity in cellular decision-making was suggested theoretically several decades ago, and it has been partially proven by indirect experimental observations, these results have sufficient impact to attract attention to the classic, yet new, cellular decision-making problem [11, 12]. Furthermore, stochastic phenotypic switching in persistence has provided an experimental evidence to support the idea that cells exploit stochasticity to survive in unpredictable environments [13–17].

Once stochasticity is introduced into the problem, however, it becomes very difficult to intuitively understand whether or not an observed randomness in cellular decision-making is actually exploited to increase the fitness advantage of a cell. Since all cellular phenomena are implemented by intra-cellular reactions that are intrinsically noisy [16, 18–21], the stochastic output of cellular decision-making could be a mere consequence of noise in the intra-cellular network that implements decision-making. Furthermore, stochasticity in intra-cellular reactions can also be a potential source of impairment, rather than improvement, of the efficiency of decision-making, by disturbing the information produced by environmental cues or by randomizing the final output of the decision-making. Theoretical analysis is indispensable for embracing this tangled relation between the constructive and destructive roles of stochasticity.

In this work, we first review the theory of cellular decision-making strategies; this provides a bird's-eye view of the problem. The mechanism of how bet–hedging strategies can be advantageous is illustrated on the basis of traditional work by Levins [22], and the effects of cue-dependent decision-making are subsequently introduced based on recent work [14, 23]. This theory together with other seminal results shows that fitness functions, statistics of environmental fluctuation, and environmental information obtained via cues are major determinants in decision-making. Among these factors, environmental information has a substantial influence on the optimality of a decision-making strategy and its maximum fitness advantage. Secondly, therefore, we review recent work on how such information is quantified

and what kind of intra-cellular networks can convey information efficiently. Finally, we discuss theoretical challenges in cellular decision-making and information-processing.

2 Cellular Decision-Making Without Environmental Sensing

The simplest situation in cellular decision-making is that a cell or cells do not have the ability to know the state of the environment they inhabit. If a certain environmental state occurs very infrequently, it is probable that a cell does not possess a sensory system for observing that state because of the cost of sustaining the system. Persistence in bacteria may be categorized as this type of cellular decision-making [7, 12].

Let $X(t)$ be the state of the environment at time $t \in \mathbb{N}$, where we assume discrete time for simplicity. Also, let us denote the phenotype or action of a cell with genotype g at t as $a^g(t)$ (Fig. 16.1a(i)). If a population of cells with genotype g is sufficiently large such that its total population size $N^g(t)$ can be approximated as a continuous variable, then we can obtain a probability distribution of the cell phenotypes at t as $\mathbb{P}_t^g(a)$. Furthermore, if we define a doubling probability $P_{\mathrm{DB}}(X(t), a(t))$ and a death probability $P_{\mathrm{DT}}(X(t), a(t))$ of a cell whose phenotype is $a(t)$ at t, then we can calculate the total population size at $t+1$ as:

$$N^g(t+1) = \sum_a \mathbb{P}_t^g(a)[1 + P_{\mathrm{DB}}(X(t), a) - P_{\mathrm{DT}}(X(t), a)]N^g(t).$$

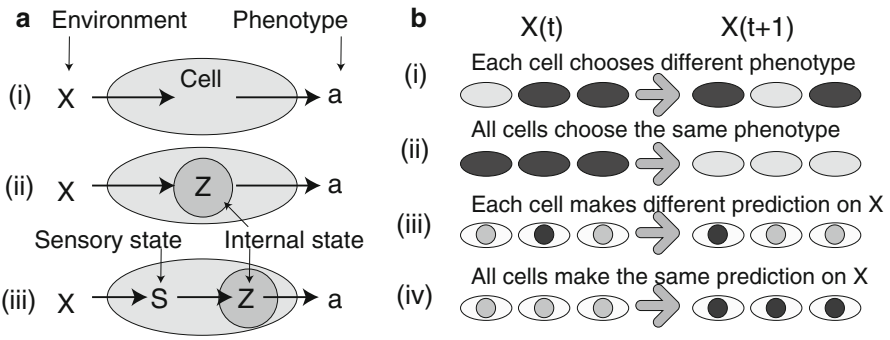


Fig. 16.1 (**a**) Schematic diagrams of models of cellular decision-making and information-processing. (i) The simplest model in which only the environmental state X and phenotype of a cell a are considered. (ii) A model in which the internal state Z, representing the information on X, is incorporated. (iii) A model in which information-processing from a sensory state S to the internal prediction Z of X is considered. (**b**) Schematic diagrams of different types of phenotypic switching (i), (ii) and information transmission errors (iii), (iv). (i) Each cell chooses phenotype stochastically and independently. (ii) All cells have the same phenotype, which changes stochastically over time. (iii) Each cell makes a different prediction $Z(t)$ on $X(t)$. (iv) All cells make the same prediction $Z(t)$ on $X(t)$

By defining the marginal growth rate $R(X,a)$ of a cell with phenotype a under the environment X as $R(X,a) := 1 + P_{\mathrm{DB}}(X,a) - P_{\mathrm{DT}}(X,a)$, we have $N^g(t+1) = [\sum_a \mathbb{P}_t^g(a) R(X(t),a)] N^g(t)$. Thus, the total population N_T^g can be described as $N^g(T) = \prod_{t=0}^{T-1} [\sum_a \mathbb{P}_t^g(a) R(X(t),a)] N^g(0)$, where $N^g(0)$ is the initial population size at $t = 0$. The logarithm of the long-term mean growth-rate of the cells with genotype g becomes:

$$L_{\mathrm{m}}(g) = \lim_{T\to\infty} \frac{\log N^g(T)/N^g(0)}{T} = \lim_{T\to\infty} \left[\frac{1}{T} \sum_{t=0}^{T-1} \log \left[\sum_a \mathbb{P}_t^g(a) R(X(t),a) \right] \right].$$

We estimate $L_{\mathrm{m}}(g)$ approximately by replacing the temporal averaging with respect to t by an ensemble averaging with respect to X, i.e.,

$$L_{\mathrm{m}}(g) = \sum_X \mathbb{P}(X) \log \left[\sum_a \mathbb{P}^g(a) R(X,a) \right],$$

where $\mathbb{P}(X)$ is the probability that X occurs and $\mathbb{P}^g(a)$ is assumed to be independent of t. As it is easily seen, $\exp[L_{\mathrm{m}}(g)]$ contains both the geometric mean with respect to X and the arithmetic mean with respect to a; this stems from the fact that a population of cells rather than an individual cell is the unit of decision-making, as shown in Fig. 16.1b(i) [22]. This property contrasts sharply with decision-making by individual humans or animals in which an agent, i.e., individual human or animal, can choose only one action at one time [24]. This situation is equivalent to the cellular decision-making where cells make the same decision at the same time (Fig. 16.1b(ii)). If the agent of decision-making is an individual, as in standard statistical decision theory, $L_{\mathrm{m}}(g)$ will become $L_{\mathrm{m}}^{\mathrm{ind}}(g) = \left[\sum_{X,a} \mathbb{P}(X)\mathbb{P}(a) \log[R(X,a)]\right]$, in which only the arithmetic mean of $\log[R(X,a)]$ appears [24]. The advantages of stochastic phenotypic switching strongly depend on this difference, as is demonstrated in the next section.

In the following, we describe $L_{\mathrm{m}}(g)$ as $L_{\mathrm{m}}(g) = \sum_i \mathbb{P}(X_i) \log R_{\mathrm{X}}(g|X_i)$, where we define $R_{\mathrm{X}}(g|X_i) := \sum_a \mathbb{P}^g(a) R(X_i,a)$. Since $L_{\mathrm{m}}(g)$ depends on g only via $R_{\mathrm{X}}(g|X_i)$, a vector of $R_{\mathrm{X}}(g|X_i)$ for i, $R_{\mathrm{X}}(g)$, is sufficient information to characterize a genotype g [22]. To intuitively illustrate the behavior of $L_{\mathrm{m}}(g)$, we focus mainly on the simplest situation, where the environment has two states, $X \in \{X_1, X_2\}$. Figure 16.2a and b shows the contour plot of $L_{\mathrm{m}}(g)$ as a function of $R_{\mathrm{X}}(g|X_1)$ and $R_{\mathrm{X}}(g|X_2)$. Thus, the value of the contour at $R_{\mathrm{X}}(g)$ on the plot defines the fitness of the genotype g. In the following, we assume that the phenotype a_i has advantages in environment X_i over the others, so $R(X_1,a_1) > R(X_1,a_2)$ and $R(X_2,a_1) < R(X_2,a_2)$, without losing generality.

Let us first consider the simplest case, in which cells with genotype g_j have a single and fixed phenotype a_j. Then, $R_{\mathrm{X}}(g_j|X_i) = R(X_i,a_j)$ holds, and the contour value at $(R(X_1,a_j), R(X_2,a_j))^{\mathbb{T}}$ corresponds to $L_{\mathrm{m}}(g_j)$. Thus, if we have two populations with different genotypes g_1 and g_2, as in Fig. 16.2a and b, then the population that has the highest $L_{\mathrm{m}}(g)$ will outcompete the others.

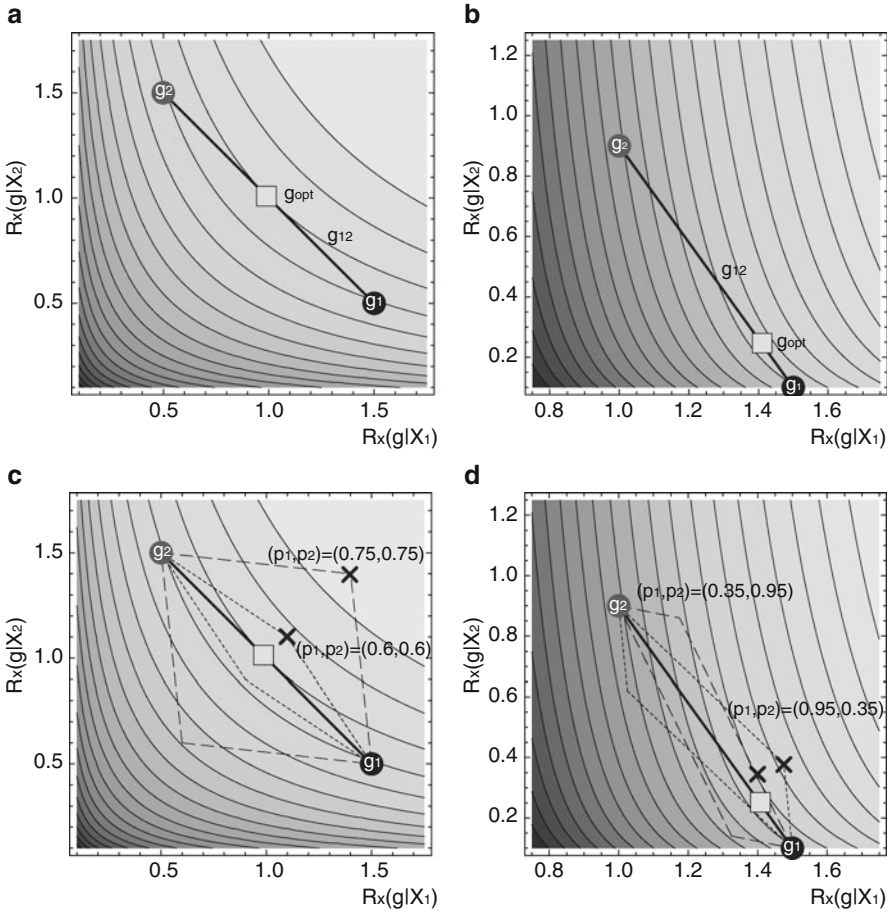


Fig. 16.2 Contour plots of $L_{\mathrm{m}}(g)$ as a function of $R_{\mathrm{X}}(g) = (R_{\mathrm{X}}(g|X_1), R_{\mathrm{X}}(g|X_2))^{\mathbb{T}}$. Lighter colors represent higher values of $L_{\mathrm{m}}(g)$. The circles represent $R_{\mathrm{X}}(g)$ of genotype g, designated within the circle. Bold lines represent the set of $R_{\mathrm{X}}(g_{12})$ swept by changing $\mathbb{P}^{g_{12}}(a)$ within $[0, 1]$. Squares are g_{12} with optimal switching rates. (**a**) and (**c**) Symmetric situation in which $(\mathbb{P}(X_1), \mathbb{P}(X_2)^{\mathbb{T}} = (1/2, 1/2)^{\mathbb{T}}$, $R_{\mathrm{X}}(g_1) = (3/2, 1/2)^{\mathbb{T}}$, and $R_{\mathrm{X}}(g_2) = (1/2, 3/2)^{\mathbb{T}}$. (**b**) and (**d**) Asymmetric situation in which $(\mathbb{P}(X_1), \mathbb{P}(X_2)^{\mathbb{T}} = (9/10, 1/10)^{\mathbb{T}}$, $R_{\mathrm{X}}(g_1) = (3/2, 1/10)^{\mathbb{T}}$, and $R_{\mathrm{X}}(g_2) = (1, 9/10)^{\mathbb{T}}$. Dashed lines in (**c**) and (**d**) define the boundaries of the regions that $R_{\mathrm{X}}(g)$ sweeps by changing q_1 and q_2 for fixed values of p_1 and p_2. Each line can be defined either by $q_1 = 0$, $q_1 = 1$, $q_2 = 0$, or $q_2 = 1$. Crosses correspond to genotypes with optimal q_1 and q_2

Next, we consider a cell with genotype g_{12} that has the ability to randomly select either phenotype a_1 or phenotype a_2. Then, its average growth rate can be described as $R_{\mathrm{X}}(g_{12}) = \sum_{i\in\{1,2\}} \mathbb{P}^{g_{12}}(a_i) R_{\mathrm{X}}(g_i)$, where $R_{\mathrm{X}}(g_i) = (R(X_1, a_i), R(X_2, a_i))^{\mathbb{T}}$. Thus, $R_{\mathrm{X}}(g_{12})$ is on the segment from $R_{\mathrm{X}}(g_1)$ to $R_{\mathrm{X}}(g_2)$, as in Fig. 16.2a and b. As we can easily see graphically, there can be a point on the segment at which $L_{\mathrm{m}}(g_{12})$ can be greater than $L_{\mathrm{m}}(g_1)$ and $L_{\mathrm{m}}(g_2)$, i.e., the squares in Fig. 16.2a and b. Thus, stochastic phenotypic switching can be advantageous. Advantage of

bet–hedging strategies (also known as mixed strategies) in fluctuating environments was described theoretically by Levins in the 1960s [22]. In the field of evolutionary biology, the conditions of the optimality have been further explored for more general situations [23], and the evolutionary stability of the bet–hedging strategies was investigated [25]. As mentioned, this phenomenon attracts renewed attention recently in the context of cellular decision-making [26–28]. The merit of stochastic switching stems from the mixture of geometric and arithmetic means in $L_{\mathrm{m}}(g)$, by which the contour of $L_{\mathrm{m}}(g)$ becomes curved, whereas the set of $R_{\mathrm{X}}(g_{12})$ is a segment, as shown in Fig. 16.2a and b. However, either g_1 or g_2 rather than g_{12} becomes advantageous when the decision-making is conducted at the individual level, as in Fig. 16.1b(ii), and thus, $L_{\mathrm{m}}(g)$ obtains only arithmetic means [24]. This result clearly illustrates that the ability of genetically identical cells to conduct population-level decisions is crucial to exploiting stochasticity for survival.

3 Cellular Decision-Making with Environmental Sensing

In this section, we consider the case where a cell can employ cues or information on the state of the environment obtained by sensing the environment. In this case, the environmental cue can be a signal from other cells as well as the actual state of the environment such as the amount of nutrients or toxic molecules. This type of decision-making includes at least metabolic switches [4, 11], apoptosis [6], differentiation [5], stress responses [29], viral latency decisions [30, 31], and λ-phage lysis/lysogenesis [3].

When environmental information is available, $L_{\mathrm{m}}(g)$ and its approximation by ensemble averaging becomes:

$$
\begin{aligned}
L_{\mathrm{m}}(g) &= \lim_{T\to\infty}\left[\sum_{t=0}^{T-1}\log\left[\sum_{a}\mathbb{P}_t^g(a|X(t))R(X(t),a)\right]\right]\Big/T, \\
&\approx \sum_{X}\mathbb{P}(X)\log\left[\sum_{a}\mathbb{P}^g(a|X)R(X,a)\right],
\end{aligned}
$$

where $\mathbb{P}_t^g(a|X(t))$ is the phenotype distribution in a population with genotype g under environment $X(t)$. As an extreme situation, let us consider that a cell can perfectly predict the state of the environment. Then, we can easily see that the strategy to deterministically select the best phenotype based on the prediction becomes the most advantageous as $L_{\mathrm{m}}(g_p) = \sum_X \mathbb{P}(X)\log[\max_a R(X,a)]$. Thus, stochastic phenotypic switching is no longer advantageous when perfect information is available.

In reality, however, a cell can neither obtain perfect information on $X(t)$ nor conduct deterministic phenotypic switching. Under an environmental state $X(t)$, each cell may have different predictions for $X(t)$ by its internal state $Z(t)$, as in

Fig. 16.1a(ii), and $Z(t)$ may not completely correlate with $X(t)$ as $\mathbb{P}(Z(t)|X(t))$. This discrepancy between $Z(t)$ and $X(t)$ can be attributed to the intra-cellular noise in the sensory and signal transduction pathways, or the inability of a cell to exploit all the relevant information obtained from its sensory system. On the basis of this internal representation of the environment Z, a cell subsequently makes a specific action or chooses a specific phenotype a deterministically or stochastically as $\mathbb{P}_t^g(a|Z)$. Then, we obtain $L_{\mathrm{m}}(g)$ as:

$$\begin{aligned} L_{\mathrm{m}}(g) &= \lim_{T\to\infty}\left[\sum_{t=0}^{T-1}\log\left[\sum_{a,Z}\mathbb{P}_t^g(a|Z)\mathbb{P}_t^g(Z|X(t))R(X(t),a)\right]\right]\Big/ T, \\ &\approx \sum_X \mathbb{P}(X)\log[R_{\mathrm{X}}(g|X)], \end{aligned}$$

where $R_{\mathrm{X}}(g|X) = \sum_{a,Z}\mathbb{P}^g(a|Z)\mathbb{P}^g(Z|X)R(X,a)$. For simplicity, assume that Z_i represents an intra-cellular state predicting that $X = X_i$. Then, $\mathbb{P}^g(Z_i|X_i)$ is the probability that the internal state Z can correctly predict the state of X as X_i, whereas $\sum_{j\neq i}\mathbb{P}^g(Z_j|X_i) = 1 - \mathbb{P}^g(Z_i|X_i)$ is the probability that the internal state fails to predict X. Furthermore, we only consider binary environments, binary internal states, and binary phenotypes such that $X \in \{X_1, X_2\}$, $Z \in \{Z_1, Z_2\}$, and $a \in \{a_1, a_2\}$; we define the following symbols for notational simplicity:

$$\begin{aligned} R_{i,j} &:= R(X_i, a_j), \quad p_i := p_{i,i} := \mathbb{P}^g(Z_i|X_i), \quad q_i := \mathbb{P}^g(a_i|Z_i), \\ \tilde{p}_i &:= \tilde{p}_{k,i} = 1 - p_i = \mathbb{P}^g(Z_k|X_i), \qquad \tilde{q}_i = 1 - q_i = \mathbb{P}^g(a_k|Z_i), \end{aligned}$$

where $i \in \{1, 2\}$, and $k = 1$ when $i = 2$ and $k = 2$ otherwise. Then, we have

$$R_{\mathrm{X}}(g|X_i) = \begin{pmatrix} R_{i,1}, & R_{i,2} \end{pmatrix}\begin{pmatrix} q_1 & \tilde{q}_2 \\ \tilde{q}_1 & q_2 \end{pmatrix}\begin{pmatrix} p_{1,i} \\ \tilde{p}_{2,i} \end{pmatrix}.$$

For constant accuracy of the sensory system represented by fixed p_1 and p_2, we can graphically describe the region that $R_{\mathrm{X}}(g)$ can sweep, as in Fig. 16.2c and d, by changing q_1 and q_2 within $[0, 1]$. We can easily see that $L_{\mathrm{m}}(g)$ becomes maximum when $q_1 = 1$, or $q_2 = 1$, or $q_1 = 1$ and $q_2 = 1$ holds. Because q_i defines the probability of choosing the phenotype a_i when Z_i, the condition $q_i = 1$ means that the phenotype a_i is always selected deterministically when $Z = Z_i$. Therefore, actively randomizing the phenotype is less encouraged when information Z on X is available (see also [14] for more detailed analysis).

To check the consistency of this result with the fact that stochastic phenotypic switching is encouraged when no information on X is available, let us define $p_1 = p_0 + \varepsilon_1$ and $p_2 = \tilde{p}_0 + \varepsilon_2$. ε_1 and ε_2 characterize the fidelity of the sensory system because no information on X_i is obtained from Z when $\varepsilon_i = 0$. Then

$$\begin{pmatrix} R_{\mathrm{X}}(g|X_1) \\ R_{\mathrm{X}}(g|X_2) \end{pmatrix} = \begin{pmatrix} R_{\mathrm{X}}(g_{12}|X_1) \\ R_{\mathrm{X}}(g_{12}|X_2) \end{pmatrix} + 2\Delta q_{12}\begin{pmatrix} \varepsilon_1(R_{1,1} - R_{1,2}) \\ \varepsilon_2(R_{2,2} - R_{2,1}) \end{pmatrix},$$

where $q_{12} := (p_0 q_1 + \tilde{p}_0 \tilde{q}_2)$, $\Delta q_{12} := \frac{q_1 - \tilde{q}_2}{2}$, and $R_X(g_{12}|X_i) = [R_{i,1} q_{12} + R_{i,2} \tilde{q}_{12}]$; $R_X(g_{12}|X_i)$ is equivalent to that used for the case of no sensory system, if $q_{12} = \mathbb{P}^{g_{12}}(a_1)$. Thus, the term $R_X(g_{12}|X_i)$ represents the factor of decision-making without sensing, and $2\varepsilon_i \Delta q_{12}(R_{i,1} - R_{i,2})$ purely describes the influence of obtaining information on X. Optimal stochastic phenotypic switching can be achieved as long as $(p_0 q_1 + \tilde{p}_0 \tilde{q}_2) = q_{12}^{\mathrm{opt}}$ is satisfied. Because p_0 is involved in q_{12}, the uncertainty in predicting X can be a passive source of stochasticity in phenotypic switching rather than an active one in which stochasticity or randomness is actively generated. Furthermore, $(p_0 q_1 + \tilde{p}_0 \tilde{q}_2) = q_{12}^{\mathrm{opt}}$ contains two free parameters other than p_0; these cannot be determined solely from this optimality. Since the optimal strategy requires $q_i = 1$ for either $i \in \{1, 2\}$ when $\varepsilon \neq 0$, it may be advantageous to have either $q_1 = 1$ or $q_2 = 1$, even when no information on X is available. Thus, active phenotype randomization by having $q_1, q_2 \neq 1$ is necessarily not advantageous, even if no information is available [14].

It should be noted that the conclusion that active randomization of phenotypes is not encouraged depends on the type of error in predicting the environment. In our case, we assume that the error in each cell is independent because we consider that the intrinsic noise of each cell is the major source of the error, as in Fig. 16.1b(iii). However, errors may be correlated among cells when there is a common disturbance of the sensing process in the environment, as in Fig. 16.1b(iv). In [14,23], it is shown that active randomization of phenotypes can be advantageous in the latter case, demonstrating that the mechanism for obtaining information on X, and the amount of information obtained, substantially influence the optimality of a strategy for phenotypic switching. Therefore, the mechanism and efficiency of cellular information-processing are also important determinants of cellular decision-making. In the next section, we focus on recent theoretical advances in the information-processing of cells by sensing environments with noisy intra-cellular reactions.

4 Information Transmission for Decision-Making

As already shown, obtaining information on the environmental state X as an intra-cellular state Z is a crucial step in making appropriate decisions in a changing environment. In general, receptor reactions and subsequent signaling pathways in a cell conduct this information transmission by relaying the received signal into a cell. Noise in intra-cellular reactions, however, makes this basic task more difficult and non-trivial than it sounds [18–21]. It is expected that intra-cellular networks and pathways with specific structures will have higher efficiency than others in transmitting relevant information, and that the efficiency is bounded by physical constraints that are specific to intra-cellular reaction dynamics. This problem is also relevant to other intra-cellular phenomena besides cellular decision-making because robust operation of intra-cellular and inter-cellular phenomena such as

metabolism, cell cycles, transcriptional regulation, and development [32, 33] may require efficient flow of information from the input component X to the output component Z. Because of its importance and broad implications, this problem has been investigated theoretically for various intra-cellular and inter-cellular phenomena [34].

For example, information transmission by intra-cellular signaling cascades has been investigated in [35–37]. Among several pathways, the most investigated are chemotactic pathways in which a cell senses information on a ligand gradient by its sensory system [38–43]. Because both data on quantitative characteristics and underlying molecular details are available, chemotactic pathways act as a benchmark platform for integrative analysis of cellular information transmission, theoretically and experimentally. Another important class of phenomena is gene expression and regulation. Because information is transmitted from the state of a gene or its regulator X to its expression level Z, gene expression and regulation are examples of information transmission between intra-cellular components [44, 45]. Specifically, the readout of positional information from a morphogen gradient has been analyzed theoretically and experimentally [46–48]. Interlocked fast and slow positive feedback loops have also been proven to produce a distinct output robustly from a noisy signal [49, 50].

Although the specific biological details differ, the performance of information transmission is characterized by quantifying the efficiency of transmission by various measures. One of the most frequently used measures is the variance of Z, $\sigma_Z^2(X)$, for a fixed X or its variants such as the coefficient of variation (CV), defined as $\mathrm{CV}(X) = \frac{\sigma_Z(X)}{\langle Z(X)\rangle}$, where $\langle Z(X)\rangle$ is the average of Z for a fixed X.

Let us consider the problem of discriminating between the state X and $X' = X + \Delta X$ on the basis of Z, when ΔX is small. Intuitively, the difference between X and X' can be determined less ambiguously when we have either larger ΔZ or smaller $\sigma_Z(X)$. Since $\frac{\Delta\langle Z\rangle}{\sigma_Z(X)} = \frac{1}{\sigma_Z(X)}\frac{\mathrm{d}\langle Z(X)\rangle}{\mathrm{d}X}\Delta X$ holds for small ΔX, $\frac{1}{\sigma_Z(X)}\frac{\mathrm{d}\langle Z(X)\rangle}{\mathrm{d}X}$ or $\frac{1}{\mathrm{CV}(X)}\frac{\mathrm{d}\log\langle Z(X)\rangle}{\mathrm{d}X}$ works as a measure of the fidelity of Z in transmitting information on X, and has been used in various applications [38–41, 46]. The advantage of this measure is that, when the molecular detail of an intra-cellular network is available, the variance and CV can be approximately calculated by linearization of a chemical master equation or Langevin equation of the network [51, 52]. Thus, we can evaluate parameter dependence and lower bound for a specific intra-cellular reaction. Furthermore, variance and CV are easier to estimate experimentally than other measures that we introduce below. Nevertheless, $\frac{1}{\sigma_Z(X)}\frac{\mathrm{d}\langle Z(X)\rangle}{\mathrm{d}X}$ has some disadvantages compared to other measures. For example, even when the value of the measure is specified, the probability of making a wrong prediction on the state of X, based on Z, is still not explicitly obvious. Furthermore, the average and standard deviation cannot exactly specify the underlying probability distribution of Z for fixed X. To resolve this problem, we have to employ more detailed information on the distribution $\mathbb{P}(Z|X)$.

Let $\{Z^k\}_{k\in\mathbb{N}}$ be an ensemble of Z generated from either $\mathbb{P}(Z|X)$ or $\mathbb{P}(Z|X + \Delta X)$. When we observe Z^k, then $\mathbb{P}(Z^k|X)$ and $\mathbb{P}(Z^k|X + \Delta X)$ are the likelihoods

that Z_k is generated from $\mathbb{P}(Z|X)$ and $\mathbb{P}(Z|X+\Delta X)$, respectively. Thus, their ratio $\mathscr{L}_k = \log[\mathbb{P}(Z^k|X+\Delta X)/\mathbb{P}(Z^k|X)]$ measures the relative likelihood of observing Z^k from $\mathbb{P}(Z|X)$ or $\mathbb{P}(Z|X+\Delta X)$. Because of the logarithm, $\mathscr{L}_k > 0$ when $X+\Delta X$ is more likely, $\mathscr{L}_k < 0$ when X is more likely, and $\mathscr{L}_k = 0$ when $X+\Delta X$ and X are equally likely. By averaging $\mathscr{L}_k$ for all $k \in \mathbb{N}$, we obtain the average log-likelihood as:

$$\langle \mathscr{L} \rangle = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathscr{L}_k \approx \int \mathbb{P}(Z) \log \frac{\mathbb{P}(Z|X+\Delta X)}{\mathbb{P}(Z|X)} \mathrm{d}Z,$$

where $\mathbb{P}(Z)$ is either $\mathbb{P}(Z|X)$ or $\mathbb{P}(Z|X+\Delta X)$. When $\mathbb{P}(Z) = \mathbb{P}(Z|X+\Delta X)$, $\langle \mathscr{L} \rangle$ becomes the Kullback–Leibler divergence $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X+\Delta X)||\mathbb{P}(Z|X)]$, defined as $\mathscr{D}_{\mathrm{KL}}[p(x)||q(x)] := \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ [53]. Similarly, $\langle \mathscr{L} \rangle = -\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X)||\ \mathbb{P}(Z|X+\Delta X)]$, when $\mathbb{P}(Z) = \mathbb{P}(Z|X)$. Thus, $|\ \langle \mathscr{L} \rangle\ |$ measures the ease of discriminating between $X+\Delta X$ and X on the basis of Z. Furthermore, $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}'(Z)||\mathbb{P}(Z)]$ is related to the probability that $\{Z^1, \ldots, Z^n\}$ happens to behave as if they are generated from $\mathbb{P}'(Z)$ even though they are actually generated from $\mathbb{P}(Z)$ [53]. By expanding $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X+\Delta X)||\mathbb{P}(Z|X)]$ with respect to ΔX, we then obtain

$$\langle \mathscr{L} \rangle \approx \mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X+\Delta X)||\mathbb{P}(Z|X)] \approx \mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))\Delta X^2,$$

where $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ is the Fisher information, defined as:

$$\mathscr{D}_{\mathrm{F}}(p(x|y)) := \int p(x|y) \left[\frac{\partial \log p(x|y)}{\partial y} \right]^2 \mathrm{d}x.$$

Thus, $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ can be used to measure the fidelity of inferring X from Z. Furthermore, $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ also works as a statistical limit for the inference such that the inverse of $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ is related to the lower bound of the variance of the inferred X [53]. If $\mathbb{P}(Z|X)$ is a Gaussian distribution whose mean and variance are $\langle Z(X) \rangle$ and σ_Z^2, then $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X)) = \left(\frac{1}{\sigma_Z} \frac{\mathrm{d}\langle Z(X) \rangle}{\mathrm{d}X} \right)^2$ holds, indicating that $\mathscr{D}_{\mathrm{F}}(p(x|y))$ is a generalization of the measure defined by variance.

Although $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ is more general and informative than the variance, it is just a local measure in the sense that it accounts for the vicinity of X. To discriminate X from all the other states, we may be able to use $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X)||\mathbb{P}(Z)]$, where $\mathbb{P}(Z) = \int \mathbb{P}(Z|X)\mathbb{P}(X))\mathrm{d}X$. Furthermore, $\mathscr{D}_{\mathrm{F}}(\mathbb{P}(Z|X))$ does not employ information on how X behaves or implicitly assume that all X occur evenly. Even though a cell does not sense X via Z, it can still predict the state of X statistically if it acquires prior information on X. Thus, by averaging $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X)||\mathbb{P}(Z)]$ over $\mathbb{P}(X)$, this prior information on X can be accounted in for the mutual information $\mathscr{I}[X;Z]$:

$$\mathscr{I}[X;Z] := \int \mathbb{P}(X)\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(Z|X)||\mathbb{P}(Z)]\mathrm{d}X$$
$$= \int\int \mathbb{P}(X,Z)\log\frac{\mathbb{P}(X,Z)}{\mathbb{P}(X)\mathbb{P}(Z)}\mathrm{d}X\,\mathrm{d}Z.$$

The mutual information has also been used to quantify the efficiency of information transmission in signaling pathways, morphogen gradient sensing, quorum sensing, and gene regulatory networks with various network structures [44, 45, 47, 54–56]. We introduce only the static version of mutual information here for simplicity, but its dynamic version, which accounts for the time-series of X_t and Z_t, has also been applied to different biological pathways [57–59]. Because only $\mathbb{P}(X,Z)$ is needed to calculate the static $\mathscr{I}[X;Z]$, $\mathscr{I}[X;Z]$ potentially has wide applicability, not only for theoretical but also for experimental evaluation of the efficiency of information transmission.

5 Information-Processing Required for Efficient Information Transmission

Although the information measures introduced above provide us with ways of quantifying the efficiency of information transmission for a given intra-cellular network, we have to specify a network structure before applying the measures. Thus, in order to reveal a specific structure or response having optimal efficiency in conducting such information transmission, other approaches than information measures become crucial.

The theory of Bayesian inference in statistics has been employed to predict an optimal way of obtaining information on X. In general, an intra-cellular state Z cannot contain perfect information on X because the intermediate state S from X to Z has stochasticity, as shown in Fig. 16.1a(iii). For example, the receptor activity on the cell membrane, S, is the first step in sensing the environmental state X. Subsequent processing of S is generally conducted before initiating the response of a cell to the environment. Thus, the processed Z will be a function of S, i.e., $Z(S)$. By Bayesian inferences, the optimal $Z^*(S)$ is the posterior probability of X, given S, defined as $Z^*(S) = \mathbb{P}(X|S)$, and is computed by Bayes's rule as $\mathbb{P}(X|S) \propto \mathbb{P}(S|X)\mathbb{P}(X)$.

This approach was adopted in [60] to derive an optimal way for *E. coli* to infer the amount of extracellular sugar from a noisy intra-cellular sugar concentration. Different genetic regulatory mechanisms were subsequently explored numerically by evaluating their ability to be tuned to the optimal response predicted from Bayes's rule. Similarly, in [61], photoreceptor networks were proposed for approximating Bayesian inferences. In [62–64], Bayes's rule was also applied to the chemotaxis of axons by gradient sensing.

For more macroscopic phenomena than cellular ones, Bayes's rule was extensively used to explain foraging behaviors of animals [65–67], and perception and cognitive processes of humans [68–70]. The applications in these fields illustrate that the advantage of Bayes's rule also lies in its ability to naturally incorporate dynamic updating of the posterior probability $\mathbb{P}(X|S)$ by obtaining new signals from the environment. By exploiting this property, it was clarified in [71–73] that a phosphorylation/dephosphorylation cycle with autoregulatory feedbacks can implement a dynamic Bayesian inference rather than a static one. If temporal dynamics is involved, Bayes's rule is extended as:

$$\mathbb{P}(X(t')|S(0:t'))=\frac{\mathbb{P}(S(t')|X(t'))\int \mathbb{P}(X(t')|X(t))\mathbb{P}(X(t)|S(0:t))\mathrm{d}X(t)}{\int\int \mathbb{P}(S(t')|X(t'))\mathbb{P}(X(t')|X(t))\mathbb{P}(X(t)|S(0:t))\mathrm{d}X(t)\mathrm{d}X(t')},$$

where $t' = t + \Delta t$, $\mathbb{P}(S(t)|X(t))$ is the probability of observing receptor activity $S(t)$ when the environment is in $X(t)$, and $\mathbb{P}(X(t')|X(t))$ defines the stochastic dynamics of the environment within a small interval $(t, t']$. $S(0:t')$ is the time-series of $S(t)$ from $t = 0$ to $t = t'$. For a binary environmental state $X(t) \in \{X_{\mathrm{on}}, X_{\mathrm{off}}\}$ whose states switch by following a two-state Markov process, it was shown that the update dynamics of the posterior probability $Z(t) = \mathbb{P}(X(t) = X_{\mathrm{on}}|S(0:t))$ is reduced to a differential equation:

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t} = [N_0\lambda_{\mathrm{r}}S(t)Z(t) + r_{\mathrm{on}}]\tilde{Z}(t) - [N_0\lambda_{\mathrm{d}}\tilde{Z}(t) + r_{\mathrm{off}}]Z(t), \qquad (16.1)$$

where $\tilde{Z}(t) = 1 - Z(t)$, N_0 is the total number of receptors, $S(t)$ is the effective activity of each receptor, and r_{on} and r_{off} are the rates at which the environmental state switches. λ_{r} and λ_{d} are determined by the probability of the receptor becoming active [71, 72]. As is easily seen, $[N_0\lambda_{\mathrm{r}}S(t)Z(t) + r_{\mathrm{on}}]\tilde{Z}(t)$ and $[N_0\lambda_{\mathrm{d}}\tilde{Z}(t) + r_{\mathrm{off}}]Z(t)$ can be regarded as phosphorylation and dephosphorylation reactions, whose rates are $[N_0\lambda_{\mathrm{r}}S(t)Z(t) + r_{\mathrm{on}}]$ and $[N_0\lambda_{\mathrm{d}}\tilde{Z}(t) + r_{\mathrm{off}}]$, if we assume that $Z(t)$ and $\tilde{Z}(t)$ are the ratios of phosphorylated and dephosphorylated molecules, respectively. Since the phosphorylation and dephosphorylation rates depend on $Z(t)$ and $\tilde{Z}(t)$, respectively, this reaction can be identified with an auto-phosphorylation/auto-dephosphorylation (aPadP) cycle (Fig. 16.3a(left)). As shown in Fig. 16.3b, the aPadP cycle can extract information on $X(t)$ as $Z(t)$, even though $S(t)$ looks extremely noisy [71, 72]. The dynamics equivalent to (16.1) can also be implemented by a dueling reaction proposed as a simplified model of T-cell responses (Fig. 16.3a(center)) and the polarity formation reaction in chemotaxis (Fig. 16.3a(right)) [73]. These results, together with other results, suggest that intra-cellular networks have the potential to conduct statistically optimal inference from noisy signals. Finally, we note that the statistically optimal inference $\mathbb{P}(X(t)|S(0:t))$ is closely linked to the information on $X(t)$ contained in $S(t)$, measured by the mutual information as:

$$\mathscr{I}[X(t); S_{0:t}] = \int \mathscr{D}_{\mathrm{KL}}[\mathbb{P}(X(t)|S_{0:t})||\mathbb{P}(X(t))]\mathbb{P}(S_{0:t})\mathrm{d}S_{0:t}.$$

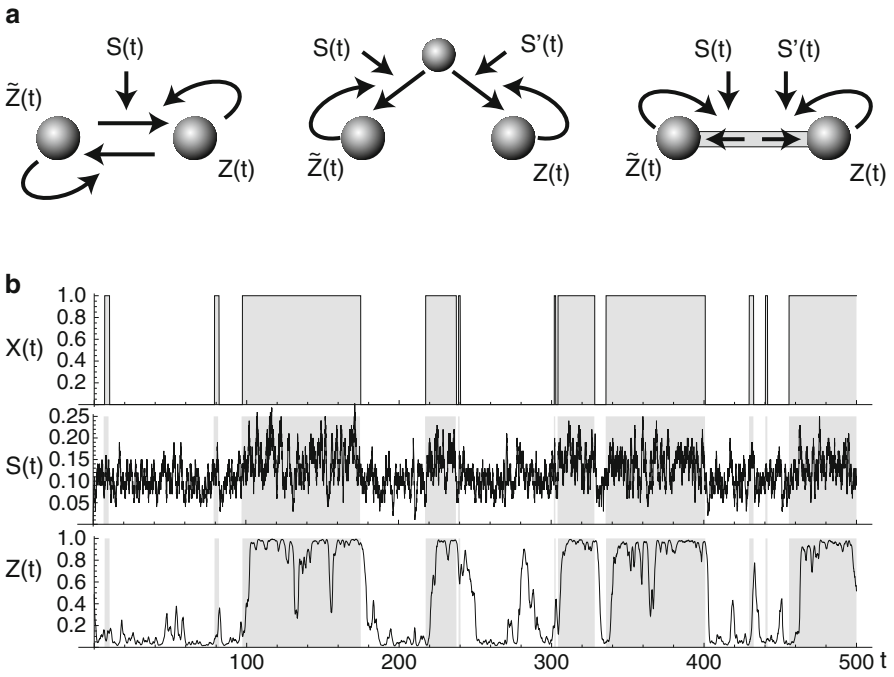


Fig. 16.3 (**a**) Schematic diagrams of reaction networks that can approximately implement dynamic Bayesian inferences (16.1): the aPadP cycle (left) [71], dueling reaction (center) [73], and one-dimensional version of polarity formation (right) [73]. (**b**) Behavior of $Z(t)$ obtained by (16.1). $X(t)$ and $S(t)$ represent the time-series of the environment and a noisy sensory system. All the parameters are the same as those in Fig. 1 of [71]

Since $\mathscr{D}_{\mathrm{KL}}[\mathbb{P}(X(t)|S(0:t))||\mathbb{P}(X(t))]$ in the right-hand side of the equation becomes large when the inferred $X(t)$, $\mathbb{P}(X(t)|S(0:t))$, is distinct from the prior information on $X(t)$ without $S(0:t)$, $\mathbb{P}(X(t))$, a larger $\mathscr{I}[X(t);S(0:t)]$ will lead to a more distinct inference of X by $\mathbb{P}(X(t)|S(0:t))$. Thus, the statistically optimal inference by $\mathbb{P}(X(t)|S(0:t))$ can effectively exploit the information on $X(t)$ that $S(t)$ contains, which in turn improves the efficiency of cellular decision-making by making more informative Z available for subsequent phenotypic choices.

6 Conclusion and Discussion

In this work, we review the theories of cellular decision-making and information-processing, together with related concepts. As demonstrated, a strategy to actively generate heterogeneity can be either advantageous or disadvantageous, depending on whether the fitness advantages of cells are determined at the population level or

the individual level. In both situations, furthermore, the availability of environmental information, when it is appropriately exploited, improves the fitness of cells, indicating that information actually has fitness value [28, 74, 75]. Information transmission is also crucial for a cell, not only in making decisions, but also in regulating noisy intra-cellular components [76, 77], because the limit of noise-suppression by an intra-cellular regulation is related to available information on the change in the target molecule [78]. However, information easily becomes degenerated when it is processed inappropriately, suggesting that some intra-cellular networks are designed to process and transmit information efficiently when the information has high fitness value.

To address this problem, as illustrated in this review, information theory enables us to quantify the efficiency of information transmission of a given network, and statistical theories, such as Bayes's theorem, provide us with a way of predicting optimal intra-cellular networks for information-processing. Although their applications are currently limited, these theories may play a more important role than before in unveiling design principles and the optimality of intra-cellular networks because single-cell-level behavior is becoming experimentally accessible by single-cell time-lapse measurements [79, 80]. Nonetheless, the existing theories developed not for cellular decision-making and information-processing are not sufficient, for example, standard statistical decision theory does not cover the situation in which the bet–hedging strategies become advantageous. Among several extensions required, incorporation of temporal continuity is of particular importance because all cellular processes are continuous-time and stochastic in nature. Although continuous-time problems are mathematically more tangled, their results can be associated easily and consistently with the actual stochastic dynamics of intra-cellular networks [26, 28, 57–59, 71–73]. Furthermore, such theories may be able to make the best use of information obtained by single-cell time-lapse measurements [81]. Integration of information and statistical theories with those of stochastic dynamics [51, 52] may be the next theoretical challenge in achieving a comprehensive understanding of cellular decision-making and information-processing.

Acknowledgements We thank Bashar Md. Khayrul, Yoshihiro Morishita, and Ryo Yokota for fruitful discussions. This work was supported by the JST PRESTO program.

References

1. Muller-Hill B (1996) The lac operon: a short history of a genetic paradigm. Berlin, Walter de Gruyter
2. Vilar JMG, Guet CC, Leibler S (2003) Modeling network dynamics: the lac operon, a case study. J Cell Biol 161(3):471
3. Ptashne M (1992) Genetic switch: phage lambda and higher organisms. Massachusetts, Blackwell Publishers
4. Acar M, Becskei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. Nature 435(7039):228

5. Kalmar T, Lim C, Hayward P, Muñoz Descalzo S, Nichols J, Garcia-Ojalvo J, Martinez Arias A (2009), Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol 7(7):e1000149
6. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature 459(7245):428
7. Balaban N, Merrin J, Chait R, Kowalik L, Leibler S (2004) Bacterial persistence as a phenotypic switch. Science 305(5690):1622
8. Süel GM, Garcia-Ojalvo J, Liberman LM, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. Nature 440(7083):545
9. Ben-Jacob E, Schultz D (2010) Bacteria determine fate by playing dice with controlled odds. Proc Natl Acad Sci USA 107(30):13197
10. Johnston R Jr (2010) Stochastic mechanisms of cell fate specification that yield random or robust outcomes. Ann Rev Cell Dev Biol 26:689
11. Kalisky T, Dekel E, Alon U (2007) Cost-benefit theory and optimal design of gene regulation functions. Phys Biol 4(4):229
12. Jayaraman R (2008) Bacterial persistence: some new insights into an old phenomenon. J Biosci 33(5):795
13. Dhar N, McKinney J (2007) Microbial phenotypic heterogeneity and antibiotic tolerance. Curr Opin Microbiol 10(1):30
14. Donaldson-Matasci M (2008) Adaptation in a changing environment: phenotypic diversity in response to environmental uncertainty and information. Ph.D. thesis, University of Washington
15. Tanase-Nicola S, ten Wolde PR (2008) Regulatory control and the costs and benefits of biochemical noise. PLoS Comput Biol 4(8):e1000125
16. Fraser D, Kaern M (2009) A chance at survival: gene expression noise and phenotypic diversification strategies. Mol Microbiol 71(6):1333
17. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467(7312):167
18. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6(6):451
19. Samoilov MS, Price G, Arkin AP (2006) From fluctuations to phenotypes: the physiology of noise. Science's STKE 2006(366):re17
20. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135(2):216
21. Shahrezaei V, Swain PS (2008) The stochastic nature of biochemical networks. Curr Opin Biotechnol 19(4):369
22. Levins R (1968) Evolution in changing environments: some theoretical explorations. New Jersey, Princeton University Press
23. Haccou P, Iwasa Y (1995) Optimal mixed strategies in stochastic environments. Theor Population Biol 47(2):212
24. Berger JO (1993) Statistical decision theory and Bayesian analysis (Springer Series in Statistics). New York, Springer
25. Sasaki A, Ellner S (1995) The evolutionarily stable phenotype distribution in a random environment. Evolution 49(2):337
26. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. Science 309(5743):2075
27. de Jong IG, Haccou P, Kuipers OP (2011) Bet hedging or not? A guide to proper classification of microbial survival strategies. BioEssays 33(3):215
28. Rivoire O, Leibler S (2011) The value of information for populations in varying environments. J Stat Phys 142(6):1124
29. Zhang XP, Liu F, Cheng Z, Wang W (2009) Cell fate decision mediated by p53 pulses. Proc Natl Acad Sci USA 106(30):12245
30. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell 122(2):169

31. Singh A (2009) Noise in viral gene expression as a molecular switch for viral latency. Curr Opin Microbiol 12(4):460
32. Maheshri N, O'Shea EK (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. Ann Rev Biophys Biomol Struct 36:413
33. Arias AM, Hayward P (2006) Filtering transcriptional noise during development: concepts and mechanisms. Nat Rev Genet 7(1):34
34. Tkačik G, Walczak AM (2011) Information transmission in genetic regulatory networks: a review. J Phys Condens Matt 23(15):153102
35. Samoilov M, Plyasunov S, Arkin A (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci USA 102(7):2310
36. Morishita Y, Kobayashi TJ, Aihara K (2006) An optimal number of molecules for signal amplification and discrimination in a chemical cascade. Biophys J 91(6):2072
37. Gomez-Uribe C, Verghese GC, Mirny La (2007) Operating regimes of signaling cycles: statics, dynamics, and noise filtering. PLoS Comput Biol 3(12):e246
38. Bialek W, Setayeshgar S (2005) Physical limits to biochemical signaling. Proc Natl Acad Sci USA 102(29):10040
39. Ueda M, Shibata T (2007) Stochastic signal processing and transduction in chemotactic response of eukaryotic cells. Biophys J 93:11
40. Shibata T, Ueda M (2008) Noise generation, amplification and propagation in chemotactic signaling systems of living cells. Biosystems 93(1–2):126
41. Endres RG, Wingreen NS (2008) Accuracy of direct gradient sensing by single cells. Proc Natl Acad Sci USA 105(41):15749
42. Rappel WJ, Levine H (2008) Receptor noise and directional sensing in eukaryotic chemotaxis. Phys Rev Lett 100(22):6
43. Rappel WJ, Levine H (2008) Receptor noise limitations on chemotactic sensing. Proc Natl Acad Sci USA 105(49):19270
44. Tkačik G, Walczak A, Bialek W (2009) Optimizing information flow in small genetic networks. Phys Rev E 80(3):031920
45. Walczak AM, Tkačik G, Bialek W (2010) Optimizing information flow in small genetic networks. II. Feed-forward interactions. Phys Rev E 81(4):1
46. Gregor T, Tank DW, Wieschaus EF, Bialek W (2007) Probing the Limits to Positional Information. Cell 130(1):153–164
47. Tkacik G, Callan CG, Bialek W (2008) Information flow and optimization in transcriptional regulation. Proc Natl Acad Sci USA 105(34):12265
48. Morishita Y, Iwasa Y (2008) Optimal placement of multiple morphogen sources. Phys Rev E 77(4):1
49. Brandman O, Ferrell JE, Li R, Meyer T (2005) Interlinked fast and slow positive feedback loops drive reliable cell decisions. Science (New York, NY) 310(5747):496
50. Brandman O, Meyer T (2008) Feedback loops shape cellular signals in space and time. Science (New York, NY) 322(5900):390
51. Gardiner C (2004) Handbook of stochastic methods: for physics, chemistry and the natural sciences, 3rd edn. Berlin Heidelberg New York, Springer-Verlag
52. van Kampen N (2007) Stochastic processes in physics and chemistry, 3rd edn. (North-Holland Personal Library) North Holland
53. Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. USA, Wiley-Interscience
54. Ziv E, Nemenman I, Wiggins CH (2007) Optimal signal processing in small stochastic biochemical networks. PloS One 2(10):e1077
55. Levine J, Kueh HY, Mirny L (2007) Intrinsic fluctuations, robustness, and tunability in signaling cycles. Biophys J 92(12):4473
56. Mehta P, Goyal S, Long T, Bassler BL, Wingreen NS (2009) Information processing and signal integration in bacterial quorum sensing. Mol Syst Biol 5(325):325

57. Tostevin F, Ten Wolde P (2009) Mutual information between input and output trajectories of biochemical networks. Phys Rev Lett 102(21):218101
58. Tostevin F, Ten Wolde P (2010) Mutual information in time-varying biochemical systems. Phys Rev E 81(6):061917
59. de Ronde W, Tostevin F, Ten Wolde P (2010) Effect of feedback on the fidelity of information transmission of time-varying signals. Phys Rev E 82(3):031914
60. Libby E, Perkins TJ, Swain PS (2007) Noisy information processing through transcriptional regulation. Proc Natl Acad Sci USA 104(17):7151
61. Houillon A, Bessière P, Droulez J (2010) The probabilistic cell: Implementation of a probabilistic inference by the biochemical mechanisms of phototransduction. Acta Biotheor 58(2–3):103–120
62. Mortimer D, Feldner J, Vaughan T, Vetter I, Pujic Z, Rosoff WJ, Burrage K, Dayan P, Richards LJ, Goodhill GJ (2009) Bayesian model predicts the response of axons to molecular gradients. Proc Natl Acad Sci USA 106(25):10296
63. Mortimer D, Dayan P, Burrage K, Goodhill GJ (2010) Optimizing chemotaxis by measuring unbound – bound transitions. Phys D 239:477
64. Mortimer D, Dayan P, Burrage K, Goodhill GJ (2011) Bayes-optimal chemotaxis. Neural Comput 23(2):336–373
65. Oaten A (1977) Optimal foraging in patches: a case for stochasticity. Theor Population Biol 12(3):263
66. Iwasa Y, Higashi M, Yamamura N (1981) Prey distribution as a factor determining the choice of optimal foraging strategy. Am Naturalist 117(5):710
67. McNamara JM, Green RF, Olsson O (2006) Bayes' theorem and its applications in animal behaviour. Oikos 112(2):243
68. Knill D (1996) Perception as Bayesian inference. Berlin Heidelberg, Cambridge University Press
69. Doya K, Ishii S, Pouget A, Rao RP (eds.) (2006) Bayesian brain: probabilistic approaches to neural coding. The MIT Press
70. Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends Cogn Sci 10(7):319
71. Kobayashi TJ (2010) Implementation of dynamic Bayesian decision making by intracellular kinetics. Phys Rev Lett 104(22):228104
72. Kobayashi TJ, Kamimura A (2011) Dynamics of intracellular information decoding. Phys Biol 8(5):055007
73. Kobayashi TJ, Connection between noise-induced symmetry breaking and an information-decoding function for intracellular networks. Phys Rev Lett 106:228101
74. Donaldson-Matasci MC, Bergstrom CT, Lachmann M (2010) The fitness value of information. Oikos 119(2):219
75. Taylor S, Tishby N, Bialek W (2007) Information and fitness. Arxiv preprint arXiv:0712.4382
76. Paulsson J (2004) Summing up the noise in gene networks. Nature 427(6973):415
77. Shibata T (2005) Noisy signal amplification in ultrasensitive signal transduction. Proc Natl Acad Sci USA 102(2):331
78. Lestas I, Vinnicombe G, Paulsson J (2010) Fundamental limits on the suppression of molecular fluctuations. Nature 467(7312):174
79. Bennett MR, Hasty J (2009) Microfluidic devices for measuring gene network dynamics in single cells. Nat Rev Genet 10(9):628
80. Muzzey D, van Oudenaarden A (2009) Quantitative time-lapse fluorescence microscopy in single cells. Ann Rev Cell Dev Biol 25:301
81. Leibler S, Kussell E (2010) Individual histories and selection in heterogeneous populations. Proc Natl Acad Sci USA 107(29):13183–13188

Chapter 17
Zooming in on Yeast Osmoadaptation

Clemens Kühn and Edda Klipp

Abstract *Saccharomyces cerevisiae* is considered as a model organism for the investigation of cellular and molecular processes and gene regulation. Specifically, the response of *S. cerevisiae* to increase in osmolarity of the external medium (osmoadaptation) is a model adaptation process. The first mathematical model of volume changes in *S. cerevisiae* due to osmolarity has been proposed as early as 1983 by Schwartz and Diller (Cryobiology 20(5):542–552). Since then, both experimental and computational methods in biology have progressed dramatically. Especially in recent years, the study of response to hyperosmotic stress in *S. cerevisiae* by systems biology approaches has advanced rapidly. However, a holistic understanding of osmoadaptation combining environmental conditions, cellular preconditions, biophysical processes, molecular and biochemical network dynamics, has not yet been reached. Here, we review recent advances in the investigation of different aspects of osmoadaptation and discuss them with respect to an integrated view. This leads us to critically evaluate how to approach the goal of such an integrated view.

1 Introduction

Understanding the osmoadaptation of eukaryotic cells and specifically of *Saccharomyces cerevisiae* is a longstanding goal that has been explicitly expressed by Hohmann in 2002 [31]. Osmoadaptation describes that cells in liquid media have to maintain their volume by equilibrating external and internal osmolarities. In *S. cerevisiae*, the internal osmolarity is maintained at a higher level than the external osmolarity. The resulting gradient leads to a swelling of the membrane enclosed

C. Kühn • E. Klipp (✉)
Theoretical Biophysics, Humboldt-Universität zu Berlin, Invalidenstr. 42,
D-10115 Berlin, Germany
e-mail: clemens.kuehn@biologie.hu-berlin.de; edda.klipp@biologie.hu-berlin.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_17,

cell volume. The outermost layer of a yeast cell is the cell wall, a structure that is more rigid than the membrane. Because of the higher intracellular osmolarity, the membrane is pushed against the cell wall. The causative difference of internal and external osmotic pressures is termed as turgor pressure. Via changes in turgor pressure, cells can maintain their volume in the face of weak perturbations of osmolarities [6].

Strong perturbations in external osmolarity as induced, for example, by the addition of salt or polyols to the medium increase the extracellular osmolarity and induce cell shrinkage through water efflux. To maintain a viable and optimal volume and to redirect water flow, cells have to increase their internal osmolarity. They accumulate compatible solutes. The solutes accumulated is species and growth condition specific, compatible solutes include, for example, potassium ions, glutamate, and trehalose in *E. coli* [14] and glycerol in *S. cerevisiae* [6]. In principle, any ion or metabolite that can be accumulated without causing toxicity is a possible compatible solute.

In our model organism *S. cerevisiae*, the adaptation to an increase in extracellular osmolarity, or a hyperosmotic shock, has been studied extensively, making osmoadaptation in *S. cerevisiae* a biological model system. We will only give a short overview over the classical view on osmoadaptation in *S. cerevisiae* here (see Fig. 17.1). A detailed review on the relevant biology can be found in [31].

The main signal in osmoadaptation is mediated by Hog1 [9], a stress activated protein (SAP) kinase activated by two signaling branches: The Sln1 branch (consisting of Sln1, Ypd1, Ssk1, Ssk2) and the Sho1 branch (consisting of Msb2, Cdc42, Ste20, Ste50, Sho1, Ste11). The two branches converge in the activation of Pbs2 that in turn activates Hog1. Active Hog1 induces glycerol accumulation via translocation to the nucleus and activation of transcription of GPD1 and regulation of the expression of other genes. Additionally, glycerol accumulation is facilitated by closure of the Fps1 channel protein upon stress. These mechanisms have been extensively studied and are summarized in Fig. 17.1. Recently, additional mechanisms have gained attention that we will introduce in the following sections.

Although the above sketch of the osmoadaptation system is very brief, it already stretches diverse fields of biological research as signal transduction, glycolysis, and gene regulation. A major task besides accumulating knowledge on individual processes contribution to osmoadaptation is to gain insight into their interplay, to integrate the available knowledge into a comprehensive picture of osmoadaptation [31]. Accordingly, we will evaluate the progress towards an integrative view in the latter sections.

Here, we will first summarize the experimental and theoretical frameworks used, then we will take the different scientific perspectives of the presented research. Is osmoadaptation perfect and what is a good model to highlight this? What is the role of physical forces and how to properly calculate them? Is the wiring and kinetics of the Hog1 signaling pathway as well as its interaction with other pathways understood? How is the production of osmolytes regulated? How does each perspective contribute to a comprehensive view?

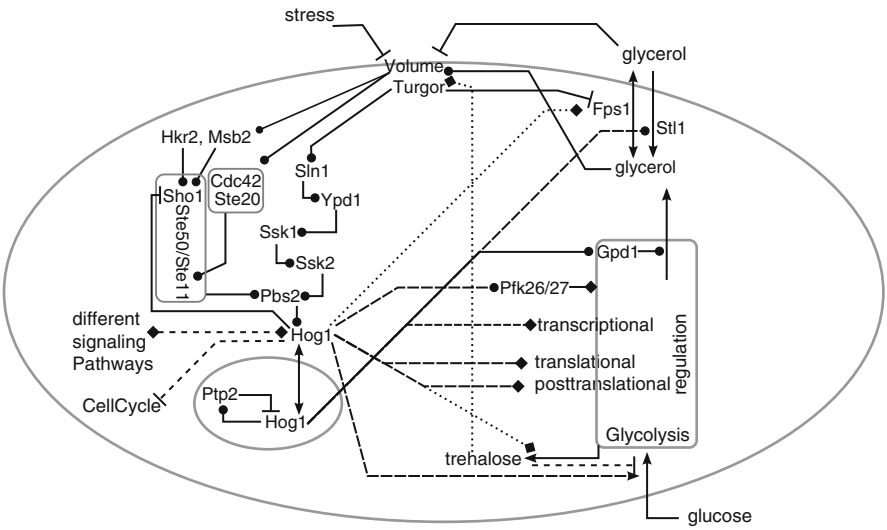


Fig. 17.1 Osmoadaptation in *S. cerevisiae*. Arrows indicate mass flow, diamonds indicate regulation, circles indicate activation, bars indicate inhibition. Solid lines indicate mechanisms observed and modeled, dashed lines indicate mechanisms observed but not modeled, dotted lines indicate presumed mechanisms that could not be mechanistically explained. See text for details

2 Setting the Scene: Experimental and Theoretical Frameworks

Before we consider relevant findings, we want to shortly introduce the experimental and theoretical approaches used in the study of osmoadaptation so far.

Osmoadaptation has been extensively studied using classical approaches such as enzyme assays, RT-qPCR, and western blotting for the observation of relevant molecules. Recently, microfluidics combined with microscopy have been introduced, refining the temporal resolution of experiments and enabling the monitoring of localized concentrations and cell volumes [17, 54]. As detailed in the following sections, the interplay of osmoadaptation with other cellular systems is achieving increased recognition. In order to generate data that allows integration of this biological knowledge into quantitative models, time resolved omics experiments are necessary.

Most of the mathematical models of osmoadaptation are based on ordinary differential equations (ODE). Other approaches include, for example, the use of Bayesian networks [24] and agent based models [42]. ODE models can be roughly divided into two different classes of approaches. The first originates from control theory and engineering (e.g., [49]): A (often small) model is constructed from basic engineering building blocks that are related to biological mechanisms. The mapping of model entities and biological entities is not always straightforward,

but these models allow for a rigid analysis of global system dynamics. The second approach employs classical kinetic modeling. Rate laws are employed to describe state transitions of biological entities (e.g., [39]). The mapping of model entities and biological entities appears straightforward and allows for the description of a wide range of biological perturbations, for example, gene knockouts. Kinetic models often employ a large number of parameters that can not be measured in vivo and hence require sound amounts of reliable experimental data for parametrization.

3 Systems Biology of Adaptation

Before understanding the details of osmoadaptation, it is imperative to understand the global behavior of yeast cells upon hyperosmotic stress. An understanding of the global dynamics can be achieved using models that are radical abstractions of the real system. Such models are convenient to handle and parametrize with only limited data while enabling a rigid analysis of the dynamical features.

Distinct global aspects of signal transduction in osmoadaptation have been investigated in model based studies of the van Oudenaarden group [49, 54]. The most prominent example is the analysis of perfect adaptation of Hog1 nuclearization employing a simplistic model [54]. This is not only an elegant exercise in applying control theory to biological processes, it also yields important results on the architecture of the adaptation process in general. The simplistic nature of the model allows for a rigid and intelligible analysis and determination of the key biological aspects, e.g., that the integration of the biological signal occurs in Hog1-dependent steps, and also allows the temporal classification of osmoadaptation. This is facilitated by the strict experimental focus on monitoring only volume, Hog1 nuclear enrichment, and intracellular glycerol. On the other hand, such a simplistic analysis omits important biological detail and does not generate predictions on biological interactions poorly characterized in experimental data. One arguable finding is that osmoadaptation exhibits perfect adaptation. In this model, the question is easy to answer, because the signal (Hog1 nuclear enrichment) does exhibit perfect adaptation. In a comprehensive view, it is arguable, however, whether the maintenance of higher intracellular solute concentrations does not come at an increased cost for the cell, contradicting perfect adaptation.

Although the aforementioned examples are presumably the most prominent applications of control theory to yeast osmoadaptation, they were preceded by another study [25]. The presented model does not simplify osmoadaptation as radical and hence contains more unknown parameters. The assignment of values to these variables is done with great care and can serve as an example of good practice.

Simplistic models can also be used for integration of different data sets, as presented in [96]. The model allows to thoroughly estimate its parameters and

reproduces data from different sources [27, 39, 47, 49] and hence allows for solid predictions and analysis of global aspects of osmoadaptation, such as signal gain and response to repeated stresses.

4 Systems Biology of Biophysical Aspects

For a more detailed dissection of osmoadaptation, we start with a fundamental aspect: its control principles cannot be fully understood without considering the biophysical laws that govern the dynamics of volume changes, since thermodynamic forces are essential for the activation of cellular signaling and for reestablishing balanced osmolarity.

Changes in cell volume due to hyperosmotic shock can be assessed using, e.g., microfluidic devices, usually inferred from measurements of cell diameter [17]. In plant cells, turgor pressure can be measured using miniaturized pressure probes [68, 69], but yeast cells are too small to be amendable to this technique.

In [39], a model of the signaling pathway and its effect on carbohydrate metabolism has been combined with a description of changes in volume, internal osmotic pressure, and turgor pressure. Here, a dependency of the activity of the receptor Sln1 and the aqua(glycerol)porin Fps1 on turgor had been postulated. This resulted in a model yielding plausible prediction, yet is not confirmed experimentally.

To challenge the effect of turgor on components of the molecular network further, Schaber et al. [67] combined modeling and experimentation to test different hypotheses on the dynamics of volume and turgor changes and their relation to Hog1 activation. In a first step, the authors propose different feasible models of turgor pressure. In a second step, the models are fitted to four data sets of minimal cell volumes after hyperosmotic stresses of various strengths. These data sets have been measured in different labs, under different conditions, and with different methods. Based on the achieved fits, models are selected according to the Akaike information criterion [1]. The selected model for turgor and volume dynamics selected fits all four data sets.

Although turgor pressure could not be measured directly, the authors present a rigid test on different models of turgor pressure in yeast cells and are able to derive further biophysical properties and properties of Hog1 activation from comparably simple data sets. This study provides a solid foundation for the description of turgor pressure in other models and exemplifies how efficient combination of experiments and modeling can lead to important new findings even without omics type of data.

Despite this recent advance, the biophysical aspects of volume maintenance remain somewhat of an enigma. This is especially true for cell volume, the key feature of osmoadaptation. Although cell volume can be measured directly, it is not at all obvious when adaptation is finished and normal cell growth becomes the main driving force of volume changes. This is important in detailed models of

osmoadaptation as described later and for arguing for or against perfect adaptation in osmoadaptation as described in [54]. Ion concentrations and changing solvation properties are also not yet taken into account.

5 Systems Biology of Signal Transduction

In the models considered so far, osmoadaptation is largely divided into a cellular signal and a response to that signal. In this section, we will zoom into the activation of the cellular signal Hog1. Hog1 is activated by two converging MAP Kinase cascades. The next section is concerned with the biochemical network responsible for the cellular response.

Hog1 is activated by a stress (or mitogen) activated protein (SAP or MAP) kinase cascade, which in turn is activated by two upstream branches, the Sln1 branch establishing a phosphorelay system and the Sho1 branch providing a SH3 anchor that ensures proximity of Pbs2 to upstream modifiers. In [39], the architecture and dynamics of the Sln1 branch are considered.

Characteristics of MAP Kinase cascades that are derived from engineering principles have been discussed on the example of Hog1 activation in [30] such as the bandwidth of the pathway and the signal integration of the individual branches.

The necessity of high basal activation of Hog1 and the role of feedback is discussed in Macia et al. [47] based on systematic time-course experiments. The high basal activity proposed here is in contrast to most other studies which assume low or almost vanishing levels of double phosphorylated Hog1 in the absence of stress.

There have been substantial findings concerning the molecular interactions inside the signaling pathways not integrated into formal mathematical models: Additional upstream sensors have been identified by Tatebayashi et al. [76]; mechanistic reason for signaling specificity by determination of dedicated binding sites have been identified for Pbs2 [74] and Hog1 [53]; details on specific steps in the signaling cascade, namely the role of Ssk1 dimerization have been determined [33]; the role of complexes is investigated in experimental studies, both with respect to signal transduction in the Hog1 pathway [75] and to specificity and crosstalk [92, 93]; the role of localization in Hog1 signaling can be assessed using mutations that tether proteins to the membrane [86, 87].

This large body consisting of often qualitative data can not easily be integrated into a quantitative model since the large number of possible interactions and phosphorylation sites bears the problem of combinatorial explosion. Nevertheless, it can be summarized using formal frameworks that allow for the unambiguous yet intelligible representation of individual molecular interactions [42].

Although the quantitative data to parametrize detailed models of osmoadaptation is not available, models can supplement experimental studies as exemplified in the work of Hao et al. [27]: Given a data set for different stimuli or perturbations, which model topologies can reproduce experimental data? In [27], the authors

could identify a feedback inhibition of upstream signaling components as a required network motif which they could confirm in experiments.

When dealing with cellular decision making, the interplay of different signaling cascades has gained attention in recent years [48, 60, 64, 84, 97]. Because of the good understanding achieved on Hog1 signaling and its many interfaces to different cellular processes [18], it is often used as one of the pathways under study for potential crosstalk. This crosstalk might seem circumstantial at first glance, but it is essential to consider in an integrative view of osmoadaptation because it determines cellular response to simultaneous stimuli, a situation that is, presumably, more common to free living organisms than to laboratory strains.

6 Systems Biology of Glycolytic Regulation

According to current literature, the main effector of osmoadaptation upon activation of Hog1 in batch cultures is intracellular glycerol. Glycerol is a by-product of glycolysis branching off at dihydroxy-acetone phosphate (DHAP) and its basal production in unstressed cells is strongly dependent on medium and growth stage [56]. Most approaches focus on one experimental setup to circumvent accounting for differences in glycolytic regulation.

Glycerol accumulation is regulated by glycerol efflux through Fps1 [72] and increased production by transcriptional activation of Gpd1 [23, 65]. Additionally, overall glycolysis undergoes Hog1 dependent and independent changes [57]. In the following, we will discuss important advances towards the understanding of these regulatory processes and highlight which questions remain (yet) unanswered.

The first model of osmoadaptation to take glycolysis into account was presented in 2005 [39]. Here, glycolysis is part of an integrative model of osmoadaptation. Although the model highlights the importance of glycolysis in osmoadaptation and provides mechanistic explanations for the adaptation processes in detail, the parameters presented are not fitted to a metabolite data.

Based on the observation that osmoadaptation reduces cell growth, Parmar et al. [62] have constructed a large model of osmoadaptation integrating cell growth and additional layers of glycolytic regulation [50]. Although the model is based on less experimental data than [39], the authors use it to predict the growth speed decrease for different stresses.

The generation of extensive metabolite data and the fit of mathematical models to these data sets (also exemplified in [80]) is indeed a bottleneck in constructing models of osmoadaptation that account for the glycolytic aspects of osmoadaptation. Hence, most published models on osmoadaptation including glycolysis yet lack extensive experimental data to which parameters could be fitted.

Following the finding that Pfk26/27 has a regulatory role in osmoadaptation, Kühn et al. [41] presented a model to assess this role more in detail. The authors conclude that an activation of upstream glycolysis is important not only to sufficiently accumulate glycerol but also to maintain pyruvate production and

subsequent energy generating reactions in glycolysis in face of increased demand for glycerol production. The question how this regulation is achieved remains unanswered and the model is not fitted to extensive metabolite data.

Besides, additional regulatory mechanisms that contribute to osmodependent regulation of glycolysis are being uncovered in experimental studies that have not been integrated into models yet. New details on the known mechanisms of osmoadaptation are generated, for example on the role, regulation, and localization of Gpd1 [36, 61, 82, 87]. Additional components of glycolysis such as sugar transporters are taken into account [26] and new techniques including phosphoproteomics are being applied [2, 22].

The integration of the novel insights in glycolytic regulation upon hyperosmotic stress into a comprehensive model of glycolysis is extremely difficult (see, for example, [10, 80] for difficulties in modeling glycolysis in itself). One method that could help to understand is the dissection of metabolic changes into a transcriptional component and an allosteric component [11, 20] that has been applied to yeast osmoregulation [7]. This view might, however, occlude additional aspects of glycolytic regulation also implicated in osmoregulation as post-transcriptional and translational changes [55, 85].

Complex in itself, glycolysis can not be considered without accounting for the general state and requirements of cells (e.g., maintaining energy production even when adapting to stress). Here, indirect regulatory effects are important as discussed in [50, 80]. Another relevant topic not discussed so far is the influence of the environment on cellular state. The role of additional metabolites implied in osmoadaptation (e.g., trehalose as discussed in [71]) gives rise to additional adaptation mechanisms that have to be discussed, but also to new layers of glycolytic regulation [79]. Possible roles of trehalose, glycolytic regulation, and the influence of cellular redox state on osmoadaptation have been reviewed [5] but have not been integrated into models.

7 Systems Biology of Glycerol Transport

Intracellular Glycerol accumulation is not possible without glycerol retention. The Fps1 channel protein closes upon hyperosmotic stress [46]. Although the dynamics of Fps1 closure upon hyperosmotic stress have been characterized by Tamàs et al. [72, 73], the mechanisms of closure are not fully understood, albeit intensely studied [4]. Under different stress conditions, Fps1 transport is regulated by Hog1 activity [51, 81, 89], hinting towards an interaction between Hog1 and Fps1 further implicated in additional studies [51, 52]. Although a direct interaction could not be observed in osmoadaptation, a strain in which Hog1 is tethered to the membrane [87] could be used to infer whether proximity of Hog1 and Fps1 and resulting induced regulatory activity affects osmoadaptation.

Besides passive transport through Fps1, *S. cerevisiae* can actively take up glycerol from the external medium through Stl1 [21]. Although not essential for osmoadaptation, this mechanism might contribute under conditions not tested so far.

8 Towards an Integrative View

Hohmann described the integrative view on osmoadaptation as "comprehensive view on the time line, spatial dynamics, interaction, and mutual dependency of the underlying cellular events" in the "foreseeable future" covering the areas of "control of transmembrane transport, the sensing of osmotic changes, the mechanisms, dynamics, and spatial organization of signal transmission, metabolic adjustments, the effects on the cytoskeleton, cell cycle progression, translation, and cell wall dynamics" [31].

So far, we have described recent advances in the study of more or less individual mechanisms that contribute to osmoadaptation. Except for two models [39, 62] that integrate biological knowledge but lack experimental data on large parts of the system, no dedicated advances towards an integrated view have been published so far. Figure 17.2 visualizes the coverage of different aspects of osmoadaptation in modeling studies. Have we lost sight of the integrative goal? Or are we on the right track? And what needs to be achieved on the road to an integrative view?

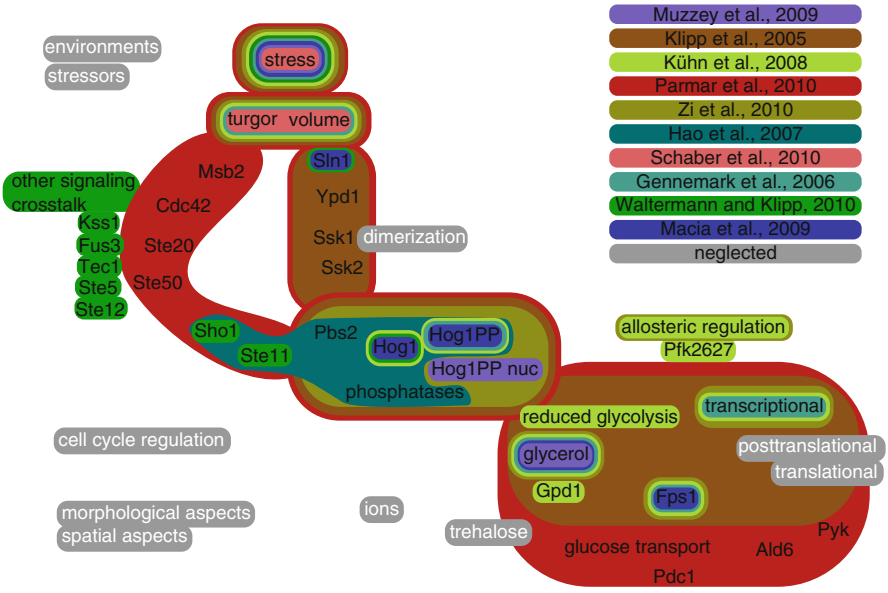


Fig. 17.2 Coverage of osmoadaptation mechanisms in formal models

To answer these questions, we must first consider what a comprehensive view exactly is. In contrast to the focused view of many individual approaches in the field, we suggest two different comprehensive views: Either a formal knowledge collection in the form of a web-based repository (including, for example, rule-based formalizations of biological knowledge [42]) or a comprehensive, detailed, dynamic, and quantitative model.

Currently, it is technically not possible to sensibly merge all models described here into a comprehensive model. The emergence of standards and tools for modeling in biology [34, 40, 43] and an increasing awareness of publishers to require well annotated models to be submitted to model databases (like BioModels [45] or JWS online [58]) indicate that in the foreseeable future, models can be efficiently merged. Also, the collaborative effort UNICELLSYS [32] is an effort to construct a comprehensive model of yeast cells. A different approach to construct a comprehensive model is to use an existing detailed model, e.g., [39,62], and improve details of these models in a collaborative way, using the large model as a scaffold.

The same holds for experimental data. Increasingly, data sets are made publicly available and their interpretation is facilitated by annotation projects [8, 77, 78]. Even if merging of data sets and models will be technically feasible, the different experimental settings used for data generation need to be accounted for. This either requires good knowledge of the cellular states in the setting used or a comprehensive data set obtained under controlled conditions.

Accordingly, a comprehensive model requires the integration of biological knowledge, experimental data, and mathematical models into a formal framework. The formal collection required for a comprehensive model is in itself a comprehensive view. It also offers more flexibility concerning the direction and revision of the current view. This is important, because, although osmoadaptation has been extensively studied, we are still facing a many open questions that must be answered before a comprehensive view can be obtained.

9 Open Questions

Experimental data is, to some extent, available on all points mentioned in [31]. Nevertheless, our knowledge of the integration of osmoadaptation with other cellular process is not sufficient and must be one of the main fields of research to obtain a comprehensive view.

Experimental evidence on interactions between osmoadaptation and other cellular pathways exist, but further data and new models are necessary to quantitatively formalize this evidence and put it into context. The most important links are links between cell cycle and osmoadaptation [3, 12, 13, 18, 19, 90, 91, 95], crosstalk with other signaling pathways prioritizing cellular response in face of multiple stresses [84] not only on the level of signaling but also on the level of downstream effectors, and to a more detailed and versatile glycolysis.

The faithful mathematical description of yeast glycolysis has been an active topic of research for decades [10, 28, 29, 35, 66, 80]. The complexity in the regulation of glycolysis has yet withstood an efficient formalization of an adaptive glycolysis. But regulation of glycolysis is a major effector of osmoadaptation and is regulated both transcriptionally, translationally, and through regulation of enzyme activity and stability, hence requiring to include all levels of regulation of glycolytic flux into a comprehensive model of osmoadaptation.

Besides established osmoadaptation-linked processes, neglected pathways require attention. This is the case for the role of trehalose [15, 37, 63, 70, 71, 88, 94] and the role of ions in osmoadaptation. Additionally, new processes involving Hog1 signaling are being uncovered [16, 38].

The aspect of an integrative view formulated in [31] for which our knowledge, up to now, is still incomplete is the role of spatial aspects of osmoadaptation. Spatial dynamics are presumably important on all levels of osmoadaptation, from biophysical changes (do cells shrink symmetrically? Does this affect activation of turgor-sensing signaling molecules?) to (co-)localization in signal transduction and metabolism (does molecular crowding boost signal transduction?) and general cellular organization (How important is the incompressible volume?). Spatial effects might also contribute to the dynamics of Fps1, as Fps1 seems to form patches on the membrane upon hyperosmotic stress [51].

Yet another topic that has received only limited attention so far is long term adaptation. Cells exposed to enduring hyperosmotic conditions do reorganize glycolysis and decrease growth [57, 59, 62]. Especially the transient nature of glycerol accumulation and the late (and also transient) trehalose accumulation [71] is generally neglected. Whether and how morphological changes, for example, actin reorganization [38], are involved in a long term response is also unknown.

10 Conclusion

We have summarized recent findings, discussed how they contribute to a comprehensive view of osmoadaptation and which aspects of osmoadaptation remain enigmatic. We have also argued that a comprehensive view is not necessarily one giant dynamic model but could be a practical comprehensive view that can also be obtained in collecting knowledge, data, and models.

Formally, a comprehensive view of osmoadaptation requires a community platform, additional biological knowledge, standardized quantitative data, and modular mathematical models. A similar approach is used to construct a comprehensive model for *S. cerevisiae*, in general, in the UNICELLSYS project [32], providing also a top level into which an integrated view of osmoadaptation can be embedded.

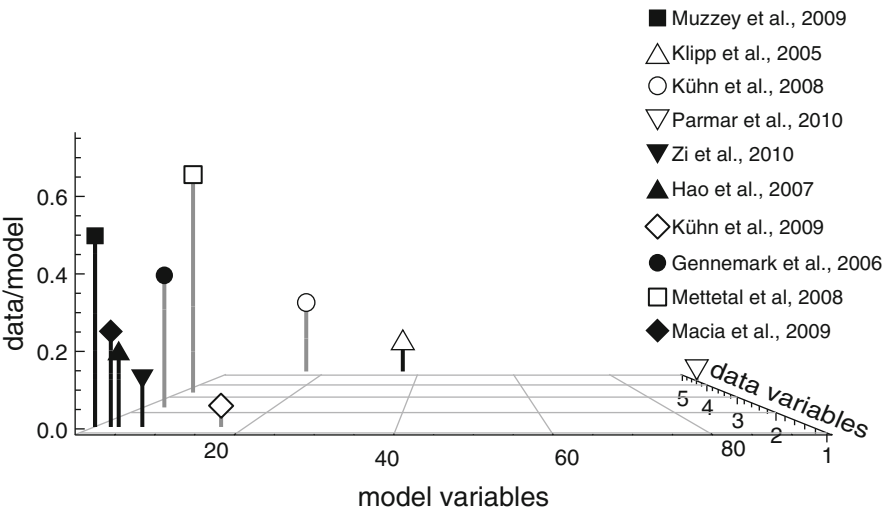


Fig. 17.3 Visualization of model diversity by comparison of the number of model variables and the number of variables measured. Black stems indicate that variables were measured under different conditions.This comparison is a simplification, neither variables measured nor the number of data points accounts for the relationship between model structure and experimental data. Hence, a high data/model ratio does not imply a better model per se. The purpose of the model must be taken into account

Biologically, a comprehensive view of osmoadaptation requires additional knowledge on interactions with cell cycle and glycolytic regulation, an increase in spatial resolution of biophysical, signaling and glycolytic dynamics, and the integration of hitherto neglected aspects.

Since models are useful abstractions of reality and different models cover different aspects in varying detail, modularity is the key feature to integrate mathematical models. Modular modeling facilitate the integration of diverse models into a comprehensive model using submodels that communicate via defined interfaces and adhere to respective standards (e.g., MIRIAM [44] or MIASE [83]). Each module could be studied in detail and, a key step, simplified. Abstract models like the control circuit presented in [54] that reliably reproduce the dynamics of a given module could then be used to describe processes that are not essential to a given research question but are essential to consider when studying this question in the context of a comprehensive view. This approach effectively yields a zoomable comprehensive model and would both strengthen and gain from a diversity of models (as exemplified in Fig. 17.3).

Acknowledgements EK is supported by UNICELLSYS (European Commission 7th Framework Programme: Contract No. 201142)

References

1. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
2. Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H (2008) A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol Cell Proteomics 7(7):1389–1396
3. Alexander MR, Tyers M, Perret M, Craig BM, Fang KS, Gustin MC (2001) Regulation of cell cycle progression by swe1p and hog1p following hypertonic stress. Mol Biol Cell 12(1):53–62
4. Beese SE, Negishi T, Levin DE (2009) Identification of positive regulators of the yeast fps1 glycerol channel. PLoS Genet 5(11):e1000738
5. Blomberg A (2000) Metabolic surprises in *Saccharomyces cerevisiae* during adaptation to saline conditions: questions, some answers and a model. FEMS Microbiol Lett 182(1):1–8
6. Blomberg A, Adler L (1992) Physiology of osmotolerance in fungi. Adv Microb Physiol 33:145–212
7. Bouwman J, Kiewiet J, Lindenbergh A, van Eunen K, Siderius M, Bakker BM (2011) Metabolic regulation rather than de novo enzyme synthesis dominates the osmo-adaptation of yeast. Yeast 28(1):43–53
8. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (miame)-toward standards for microarray data. Nat Genet 29(4):365–371
9. Brewster JL, de Valoir T, Dwyer ND, Winter E, Gustin MC (1993) An osmosensing signal transduction pathway in yeast. Science 259(5102):1760–1763
10. Bruck J, Liebermeister W, Klipp E (2008) Exploring the effect of variable enzyme concentrations in a kinetic model of yeast glycolysis. Genome Inform 20:1–14
11. Bruggeman F, De Haan J, Hardin H, Bouwman J, Rossell S, Van Eunen K, Bakker B, Westerhoff H (2006) Time-dependent hierarchical regulation analysis: deciphering cellular adaptation. IEE Proc Syst Biol 153(5):318
12. Clotet J, Posas F (2007) Control of cell cycle in response to osmostress: lessons from yeast. Methods Enzymol 428:63–76
13. Clotet J, Escoté X, Adrover MA, Yaakov G, Garí E, Aldea M, de Nadal E, Posas F (2006) Phosphorylation of hsl1 by hog1 leads to a g2 arrest essential for cell survival at high osmolarity. EMBO J 25(11):2338–2346
14. Dinnbier U, Limpinsel E, Schmid R, Bakker EP (1988) Transient accumulation of potassium glutamate and its replacement by trehalose during adaptation of growing cells of escherichia coli k-12 to elevated sodium chloride concentrations. Arch Microbiol 150(4):348–357
15. Elbein AD, Pan YT, Pastuszak I, Carroll D (2003) New insights on trehalose: a multifunctional molecule. Glycobiology 13(4):17R–27R
16. Eraso P, Mazón MJ, Posas F, Portillo F (2011) Gene expression profiling of yeasts overexpressing wild type or misfolded pma1 variants reveals activation of the hog1 mapk pathway. Mol Microbiol 79(5):1339–1352
17. Eriksson E, Enger J, Nordlander B, Erjavec N, Ramser K, Goksör M, Hohmann S, Nyström T, Hanstorp D (2007) A microfluidic system in combination with optical tweezers for analyzing rapid and reversible cytological alterations in single cells upon environmental changes. Lab Chip 7(1):71–76
18. Escoté X, Zapater M, Clotet J, Posas F (2004) Hog1 mediates cell-cycle arrest in g1 phase by the dual targeting of sic1. Nat Cell Biol 6(10):997–1002
19. Escoté X, Miranda M, Rodríguez-Porrata B, Mas A, Cordero R, Posas F, Vendrell J (2011) The stress-activated protein kinase hog1 develops a critical role after resting state. Mol Microbiol 80(2):423–435

20. van Eunen K, Bouwman J, Lindenbergh A, Westerhoff HV, Bakker BM (2009) Time-dependent regulation analysis dissects shifts between metabolic and gene-expression regulation during nitrogen starvation in baker's yeast. FEBS J 276(19):5521–5536
21. Ferreira C, van Voorst F, Martins A, Neves L, Oliveira R, Kielland-Brandt MC, Lucas C, Brandt A (2005) A member of the sugar transporter family, stl1p is the glycerol/h+ symporter in *Saccharomyces cerevisiae*. Mol Biol Cell 16(4):2068–2076
22. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nat Biotechnol 20(3):301–305
23. Gasch A (2007) Comparative genomics of the environmental stress response in ascomycete fungi. Yeast 24(11):961–976, DOI 10.1002/yea
24. Gat-Viks I, Shamir R (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast. Genome Res 17(3):358–367
25. Gennemark P, Nordlander B, Hohmann S, Wedelin D (2006) A simple mathematical model of adaptation to high osmolarity in yeast. In Silico Biol 6(3):193–214
26. Greatrix BW, van Vuuren AHJJ (2006) Expression of the hxt13 , hxt15 and hxt17 genes in *Saccharomyces cerevisiae* and stabilization of the hxt1 gene transcript by sugar-induced osmotic stress. Curr Genet 49(4):205–217
27. Hao N, Behar M, Parnell SC, Torres MP, Borchers CH, Elston TC, Dohlman HG (2007) A systems-biology analysis of feedback inhibition in the sho1 osmotic-stress–response pathway. Curr Biol 17(8):659–667
28. Heinrich R, Rapoport TA (1974) A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. Eur J Biochem 42(1):89–95
29. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BO, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat biotechnol 26(10):1155–1160
30. Hersen P, McClean MN, Mahadevan L, Ramanathan S (2008) Signal processing by the hog map kinase pathway. Proc Natl Acad Sci USA 105(20):7165–7170
31. Hohmann S (2002) Osmotic stress signaling and osmoadaptation in yeasts. Microbiol Mol Biol Rev 66(2):300
32. Hohmann S (2010) Unicellsys – understanding the cells functional organization. J Biotechnol 150:545
33. Horie T, Tatebayashi K, Yamada R, Saito H (2008) Phosphorylated ssk1 prevents unphosphorylated ssk1 from activating the ssk2 mitogen-activated protein kinase kinase kinase in the yeast high-osmolarity glycerol osmoregulatory pathway. Mol Cell Biol 28(17):5172–5183
34. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J, SBML Forum (2003) The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4):524–531
35. Hynne F, Danø S, Sørensen PG (2001) Full-scale model of glycolysis in *Saccharomyces cerevisiae*. Biophys Chem 94(1–2):121–163
36. Jung S, Marelli M, Rachubinski Ra, Goodlett DR, Aitchison JD (2010) Dynamic changes in the subcellular distribution of gpd1p in response to cell stress. J Biol Chem 285(9):6739–6749
37. Kandror O, Bretschneider N, Kreydin E, Cavalieri D, Goldberg AL (2004) Yeast adapt to near-freezing temperatures by stre/msn2,4-dependent induction of trehalose synthesis and certain molecular chaperones. Mol Cell 13(6):771–781

38. Kim S, Shah K (2007) Dissecting yeast hog1 map kinase pathway using a chemical genetic approach. FEBS Lett 581(6):1209–1216
39. Klipp E, Nordlander B, Krüger R, Gennemark P, Hohmann S (2005) Integrative model of the response of yeast to osmotic shock. Nat Biotechnol 23(8):975–982
40. Krause F, Uhlendorf J, Lubitz T, Schulz M, Klipp E, Liebermeister W (2010) Annotation and merging of sbml models with semanticsbml. Bioinformatics 26(3):421–422
41. Kühn C, Petelenz E, Nordlander B, Schaber J, Hohmann S, Klipp E (2008) Exploring the impact of osmoadaptation on glycolysis using time-varying response-coefficients. Genome Inform 20:77–90
42. Kühn C, Prasad KVS, Klipp E, Gennemark P (2010) Formal representation of the high osmolarity glycerol pathway in yeast. Genome Inform 22:69–83
43. Laibe C, Le Novère N (2007) Miriam resources: tools to generate and resolve robust cross-references in systems biology. BMC Syst Biol 1:58
44. Le Novère N, Finney A, Hucka M, Bhalla U, Campagne F, Collado-Vides J, Crampin E, Halstead M, Klipp E, Mendes P, et al (2005) Minimum information requested in the annotation of biochemical models (miriam). Nat Biotechnol 23(12):1509–1515
45. Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34(Database issue):D689–D691
46. Luyten K, Albertyn J, Skibbe WF, Prior BA, Ramos J, Thevelein JM, Hohmann S (1995) Fps1, a yeast member of the mip family of channel proteins, is a facilitator for glycerol uptake and efflux and is inactive under osmotic stress. EMBO J 14(7):1360–1371
47. Macia J, Regot S, Peeters T, Conde N, Solé R, Posas F (2009) Dynamic signaling in the hog1 mapk pathway relies on high basal signal transduction. Sci Signal 2(63):ra13
48. McClean MN, Mody A, Broach JR, Ramanathan S (2007) Cross-talk and decision making in map kinase pathways. Nat Genet 39(3):409–414
49. Mettetal JT, Muzzey D, Gómez-Uribe C, van Oudenaarden A (2008) The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. Science 319(5862): 482–484
50. Modig T, Granath K, Adler L, Lidn G (2007) Anaerobic glycerol production by *Saccharomyces cerevisiae* strains under hyperosmotic stress. Appl Microbiol Biotechnol 75(2): 289–296
51. Mollapour M, Piper PW (2007) Hog1 mitogen-activated protein kinase phosphorylation targets the yeast fps1 aquaglyceroporin for endocytosis, thereby rendering cells resistant to acetic acid. Mol Cell Biol 27(18):6446–6456
52. Mollapour M, Shepherd A, Piper PW (2009) Presence of the fps1p aquaglyceroporin channel is essential for hog1p activation, but suppresses slt2(mpk1)p activation, with acetic acid stress of yeast. Microbiology 155(Pt 10):3304–3311
53. Murakami Y, Tatebayashi K, Saito H (2008) Two adjacent docking sites in the yeast hog1 mitogen-activated protein (map) kinase differentially interact with the pbs2 map kinase kinase and the ptp2 protein tyrosine phosphatase. Mol Cell Biol 28(7):2481–2494
54. Muzzey D, Gómez-Uribe Ca, Mettetal JT, van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. Cell 138(1):160–171
55. de Nadal E, Posas F (2010) Multilayered control of gene expression by stress-activated protein kinases. EMBO J 29(1):4–13
56. Nevoigt E, Stahl U (1997) Osmoregulation and glycerol metabolism in the yeast *Saccharomyces cerevisiae*. FEMS Microbiol Rev 21(3):231–241
57. Nordlander B, Krantz M, Hohmann S (2008) Hog1-mediated metabolic adjustments following hyperosmotic shock in the yeast *Saccharomyces cerevisiae*. In: Posas F, Nebreda A (eds) Stress-activated protein kinases, Top Curr Genet, vol 20, Springer Berlin/Heidelberg, pp. 141–158
58. Olivier BG, Snoep JL (2004) Web-based kinetic modelling using jws online. Bioinformatics 20(13):2143–2144

59. Olz R, Larsson K, Adler L, Gustafsson L (1993) Energy flux and osmoregulation of *Saccharomyces cerevisiae* grown in chemostats under nacl stress. J Bacteriol 175(8): 2205–2213
60. O'Rourke SM, Herskowitz I (1998) The hog1 mapk prevents cross talk between the hog and pheromone response mapk pathways in *Saccharomyces cerevisiae*. Genes Dev 12(18): 2874–2886
61. Ou X, Ji C, Han X, Zhao X, Li X, Mao Y, Wong LL, Bartlam M, Rao Z (2006) Crystal structures of human glycerol 3-phosphate dehydrogenase 1 (gpd1). J Mol Biol 357(3): 858–869
62. Parmar JH, Bhartiya S, Venkatesh KV (2011) Characterization of the adaptive response and growth upon hyperosmotic shock in *Saccharomyces cerevisiae*. Mol BioSyst 7(4):1138–1148
63. Parrou JL, Teste MA, François J (1997) Effects of various types of stress on the metabolism of reserve carbohydrates in *Saccharomyces cerevisiae*: genetic evidence for a stress-induced recycling of glycogen and trehalose. Microbiology 143(6):1891–1900
64. Rensing L, Ruoff P (2009) How can yeast cells decide between three activated map kinase pathways? A model approach. J Theor Biol 257(4):578–587
65. Rep M, Albertyn J, Thevelein JM, Prior Ba, Hohmann S (1999) Different signalling pathways contribute to the control of gpd1 gene expression by osmotic stress in *Saccharomyces cerevisiae*. Microbiology 145 (Pt 3):715–727
66. Rizzi M, Baltes M, Theobald U, Reuss M (1997) In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: Ii. mathematical model. Biotechnol Bioeng 55(4):592–608
67. Schaber J, Adrover MA, Eriksson E, Pelet S, Petelenz-Kurdziel E, Klein D, Posas F, Goksör M, Peter M, Hohmann S, Klipp E (2010) Biophysical properties of *Saccharomyces cerevisiae* and their relationship with hog pathway activation. Eur Biophys J 39(11):1547–1556
68. Shabala SN, Lew RR (2002) Turgor regulation in osmotically stressed arabidopsis epidermal root cells. direct support for the role of inorganic ion uptake as revealed by concurrent flux and cell turgor measurements. Plant Physiol 129(1):290–299
69. Shackel KA, Brinckmann E (1985) In situ measurement of epidermal cell turgor, leaf water potential, and gas exchange in *Tradescantia virginiana* l. Plant Physiol 78(1):66–70
70. Singer MA, Lindquist S (1998) Thermotolerance in *Saccharomyces cerevisiae*: The yin and yang of trehalose. Trends Biotechnol 16(11):460–468
71. Singh K, Norton R (1991) Metabolic changes induced during adaptation of *Saccharomyces cerevisiae* to a water stress. Arch Microbiol 156(1):38–42
72. Tamás MJ, Luyten K, Sutherland FC, Hernandez a, Albertyn J, Valadi H, Li H, Prior Ba, Kilian SG, Ramos J, Gustafsson L, Thevelein JM, Hohmann S (1999) Fps1p controls the accumulation and release of the compatible solute glycerol in yeast osmoregulation. Mol Microbiol 31(4):1087–1104
73. Tamás MJ, Rep M, Thevelein JM, Hohmann S (2000) Stimulation of the yeast high osmolarity glycerol (hog) pathway: evidence for a signal generated by a change in turgor rather than by water stress. FEBS Lett 472(1):159–165
74. Tatebayashi K, Takekawa M, Saito H (2003) A docking site determining specificity of pbs2 mapkk for ssk2/ssk22 mapkkks in the yeast hog pathway. EMBO J 22(14):3624–3634
75. Tatebayashi K, Yamamoto K, Tanaka K, Tomida T, Maruoka T, Kasukawa E, Saito H (2006) Adaptor functions of cdc42, ste50, and sho1 in the yeast osmoregulatory hog mapk pathway. EMBO J 25(13):3033–3044
76. Tatebayashi K, Tanaka K, Yang HY, Yamamoto K, Matsushita Y, Tomida T, Imai M, Saito H (2007) Transmembrane mucins hkr1 and msb2 are putative osmosensors in the sho1 branch of yeast hog pathway. EMBO J 26(15):3521–3533
77. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJR, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR, Hermjakob H (2007) The minimum information about a proteomics experiment (miape). Nat Biotechnol 25(8): 887–893

78. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Clark AM, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. Nat Biotechnol 26(8):889–896, DOI 10.1038/nbt.1411
79. Teusink B, Walsh MC, van Dam K, Westerhoff HV (1998) The danger of metabolic pathways with turbo design. Trends Biochem Sci 23(5):162–169
80. Teusink B, Passarge J, Reijenga Ca, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267(17):5313–5329
81. Thorsen M, Di Y, Tangemo C (2006) The mapk hog1p modulates fps1p-dependent arsenite uptake and tolerance in yeast. Mol Biol Cell 17(October):4400–4410
82. Valadi A, Granath K, Gustafsson L, Adler L (2004) Distinct intracellular localization of gpd1p and gpd2p, the two yeast isoforms of nad+-dependent glycerol-3-phosphate dehydrogenase, explains their different contributions to redox-driven glycerol production. J Biol Chem 279(38):39677–39685
83. Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, Chelliah V, Cooling MT, Cooper J, Crampin EJ, Garny A, Hoops S, Hucka M, Hunter P, Klipp E, Laibe C, Miller AK, Moraru I, Nickerson D, Nielsen P, Nikolski M, Sahle S, Sauro HM, Schmidt H, Snoep JL, Tolle D, Wolkenhauer O, Le Novère N (2011) Minimum information about a simulation experiment (miase). PLoS Comput Biol 7(4):e1001122+
84. Waltermann C, Klipp E (2010) Signal integration in budding yeast. Biochem Soc Trans 38(5):1257–1264
85. Warringer J, Hult M, Regot S, Posas F, Sunnerhagen P (2010) The hog pathway dictates the short-term translational response after hyperosmotic shock. Mol Biol Cell 21(17):3080–3092
86. Westfall PJ, Thorner J (2006) Analysis of mitogen-activated protein kinase signaling specificity in response to hyperosmotic stress: use of an analog-sensitive hog1 allele. Eukaryotic Cell 5(8):1215–1228
87. Westfall PJ, Patterson JC, Chen RE, Thorner J (2008) Stress resistance and signal fidelity independent of nuclear mapk function. Proc Natl Acad Sci USA 105(34):12212–12217
88. Wiemken A (1990) Trehalose in yeast, stress protectant rather than reserve carbohydrate. Antonie van Leeuwenhoek 58(3):209–217
89. Wysocki R, Chéry CC, Wawrzycka D, Hulle MV, Cornelis R, Thevelein JM, Tamás MJ (2001) The glycerol channel fps1p mediates the uptake of arsenite and antimonite in *Saccharomyces cerevisiae*. Mol Microbiol 40(6):1391–1401
90. Yaakov G, Bell M, Hohmann S, Engelberg D (2003) Combination of two activating mutations in one hog1 gene forms hyperactive enzymes that induce growth arrest. Mol Cell Biol 23(14):4826–4840
91. Yaakov G, Duch A, Garca-Rubio M, Clotet J, Jimenez J, Aguilera A, Posas F (2009) The stress-activated protein kinase hog1 mediates s phase delay in response to osmostress. Mol Biol Cell 20(15):3572–3582
92. Yamamoto K, Tatebayashi K, Tanaka K, Saito H (2010) Dynamic control of yeast map kinase network by induced association and dissociation between the ste50 scaffold and the opy2 membrane anchor. Mol Cell 40(1):87–98
93. Yang HY, Tatebayashi K, Yamamoto K, Saito H (2009) Glycosylation defects activate filamentous growth kss1 mapk and inhibit osmoregulatory hog1 mapk. EMBO J 28(10):1380–1391

94. Zähringer H, Thevelein JM, Nwaka S (2000) Induction of neutral trehalase nth1 by heat and osmotic stress is controlled by stre elements and msn2/msn4 transcription factors: variations of pka effect during stress and growth. Mol Microbiol 35(2):397–406
95. Zapater M, Clotet J, Escoté X, Posas F (2005) Control of cell cycle progression by the stress-activated hog1 mapk. Cell Cycle 4(1):6–7
96. Zi Z, Liebermeister W, Klipp E (2010) A quantitative study of the hog1 MAPK response to fluctuating osmotic stress in *Saccharomyces cerevisiae*. PLoS One 5(3):e9522
97. Zou X, Peng T, Pan Z (2008) Modeling specificity in the yeast mapk signaling networks. J Theor Biol 250(1):139–155

Part III
Spatial and Temporal Dymensions of Intracellular Dynamics

Chapter 18
Receptor Dynamics in Signaling

Verena Becker, Jens Timmer, and Ursula Klingmüller

Abstract Reliable inter- and intracellular communication is central to both the development and the integrity of multicellular organisms. Key mediators of these processes are cell surface receptors that perceive and convert extracellular cues to trigger intracellular signaling networks and ultimately a phenotypic response. Deregulation of signal transduction leads to a variety of diseases, and aberrations in receptor proteins are very common in various cancer types. Therefore, cell surface receptors have been established as major targets in drug discovery. However, in order to efficiently apply therapeutics, it is crucial to gain knowledge about design principles of receptor signaling. In this chapter, we will discuss signal transduction at the receptor level for examples from different receptor classes.

1 Introduction

Tightly regulated cellular communication is key not only to the development of multicellular organisms but also to the functional integrity of tissues, organs,

V. Becker (✉)
Division Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany

Bioquant, Heidelberg University, Germany Present address: Department of Systems Biology, Harvard Medical School, Boston, MA, USA
e-mail: verena_becker@hms.harvard.edu

J. Timmer
BIOSS Centre for Biological Signalling Studies, Freiburg Institute for Advanced Studies, Institute of Physics, Center for Systems Biology, University of Freiburg, Freiburg, Germany
e-mail: jeti@fdm.uni-freiburg.de

U. Klingmüller
Division Systems Biology of Signal Transduction, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany; Bioquant, Heidelberg University, Germany
e-mail: u.klingmueller@dkfz-heidelberg.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_18,

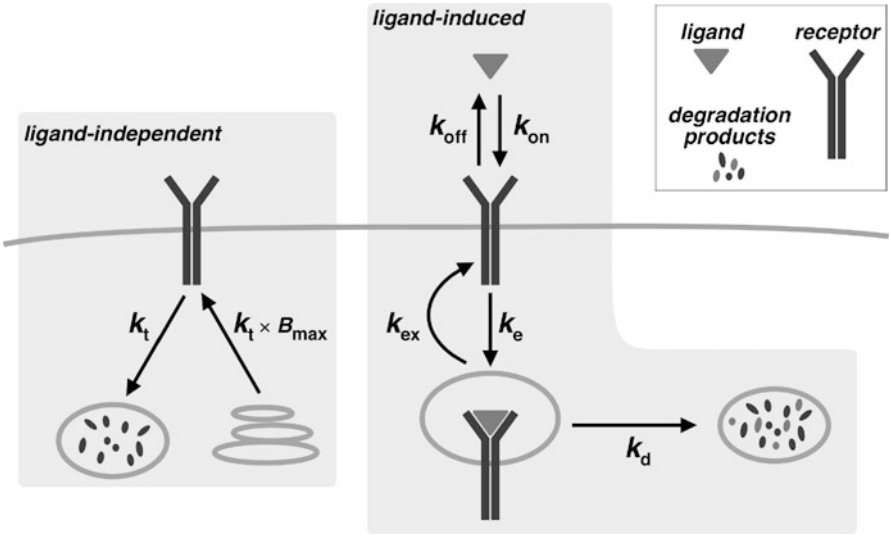


Fig. 18.1 Generalized scheme of ligand and receptor interaction and trafficking processes

and the whole body. There are a plethora of mediators involved in cell-to-cell communications such as small molecules, peptides, cytokines, growth factors, lipid hormones, and physical signals. These molecules bind to specific cell surface receptors, which initiate signal transmission by linking extracellular cues to intracellular cascades of signaling molecules. Integration of different signal transduction networks via crosstalk of intersecting pathways processes the information and finally leads to appropriate phenotypic responses of the cell such as proliferation, differentiation, migration, survival, or apoptosis.

Aberrations in signaling cascades are linked to various disease types including cancer, infections, as well as immunological and metabolic disorders. In the advent of targeted therapeutics, cell surface receptors have become prime objectives in drug discovery [1], and various antibodies impeding ligand binding or small molecule inhibitors interfering with the enzymatic activity of receptor proteins undergo development or are already used in cancer therapy.

However, to efficiently apply targeted therapeutics, it is crucial to understand the complex regulation of the underlying biochemical networks [2–4]. Therefore, the identification of design principles for cell surface receptor signaling holds great promise in furthering rational drug discovery and personalized therapy strategies. Mathematical models have been established to aid the understanding of how ligand–receptor interaction and trafficking shape receptor activation kinetics [5–8]. In a generalized scheme (Fig. 18.1), ligand undergoes binding to receptor proteins with distinct association (k_{on}) and dissociation (k_{off}) rates. Trafficking of receptors can be both ligand-independent and ligand-induced. Receptor transport to the plasma membrane ($k_t \times B_{max}$) can be described by ligand-independent

endocytosis (k_t) and the receptor abundance in the absence of ligand (B_{max}), i.e. at steady state. Endocytosis of ligand–receptor complexes (k_e) can either be followed by recycling (k_{ex}) or by degradation processes (k_d). This generalized model varies with the receptor system under study, and additional processes might be taken into account such as ligand-induced mobilization of newly synthesized receptor from intracellular pools to the plasma membrane.

In this review, we will discuss information processing at the receptor level, exemplified by the erythropoietin receptor (EpoR), the interleukin 3 receptor (IL3R), the epidermal growth factor receptor (EGFR), and the receptor for transforming growth factor beta (TGFβ).

2 Cytokine Receptors

Cytokine receptors are involved in diverse physiological processes such as the development of the hematopoietic system or in pro- as well as anti-inflammatory cellular responses [9, 10]. Members of the cytokine receptor family are single membrane-spanning proteins that lack intrinsic enzymatic activity and, therefore, associate with cytoplasmic Janus kinases (JAK). Mutations that constitutively activate cytokine receptors have been described for a variety of hematological disorders, and they are found either in receptor proteins such as the EpoR [11, 12], the granulocyte colony-stimulating factor (GCSF) receptor [13], and the thrombopoietin receptor [14], or in receptor-associated kinases such as JAK2 [15] and JAK1 [16].

2.1 Erythropoietin Receptor

Erythropoietin (Epo) signaling [17] is crucial for the survival, proliferation, and differentiation of erythroid progenitors at the colony-forming unit-erythroid (CFU-E) stage [18]. Crystallographic studies revealed that the EpoR is expressed as a preformed homodimer [19]. The majority of receptor protein resides in intracellular compartments of the endoplasmic reticulum and the Golgi apparatus as shown for both endogenous EpoR in CFU-E cells as well as exogenous EpoR expression in various cell lines [20–24].

Endocytosis and subsequent degradation of ligand–receptor complexes have been proposed to downregulate EpoR activity [25]. Using a kinetic model, ligand-induced endocytosis could be identified as a mechanism to clear Epo from the extracellular space, and differences in clearance rates between Epo derivatives were assigned to distinct ligand binding rates [26].

By combining time-resolved quantitative data for ligand-independent and ligand-induced endocytosis with ordinary differential equation-based modeling, design principles of EpoR signaling could be further refined [8]. Whereas ligand-induced endocytosis plays a major role in shaping early-response kinetics of EpoR phosphorylation, ligand-independent EpoR turnover at the plasma membrane is crucial

for a linear conversion of extracellular Epo levels into receptor activation. Both computational and experimental evidence showed that intracellular EpoR pools constitute a reservoir for a continuous replenishment of cell surface receptor, a process that is key to linear information processing. While peak levels of EpoR and JAK2 phosphorylation are saturated at higher ligand concentrations, the duration and thereby the integral of signaling activity of these proteins is increased under such conditions.

This principle of dose-to-duration signaling has been analyzed as a means to decode ligand levels beyond saturation and subsequently shown for pheromone signaling in yeast at the level of mitogen-activated protein kinases [27]. In light of this, it will be interesting to examine if the linear relation between extracellular ligand concentration and activation of signaling molecules might be abrogated downstream of the EpoR. Such an observation could indicate at which level EpoR-mediated signaling interacts with other signaling networks through pathway crosstalk, thereby allowing for integration and interpretation of the cellular signaling status.

2.2 Interleukin 3 Receptor

In contrast to the EpoR, the IL3R consists of a cytokine-specific alpha chain and the common beta chain, which is shared with cytokine receptors for IL5 and the granulocyte–macrophage colony-stimulating factor (GM-CSF) [28].

Studying the characteristics of IL3R activation showed that, comparable to the EpoR system, IL3 is rapidly depleted from the medium within the early phase of stimulation [8]. A second key feature shared by the EpoR and the IL3R is the restimulation capacity of both the receptor and the receptor-associated JAK2, demonstrating that cells remain ligand-responsive (Fig. 18.2). However, treatment of cells with IL3 resulted in a massive degradation of the common beta chain and JAK2 (Fig. 18.2b). This observation indicates that in contrast to the EpoR, the majority of IL3R resides at the plasma membrane where it is accessible for ligand binding. Another key difference between these receptor systems is the IL3-induced increase of beta chain expression, which may compensate for dramatic receptor degradation after ligand engagement and prevent a refractory state of the cell.

In summary, the EpoR and the IL3R reveal comparable characteristics of signaling at the receptor level, i.e. (1) rapid clearance of ligand from the medium and (2) receptor recovery at the plasma membrane. However, both receptor systems evolved distinct strategies to accomplish this systems behavior, either employing a constant rapid ligand-independent turnover of the EpoR or a massive ligand-stimulated synthesis of the IL3R (Fig. 18.2). Rapid uptake of ligand from the medium by ligand-induced endocytosis has been discussed to facilitate temporal fidelity of receptor signaling [29, 30]. Thus, the combination of rapid ligand depletion with fast cell surface recovery of the EpoR or the IL3R enables the cell to stay in a ligand-responsive state and at the same time promotes a high temporal resolution of extracellular signaling cues.

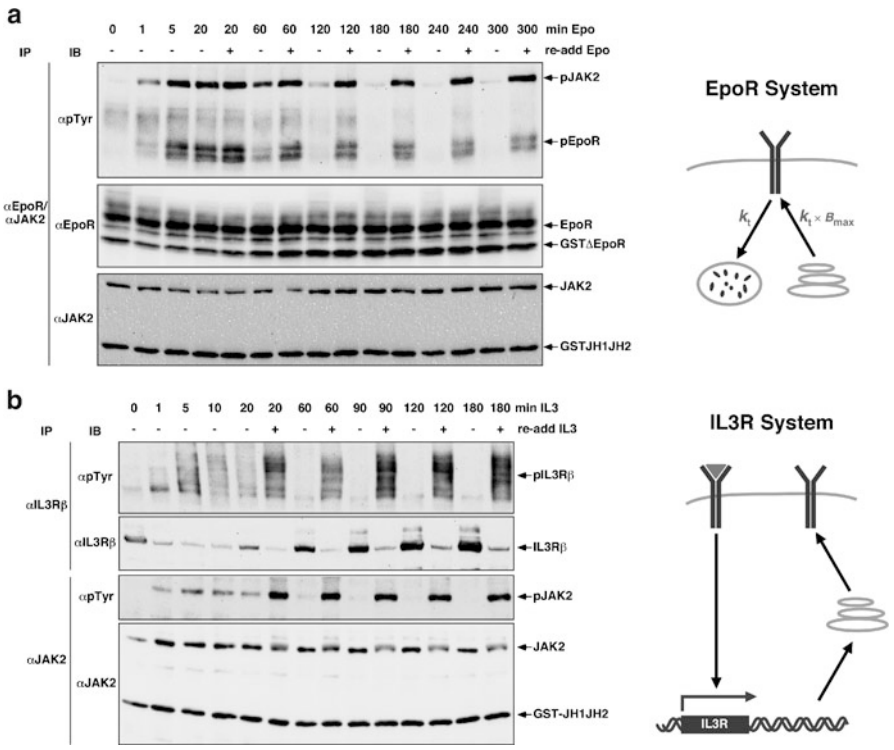


Fig. 18.2 Comparison of overall systems behavior and strategies employed in (**a**) the EpoR and (**b**) the IL3R system. Immunoblot analysis shows that both receptor systems stay in a ligand-responsive state as judged by receptor and JAK2 phosphorylation after re-addition of ligand. (**b**) Left panel adapted from [8]

3 Epidermal Growth Factor Receptor

Members of the receptor tyrosine kinase (RTK) family are single-pass transmembrane proteins that regulate multiple cellular processes such as proliferation, differentiation, migration, angiogenesis, and metabolism [31]. Conversely, deregulation of RTK signaling pathways has been assigned to various human cancers as well as non-malignant diseases [32, 33]. After completion of the Humane Genome Project, 58 RTKs have been identified [34] including ErbB receptors, vascular endothelial growth factor receptor, and c-Met. The EGFR (ErbB1, Her1) is a member of the ErbB receptor family and as the prototypical RTK probably the best-studied receptor, also from a systems point of view [6]. EGFR signaling regulates proliferation and survival in a variety of epithelial cell types, and deregulated signaling through the EGFR is associated with numerous solid tumors [35].

Biochemical studies showed that the EGFR is rapidly internalized from the plasma membrane upon epidermal growth factor (EGF) stimulation and subsequently degraded in the lysosomal compartment. This downregulation is proposed to contribute to signal attenuation [36, 37]. However, this observation is context-dependent since stimulation of the EGFR with transforming growth factor α (TGFα) results in receptor recycling rather than in downregulation [38] due to a higher pH sensitivity of ligand–receptor binding [39]. Differential binding and trafficking of EGF and TGFα have been shown to result in distinct mitogenic potency of EGFR signaling [40]. This knowledge has also been employed to engineer a more effective variant of EGF [41], and a similar study has been carried out for the cytokine GCSF [42]. Distinct receptor trafficking or binding properties also account for the altered biology of IL2 [43] and Epo [26] derivatives, respectively.

Comparing the regulatory role of endocytosis in EGFR and EpoR signaling shows that the contribution of endocytic downregulation D, i.e. the ratio of ligand-induced (k_e) to ligand-independent (k_t) receptor endocytosis, is approximately threefold higher for EGF-stimulated EGFR ($D = 7.5$) [30] than for the EpoR system ($D = 2.3$) [8]. This is due to both a lower rate of ligand-independent endocytosis and a higher rate for ligand-induced endocytosis of the EGFR compared to the EpoR. Whereas EGF mediated a substantial decrease in half-life and total expression of its receptor [44], neither higher levels of Epo nor prolonged exposure to ligand resulted in a change of total EpoR expression [8]. Thus, ligand-mediated loss of receptor protein at the plasma membrane is much more likely to play a role in attenuation of EGF-stimulated EGFR signaling [36, 37] compared to EpoR signaling. In addition, the ratio of ligand-induced endocytosis k_e to ligand–receptor dissociation k_{off} is considerably higher for Epo–EpoR compared to EGF–EGFR complexes [30]. This, in combination with a rapid constitutive receptor turnover, allows the EpoR system to reach a high temporal resolution of sampling extracellular cues, while, at the same time, staying in a ligand-responsive state.

4 Transforming Growth Factor β Receptor

In contrast to cytokine receptors and the EGFR, the TGFβ receptor belongs to the serine/threonine kinase receptor family. Binding of TGFβ ligand induces cooperative complex formation of two receptor subunits, the TGFβ type I and type II receptors. The type II receptor is a constitutively active serine/threonine kinase that, upon ligand binding, activates the dormant TGFβ type I receptor. The type I receptor in turn phosphorylates serine residues of receptor-associated SMAD2 and SMAD3 transcription factors [45]. TGFβ is mainly involved in the development as well as homeostasis of tissues. Although TGFβ signaling is typically thought of mediating anti-proliferative cues and, therefore, being a tumor suppressor, it can fuel tumor progression at later stages by stimulation of tumor angiogenesis and metastasis [46].

Signaling through SMAD transcription factors is promoted by clathrin-mediated endocytosis, whereas endocytosis via caveolae mediates receptor turnover [47, 48].

A recent study suggested that caveolae are also involved to differentially trigger the mitogen-activated protein kinase cascade [49]. Thus, receptor trafficking possesses the capacity to induce distinct biological responses, thereby establishing an additional layer of regulation to TGFβ signal transduction. Mathematical analysis of the TGFβ pathway showed that the connection of receptor activation and trafficking processes allows for sensing absolute and temporal changes in ligand concentrations, regulating signal duration, and controlling cellular responses upon stimulation with multiple ligands [7]. Another study suggested that the ratio of clathrin- and caveolae-mediated endocytosis controls transient versus sustained responses [50]. Similar to the TGFβ receptor, signaling from endosomes has also been proposed for signaling downstream of RTKs as well as G-protein coupled receptors (GPCR) as a mechanism to facilitate temporal and spatial regulation [51].

5 Concluding Remarks

The examples discussed in this review show that various strategies have evolved to shape signal initiation at the receptor level by ligand–receptor interaction and trafficking kinetics. The physiological impact of distinct trafficking routes and signaling endosomes is still not fully explored as illustrated by controversial results for caveolae-mediated EGFR internalization [52, 53]. Deciphering these processes might give rise to an even more complicated picture of how receptor dynamics set the stage for selective regulation of downstream signaling. However, despite these distinct strategies, a unifying regulator of signal transduction at the receptor level appears to be the ratio of ligand-independent and ligand-induced endocytosis and subsequent receptor degradation [7, 8, 30].

Different from homodimeric EpoR, many cytokine receptors are composed of heterotypic subunits. Besides the IL3R that shares its common beta chain with receptors for IL5 and GM-CSF, another subset of cytokine receptors including receptors for IL2, IL4, IL7, IL9, IL13, IL15, and IL21 have a common gamma chain, whereas receptors for e.g. IL6, IL11, or LIF engage the gp130 subunit [28]. This gives rise to potential competition between different receptors for their common chain and additionally, these receptors often signal through the same JAK–STAT cascade. Moreover, induced feedback regulators, for instance members of the suppressor of cytokine signaling (SOCS) family, can affect multiple cytokine receptors either directly or indirectly at the level of JAKs or downstream pathway components. Thus, there are numerous layers of cross-regulation in cytokine signaling as exemplified by studies of IL7 signaling [54]. These phenomena create the necessity to generate complex data and mathematical models, studying the effects of multiple cytokine stimuli or of a specific stimulus on the activity of various cytokine receptors.

Crosstalk also plays a crucial role for EGFR signaling in cancer. The EGFR does not only form hetero-oligomeric structures with other members of the ErbB receptor family, but it is also suggested to directly interact with c-Met [55, 56] and to exhibit

transactivation with c-Met [57] and GPCRs [58] at multiple levels. Such interactions are relevant for both drug resistance and cancer progression.

Although studies of cell lines exposed to single stimuli give rise to important insights, it will be crucial to expand the analysis of cell signaling towards more physiological conditions of multi-factor stimulation for understanding in vivo signaling through cell surface receptors. This also holds true for the repertoire of stimulation schemes: bolus stimulation is a rather non-physiological, yet practical means to examine signal transduction in cell lines. However, the investigation of autocrine or paracrine signaling in the cellular microenvironment or the administration of a constant stimulus at physiological concentrations promises to advance the field of signaling research. Here, technical developments such as microfluidics [59, 60] in combination with mathematical modeling may greatly impact the success of such endeavors and finally refine strategies for drug discovery [3].

Acknowledgments This work was supported by the Helmholtz Alliance on Systems Biology (SBCancer) (VB, JT, UK), the German Federal Ministry of Education and Research (BMBF)-funded MedSys-Network LungSys (JT, UK), and the Excellence Initiative of the German Federal and State Governments (EXC 294) (JT).

References

1. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5(12):993–996
2. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
3. Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. Nat Biotechnol 22(10):1253–1259
4. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J (2006) Cancer: a systems biology disease. Biosystems 83(2–3):81–90
5. Wiley HS, Cunningham DD (1981) A steady state model for analyzing the cellular binding, internalization and degradation of polypeptide ligands. Cell 25(2):433–440
6. Wiley HS, Shvartsman SY, Lauffenburger DA (2003) Computational modeling of the EGF–receptor system: a paradigm for systems biology. Trends Cell Biol 13(1):43–50
7. Vilar JM, Jansen R, Sander C (2006) Signal processing in the TGF-beta superfamily ligand-receptor network. PLoS Comput Biol 2(1):e3
8. Becker V, Schilling M, Bachmann J, Baumann U, Raue A, Maiwald T, Timmer J, Klingmüller U (2010) Covering a broad dynamic range: information processing at the erythropoietin receptor. Science 328(5984):1404–1408
9. Baker SJ, Rane SG, Reddy EP (2007) Hematopoietic cytokine receptor signaling. Oncogene 26(47):6724–6737
10. O'Shea JJ, Murray PJ (2008) Cytokine signaling modules in inflammatory responses. Immunity 28(4):477–487
11. Longmore GD, Lodish HF (1991) An activating mutation in the murine erythropoietin receptor induces erythroleukemia in mice: a cytokine receptor superfamily oncogene. Cell 67(6): 1089–1102
12. Arcasoy MO, Degar BA, Harris KW, Forget BG (1997) Familial erythrocytosis associated with a short deletion in the erythropoietin receptor gene. Blood 89(12):4628–4635

13. Forbes LV, Gale RE, Pizzey A, Pouwels K, Nathwani A, Linch DC (2002) An activating mutation in the transmembrane domain of the granulocyte colony-stimulating factor receptor in patients with acute myeloid leukemia. Oncogene 21(39):5981–5989
14. Ding J, Komatsu H, Wakita A, Kato-Uranishi M, Ito M, Satoh A, Tsuboi K, Nitta M, Miyazaki H, Iida S, Ueda R (2004) Familial essential thrombocythemia associated with a dominant-positive activating mutation of the c-MPL gene, which encodes for the receptor for thrombopoietin. Blood 103(11):4198–4200
15. James C, Ugo V, Le Couedic JP, Staerk J, Delhommeau F, Lacout C, Garcon L, Raslova H, Berger R, Bennaceur-Griscelli A, Villeval JL, Constantinescu SN, Casadevall N, Vainchenker W (2005) A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature 434(7037):1144–1148
16. Flex E, Petrangeli V, Stella L, Chiaretti S, Hornakova T, Knoops L, Ariola C, Fodale V, Clappier E, Paoloni F, Martinelli S, Fragale A, Sanchez M, Tavolaro S, Messina M, Cazzaniga G, Camera A, Pizzolo G, Tornesello A, Vignetti M, Battistini A, Cave H, Gelb BD, Renauld JC, Biondi A, Constantinescu SN, Foa R, Tartaglia M (2008) Somatically acquired JAK1 mutations in adult acute lymphoblastic leukemia. J Exp Med 205(4):751–758
17. Richmond TD, Chohan M, Barber DL (2005) Turning cells red: signal transduction mediated by erythropoietin. Trends Cell Biol 15(3):146–155
18. Wu H, Liu X, Jaenisch R, Lodish HF (1995) Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. Cell 83(1):59–67
19. Livnah O, Stura EA, Middleton SA, Johnson DL, Jolliffe LK, Wilson IA (1999) Crystallographic evidence for preformed dimers of erythropoietin receptor before ligand activation. Science 283(5404):987–990
20. Yoshimura A, D'Andrea AD, Lodish HF (1990) Friend spleen focus-forming virus glycoprotein gp55 interacts with the erythropoietin receptor in the endoplasmic reticulum and affects receptor metabolism. Proc Natl Acad Sci USA 87(11):4139–4143
21. Neumann D, Wikström L, Watowich SS, Lodish HF (1993) Intermediates in degradation of the erythropoietin receptor accumulate and are degraded in lysosomes. J Biol Chem 268(18):13639–13649
22. Hilton DJ, Watowich SS, Murray PJ, Lodish HF (1995) Increased cell surface expression and enhanced folding in the endoplasmic reticulum of a mutant erythropoietin receptor. Proc Natl Acad Sci USA 92(1):190–194
23. Ketteler R, Heinrich AC, Offe JK, Becker V, Cohen J, Neumann D, Klingmüller U (2002) A functional green fluorescent protein-erythropoietin receptor despite physical separation of JAK2 binding site and tyrosine residues. J Biol Chem 277(29):26547–26552
24. Becker V, Sengupta D, Ketteler R, Ullmann GM, Smith JC, Klingmüller U (2008) Packing density of the erythropoietin receptor transmembrane domain correlates with amplification of biological responses. Biochemistry 47(45):11771–11782
25. Walrafen P, Verdier F, Kadri Z, Chretien S, Lacombe C, Mayeux P (2005) Both proteasomes and lysosomes degrade the activated erythropoietin receptor. Blood 105(2):600–608
26. Gross AW, Lodish HF (2006) Cellular trafficking and degradation of erythropoietin and novel erythropoiesis stimulating protein (NESP). J Biol Chem 281(4):2024–2032
27. Behar M, Hao N, Dohlman HG, Elston TC (2008) Dose-to-duration encoding and signaling beyond saturation in intracellular signaling networks. PLoS Comput Biol 4(10):e1000197
28. Wang X, Lupardus P, Laporte SL, Garcia KC (2009) Structural biology of shared cytokine receptors. Annu Rev Immunol 27:29–60
29. Shankaran H, Wiley HS, Resat H (2007) Receptor downregulation and desensitization enhance the information processing ability of signalling receptors. BMC Syst Biol 1:48
30. Shankaran H, Resat H, Wiley HS (2007) Cell surface receptors for signal transduction and ligand transport: a design principles study. PLoS Comput Biol 3(6):e101
31. Lemmon MA, Schlessinger J (2010) Cell signaling by receptor tyrosine kinases. Cell 141(7):1117–1134

32. Lamorte L, Park M (2001) The receptor tyrosine kinases: role in cancer progression. Surg Oncol Clin N Am 10(2):271–288, viii
33. Grimminger F, Schermuly RT, Ghofrani HA (2010) Targeting non-malignant disorders with tyrosine kinase inhibitors. Nat Rev Drug Discov 9(12):956–970
34. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298(5600):1912–1934
35. Holbro T, Hynes NE (2004) ErbB receptors: directing key signaling networks throughout life. Annu Rev Pharmacol Toxicol 44:195–217
36. Wells A, Welsh JB, Lazar CS, Wiley HS, Gill GN, Rosenfeld MG (1990) Ligand-induced transformation by a noninternalizing epidermal growth factor receptor. Science 247(4945):962–964
37. Wiley HS, Herbst JJ, Walsh BJ, Lauffenburger DA, Rosenfeld MG, Gill GN (1991) The role of tyrosine kinase activity in endocytosis, compartmentation, and down-regulation of the epidermal growth factor receptor. J Biol Chem 266(17):11083–11094
38. Decker SJ (1990) Epidermal growth factor and transforming growth factor-alpha induce differential processing of the epidermal growth factor receptor. Biochem Biophys Res Commun 166(2):615–621
39. French AR, Tadaki DK, Niyogi SK, Lauffenburger DA (1995) Intracellular trafficking of epidermal growth factor family ligands is directly influenced by the pH sensitivity of the receptor/ligand interaction. J Biol Chem 270(9):4334–4340
40. Reddy CC, Wells A, Lauffenburger DA (1998) Comparative mitogenic potencies of EGF and TGF alpha and their dependence on receptor-limitation versus ligand-limitation. Med Biol Eng Comput 36(4):499–507
41. Reddy CC, Niyogi SK, Wells A, Wiley HS, Lauffenburger DA (1996) Engineering epidermal growth factor for enhanced mitogenic potency. Nat Biotechnol 14(13):1696–1699
42. Sarkar CA, Lowenhaupt K, Horan T, Boone TC, Tidor B, Lauffenburger DA (2002) Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". Nat Biotechnol 20(9):908–913
43. Fallon EM, Liparoto SF, Lee KJ, Ciardelli TL, Lauffenburger DA (2000) Increased endosomal sorting of ligand to recycling enhances potency of an interleukin-2 analog. J Biol Chem 275(10):6790–6797
44. Stoscheck CM, Carpenter G (1984) Down regulation of epidermal growth factor receptors: direct demonstration of receptor degradation in human fibroblasts. J Cell Biol 98(3): 1048–1053
45. Moustakas A, Heldin CH (2009) The regulation of TGFbeta signal transduction. Development 136(22):3699–3714
46. Ikushima H, Miyazono K (2010) TGF-beta signalling: a complex web in cancer progression. Nat Rev Cancer 10(6):415–424
47. Di Guglielmo GM, Le Roy C, Goodfellow AF, Wrana JL (2003) Distinct endocytic pathways regulate TGF-beta receptor signalling and turnover. Nat Cell Biol 5(5):410–421
48. Mitchell H, Choudhury A, Pagano RE, Leof EB (2004) Ligand-dependent and -independent transforming growth factor-beta receptor recycling regulated by clathrin-mediated endocytosis and Rab11. Mol Biol Cell 15(9):4166–4178
49. Zuo W, Chen YG (2009) Specific activation of mitogen-activated protein kinase by transforming growth factor-beta receptors in lipid rafts is required for epithelial cell plasticity. Mol Biol Cell 20(3):1020–1029
50. Zi Z, Klipp E (2007) Constraint-based modeling and kinetic analysis of the Smad dependent TGF-beta signaling pathway. PLoS One 2(9):e936
51. Miaczynska M, Pelkmans L, Zerial M (2004) Not just a sink: endosomes in control of signal transduction. Curr Opin Cell Biol 16(4):400–406
52. Sigismund S, Woelk T, Puri C, Maspero E, Tacchetti C, Transidico P, Di Fiore PP, Polo S (2005) Clathrin-independent endocytosis of ubiquitinated cargos. Proc Natl Acad Sci USA 102(8):2760–2765
53. Rappoport JZ, Simon SM (2009) Endocytic trafficking of activated EGFR is AP-2 dependent and occurs through preformed clathrin spots. J Cell Sci 122(Pt 9):1301–1305

54. Palmer MJ, Mahajan VS, Trajman LC, Irvine DJ, Lauffenburger DA, Chen J (2008) Interleukin-7 receptor signaling network: an integrated systems perspective. Cell Mol Immunol 5(2):79–89
55. Jo M, Stolz DB, Esplen JE, Dorko K, Michalopoulos GK, Strom SC (2000) Cross-talk between epidermal growth factor receptor and c-Met signal pathways in transformed cells. J Biol Chem 275(12):8806–8811
56. Guo A, Villen J, Kornhauser J, Lee KA, Stokes MP, Rikova K, Possemato A, Nardone J, Innocenti G, Wetzel R, Wang Y, MacNeill J, Mitchell J, Gygi SP, Rush J, Polakiewicz RD, Comb MJ (2008) Signaling networks assembled by oncogenic EGFR and c-Met. Proc Natl Acad Sci USA 105(2):692–697
57. Jänne PA, Gray N, Settleman J (2009) Factors underlying sensitivity of cancers to small-molecule kinase inhibitors. Nat Rev Drug Discov 8(9):709–723
58. Lappano R, Maggiolini M (2011) G protein-coupled receptors: novel targets for drug discovery in cancer. Nat Rev Drug Discov 10(1):47–60
59. Breslauer DN, Lee PJ, Lee LP (2006) Microfluidics-based systems biology. Mol BioSyst 2(2):97–112
60. Wang CJ, Levchenko A (2009) Microfluidics technology for systems biology research. Meth Mol Biol 500:203–219

Chapter 19
A Systems-Biology Approach to Yeast Actin Cables

Tyler Drake, Eddy Yusuf, and Dimitrios Vavylonis

Abstract We focus on actin cables in yeast as a model system for understanding cytoskeletal organization and the workings of actin itself. In particular, we highlight quantitative approaches on the kinetics of actin-cable assembly and methods of measuring their morphology by image analysis. Actin cables described by these studies can span greater lengths than a thousand end-to-end actin-monomers. Because of this difference in length scales, control of the actin-cable system constitutes a junction between short-range interactions – among actin-monomers and nucleating, polymerization-facilitating, side-binding, severing, and cross-linking proteins – and the emergence of cell-scale physical form as embodied by the actin cables themselves.

1 Introduction

Many basic cell functions such as cell motility, endocytosis, cytokinesis, and establishment of cell polarity depend on actin filaments [1]. Actin filament nucleation, polymerization, and controlled disassembly keep actin subunits in a state of constant turnover between the monomer and filament states. Groups of regulating proteins marshal this adaptable actin cytoskeleton for diverse tasks. A huge body of work considers the actin system both in controlled *in vitro* situations and in eukaryotic and bacteria cells that use actin or its homologs [1, 2]. These studies raise questions about how cells employ the actin system to move, polarize, divide, transport

T. Drake • D. Vavylonis (✉)
Department of Physics, Lehigh University, Bethlehem, PA 18015, USA
e-mail: tyler.gates.drake@gmail.com; vavylonis@lehigh.edu

E. Yusuf
Physics Department, Surya College of Education, Surya Research and Education (SURE) Center, Jln. Scientia Boulevard U7, Gading Serpong, Tangerang 15233, Indonesia
e-mail: yusuf.eddy@stkripsurya.ac.id

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_19,

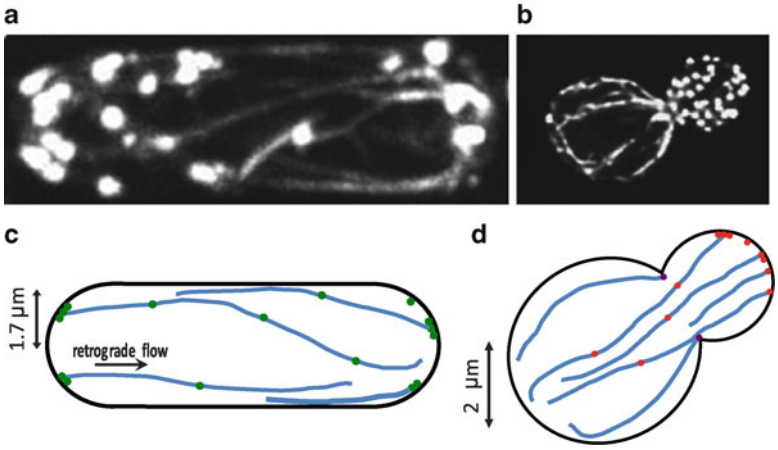


Fig. 19.1 Actin cables in fission and budding yeast. (**a**) Image of interphase yeast cells expressing actin marker CHD–GFP showing actin patches (bright spots) and actin cables (linear elements) [5]. (**b**) Actin cables (blue) in fission yeast polymerize away from the cell tips where formin For3p (green) nucleates actin filament assembly. (**c**) Image of budding yeast showing actin cables and actin patches (stained with rhodamine–phalloidin) from [12]. (**d**) Budding yeast actin cables (blue) are nucleated by formins Bni1p (red) and Bnr1p (purple) that localize at the bud and bud neck

material, resist stress, contract, and signal. Here, we discuss how budding and fission yeast can serve as model systems for universal molecular mechanisms of the actin cytoskeleton.

During growth, yeast cells build two actin structures: patches and cables [1,3–5]. The actin patches, dense dendritic networks of actin filaments nucleated by the Arp2/3 complex, assist endocytosis. The actin cables, which we focus on here, are bundles of actin filaments that run across the cell and guide the transport of secretory vesicles and organelles (see Fig. 19.1). Formins proteins at the cell cortex generate these bundles by promoting nucleation of new filaments out of monomers and by processive polymerization [6–8]. As actin filaments polymerize away from the cell tip, they become bundles held together by cross-linking proteins, often as long as the cell. Budding and fission yeast differ in shape and, accordingly, in where they direct formins to sow cables at specific sites. Two formins, Bni1p and Bnr1p, activate cable growth in budding yeast [3]. A single formin, For3p, initiates actin-cable growth in fission yeast [4]. The cell breaks down and disassembles long cables through the coordinated action of a set of proteins. As a whole, the actin-cable system manages an interface between actin biochemistry and cell geometry.

As a system to study how cells respond to information about location, actin cables offer many advantages. Yeast serves well for genetic manipulation, allowing researchers to exploit homologous recombination and deletion libraries. Regulating proteins and the cables themselves can be monitored by fluorescence microscopy. Methods have been developed to measure protein concentrations in yeast [9, 10]. Also, the cables may be one of the simpler actin structures: in fission yeast, they

appear to require far fewer assisting proteins than actin patches or contractile rings for division [4]. But although actin cables may be a relatively simple system, they certainly behave in ways that would be hard to predict by knowing only the interactions among components. Understanding the complexity of regulating these dynamic structures by timing and location seems to require the approach of systems-biology: simultaneous measurements of multiple components and rigorous statistical analyses combined with mathematical models [11]. Here we highlight recent quantitative studies of yeast actin cables and discuss our view of the direction of this field.

2 Quantifying the Polymerization Kinetics of Actin-Cable Assembly

Actin cables are very dynamic structures with lifetimes of order one minute. The constituent filaments grow by adding monomers from the cytoplasm, at their barbed ends, and losing monomers to the cytoplasm by severing and depolymerization. Many actin-binding proteins modulate these kinetics. For instance, formins seed cables and increase the rate of actin-monomer addition at the barbed end. Reaction rates and protein concentrations, and their regulation, can affect qualitative behavior and so understanding the actin-cable system depends on measuring their values.

Recent studies suggested a detailed molecular mechanism for formin-mediated actin-cable assembly [13, 14]. Formin proteins promote actin-filament nucleation and elongation by processive association with the polymerizing end of actin filaments [6–8]. In fission yeast, formin For3p localizes in cortical foci at the growing tips of the cell (see Fig. 19.1b). Budding-yeast actin cables are nucleated by formins Bni1p and Bnr1p. Bnr1p localizes at the bud neck (see Fig. 19.1d). Bni1p localizes as foci at the tip of the growing bud and subsequently joins Bnr1p at the bud neck (see Fig. 19.1d). Both Bni1p and For3p associate with large cortical macromolecular structures where they nucleate actin filaments for cables. These filaments are bundled by actin cross-linking proteins such as fimbrin [3], and undergo retrograde flow away from the bud (or away from the cell tips in fission yeast) at speeds of order 0.3 μm/s and larger [15, 16]. Long cables disassemble through the coordinated action of tropomyosin, cofilin, actin-interacting protein Aip1, coronin, and twinfillin [3, 17].

The association of Bni1p and For3p with the cortex is transient: within seconds, these formins dissociate from the cortex and passively follow actin cable retrograde flow and disassembly, thus following a turnover cycle similar to actin. Based on these observations, Buttery and Pellman [13] and Martin and Chang [14] proposed the mechanism shown in Fig. 19.2a. The movement of the formins away from the cortex (process 4 in Fig. 19.2a) was found to be dependent on actin-polymerization, indicating the existence of coupled control mechanisms between actin and formins. This feedback mechanism indicates the possibility for rich dynamical behavior by the cable system. Unlike Bni1p, Bnr1p appeared to remain associated with the

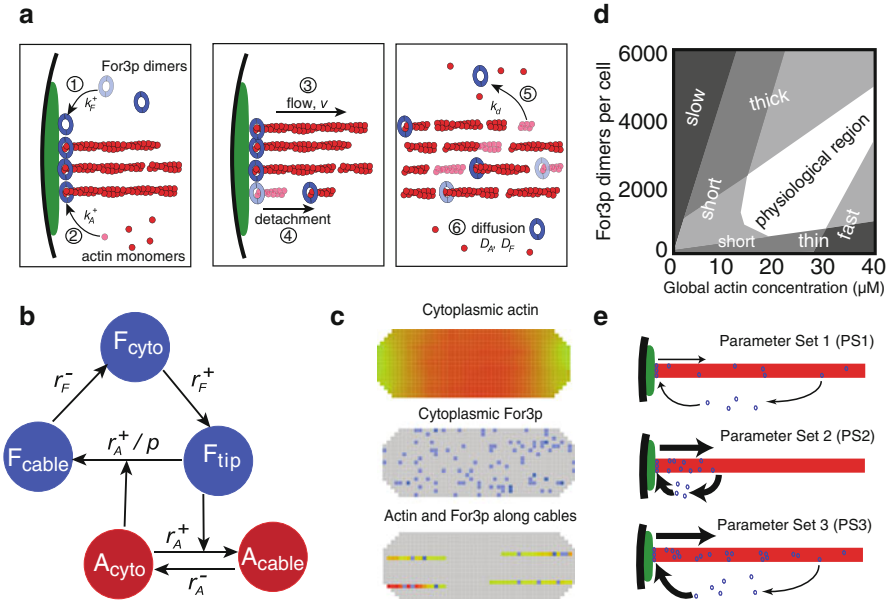


Fig. 19.2 Model of actin cables in fission yeast [18]. (**a**) Schematic showing the basic processes (1–6) of the model. (**b**) Ordinary differential equations model. (**c**) 3D computational lattice model accounting for the small number of For3p which are treated as discrete units. (**d**) Qualitative dynamical phase diagram describing the morphology of the actin-cable system as a function of actin and For3p concentration. (**e**) The model predicts different distributions of actin and For3p depending on rate constants

neck [13]. The cartoon in Fig. 19.2a appears to describe a summary of what is seen in experiments. But, without the support of a quantitative model, it is unclear if the model is even a consistent representation of an actin-cable assembly mechanism.

To explore the quantitative implications of the proposed model in Fig. 19.2a, Wang and Vavylonis added rate constants, protein concentrations, and diffusion coefficients in an analytical and computational model [18]. They considered fission yeast due to its simpler geometry as compared to budding yeast, and the fact that actin cables are nucleated by only one formin, For3p. Figure 19.2b shows the processes described by the rate equations of the model, with A_{cyto} and A_{cable} being the numbers of actin subunits in the cytoplasm and in actin cables, respectively, and F_{cyto}, F_{cable}, and F_{tip}, are the numbers of For3p in the cytoplasm, along the body of actin cables, and at cable tips, respectively. One rate constant depends on the processivity parameter, p, the average number of actin subunits polymerized per cortical For3p before its detachment into the cable. Whole-cell numerical simulations of actin and For3p reaction and diffusion were performed in 3D (Fig. 19.2c). The model considers a continuous field of cytoplasmic actin due to its abundance and individual For3p particles moving on a lattice due to their rarity.

The simulations validated the cartoon model. Using a combination of measured and fitted parameters, the model could explain experimental results, such as fluorescence recovery after photobleaching curves (FRAP) of For3p-3GFP and the response of the actin cables to treatments with the drug Latrunculin A (LatA), which promotes cable disassembly by sequestering available actin-monomers.

In addition, the model suggests a description of the system in the form of "dynamical phase diagrams" (Fig. 19.2d and e) that describe how parameter values (concentrations, rate constants) affect physiological properties of actin cables: polymerization rate, thickness, and length. Do cells tweak these parameter values to manipulate form (as in Fig. 19.2d) and thus optimize function? The facility of genetic engineering in yeast may allow future tests of these results. For example, systematic For3p overexpression and/or reduction of For3p expression levels are possible. Similarly, changes in the polymerization rate constant and processivity parameter could be tested by targeted changes in the FH2 and FH1 domains of For3p that mediate polymerization and processive motion [7, 16, 19, 20]. The above could be combined with treatments with drugs such as LatA, which effectively reduces the actin-polymerization rate constant.

With a constant processivity parameter, this model admits a single steady state for the actin-cable system. A cooperative mechanism for For3p detachment, meaning that the detachment rate depends sensitively on the polymerization rate, would introduce nonlinearities that could lead to additional steady states of actin-cable organization. This suggests the intriguing possibility that the cell might gain fitness through the ability to signal a switch between these states, allowing a rapid reorganization of the actin-cable system.

So far there have been no detailed quantitative models of actin-cable dynamics in budding yeast. However, some experimental studies treat the actin cables from a systems point of view [21]. For example, upon Bni1p overexpression the actin cables become shorter and more dense within the bud [22, 23]. In these overexpression studies, the actin cables within the mother cell (presumably nucleated by Bnr1p) become short and thin [22], though some mother cells become unusually large and contain multiple cable-like fragments [23]. This change in the actin cables in the mother cell could be due to the Bni1p-induced depletion of the actin-monomer pool available to Bnr1p. Because of uncertainties in the mechanisms of Bnr1p cortical dissociation and association, the effects of Bnr1p overexpression [21] are harder to interpret. Full length Bnr1p overexpression has small effects [21], though overexpression of unregulated Bnr1p leads to serious defects that can be rescued by an increase in the concentrations of proteins that bind to actin-monomers or with treatment with LatA, possibly by reducing Bnr1p-mediated nucleation of actin filaments in the cytoplasm [21]. A more recent study showed that the two formins, Bni1p and Bnr1p, assemble kinetically in separable cable populations [16]. A future quantitative modeling approach may help to provide an insight into the importance of these experimental findings.

3 Analyzing the Morphology of Actin Cables

The model in Fig. 19.2a described the kinetics of nucleation, assembly, disassembly, and severing in 3D but treated each cable as a 1D object. Further cable traits arise however in 3D: cables bend, twist, and buckle; organelles deflect the cables and cell geometry confines them; and cables cross or become bundled with one another. Actin-binding proteins regulate this cable morphology. Cross-linking proteins, such as fimbrin [3, 4], can bind to multiple filaments and stiffen a single cable or help to bundle multiple cables. Together with side-binding proteins, such as tropomyosin [24] they mediate bending and twisting and may make filaments more or less susceptible to severing [25]. Actin cables also appear to associate with actin patches [26]; further, association with myosin V [16] may attach cables to organelles. Mutations in these proteins can disrupt normal actin-cable morphology, indicating that these proteins regulate the spatial distribution of actin cables.

Measuring actin-cable morphology requires clear images of actin cables. This is possible since yeast actin cables are dilute as compared to actin structures in other cells. Several fluorescent markers can illuminate cables in live cells. For example, a fusion of the calponin homology domain from the IQGAP Rng2p to GFP (GFP–CHD) [14], the seventeen-amino-acid peptide Lifeact [30], and actin-binding protein 140 tagged with GFP [15] all mark filamentous actin. Confocal microscopy allows for the reconstruction of the cable position in three dimensions. Sub-diffraction microscopy could enhance the precision with which cables can be located.

Extracting the numbers that describe morphology from images and movies of the cables presents another challenge, but recent work shows that this can be done. Smith and others provide an open-source tool, JFilament, which fits this task [28]. JFilament uses stretching open active contours [27] to find flexible filaments in a noisy image. The algorithm starts with a proposed filament skeleton and modifies it to minimize an energy, which includes internal terms that penalize bending and stretching and external terms that account for crossing a gradient in the image (see Fig. 19.3a). With adjustments to the relative contribution of these terms and some manual interaction, JFilament allows efficient capture of many actin-cable statistics. Using the program, Smith and others analyzed cables with a clear trajectory across the cell (Fig. 19.3b and c). They found two length scales that described the cables, one less than the persistence length of single actin filaments and one closer to the persistence length of microtubules. The smaller length scale could correspond to short-scale deformations from pulling and motor buckling [31, 32], interaction of cables with patches [33], or fixed fluctuations during actin-cable assembly [18, 34]. The longer length scale could reflect the stiffness of the bundles and the fact that the actin cables are confined to the cell interior, which behaves as a rigid tube [35]. In any case, these analyses suggest that the equilibrium semiflexible-polymer description needs a few additions to capture the behavior of actin cables.

Looking forward, work towards complete, reliable automation of the data extraction should allow for the leveraging of the statistics of thousands of cells' worth of actin cables, possibly to reveal subtle changes in the regulation of morphology

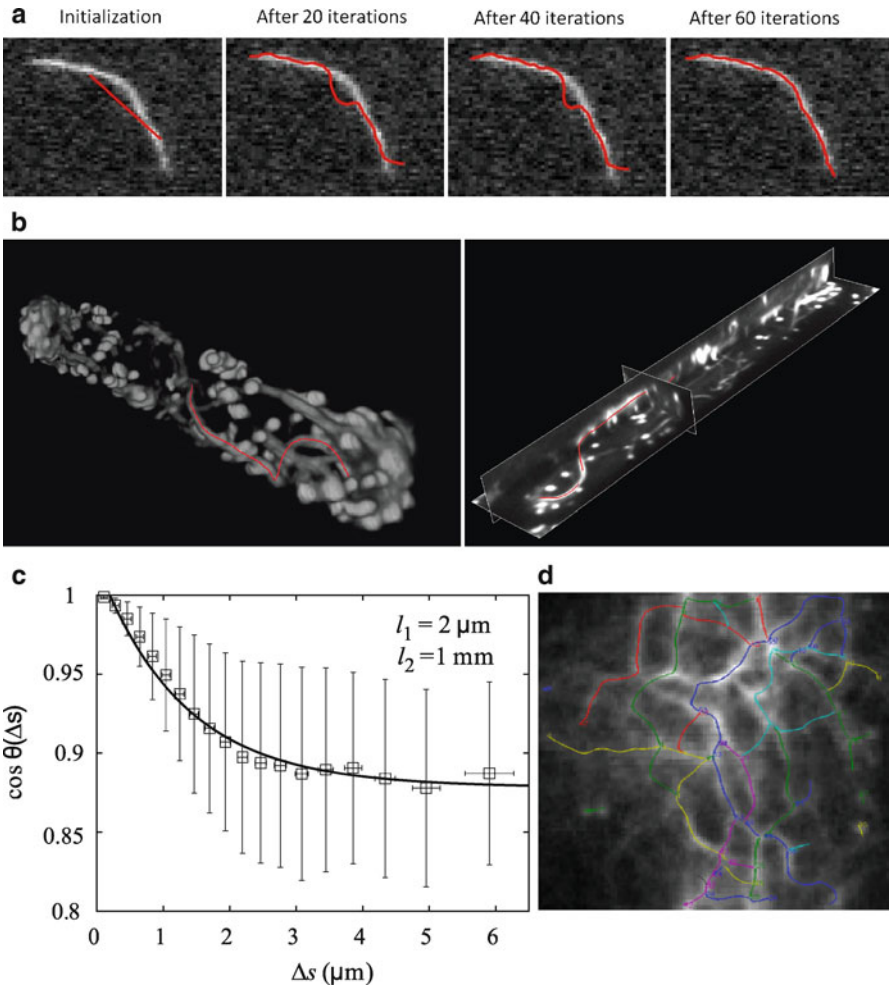


Fig. 19.3 Segmentation and tracking of actin cables using active contours. (**a**) Example of segmentation in a total internal reflection microscopy image of a single actin filament [27]. Initialization of the active contour away from the central line of the filament and position of active contour after 20, 40, and 60 iterations of deformation. Scale: 1 pixel = 0.17 μm. (**b**) Images showing a fission yeast *cdc25-22* cell expressing GFP–CHD that marks actin cables and actin patches [28]. Left: 3D volume view and active contour of a segmented actin cable. Right: Image of an active contour together with x, y and z cross-sections of the image. Cell diameter is $\sim$3.5 μm. (**c**) The tangent correlation function of actin cables from images as in panel **b**. A fit to a double exponential (continuous line) leads to length scales $l_1 = 2\,\mu$m and $l_2 = 1$ mm [28]. (**d**) Automatic segmentation of 2D filament network using multiple active contours from [29]. A meshwork is generated by initialization of multiple active contours at ridge points followed by growth, merging and splitting of active contours. Grouping analysis is used to classify segments. Image shows application to a 2D radial projection of a 3D confocal microscopy volume of a dividing *cdc25-22* fission yeast cell expressing GFP–CHD. Vertical axis is arc length

across the breadth and cycle of the cell. Reported automated two-dimensional methods that distill stacks of images into filament locations and network topologies [29] (Fig. 19.3d) could be extended to three dimensions. Also, automated separation of actin patches and cables remains challenging.

From this data, modeling studies will attempt to answer open questions: Do actin cables interact with or attach to the membrane? If so, can they grow while attached? Is cable position tightly regulated, or do cells allow random processes to determine their precise location? Are multiple nucleators required in budding yeast because it has a more complex shape than fission yeast? What minimal set of regulating processes can capture the salient aspects of actin-cable morphology? Theoretical work addresses the behavior of semi-flexible polymer bundles under confinement [35–37], but this has yet to be applied to actin cables in yeast. A model may show that only a few simple assumptions are necessary to reproduce most characteristics of measurable behavior, and this could become a framework for understanding how cells regulate the morphology of actin cables. These mathematical models will help us to understand how proteins guide this measured morphology through collective behavior.

4 Outlook

Cells may have optimized the actin-cable parameter values to be robust [38], corresponding to a large parcel of parameter space for the physiological region in Fig. 19.2d. However, the actin-cable system adapts for reorganization, as when the cables disassemble and actin filaments move to the division site for cytokinesis in fission yeast [5, 39]. The size of the physiological region may balance robust behavior for actin cables, which requires a large region, with the need for a malleable actin system that may be adapted to many purposes, which may require a small region. Here we motivate a systematic experimental exploration of parameter space to test these issues. Such studies should also reveal quantitative details on the role of other components of actin cables, such as regulatory pathways and bundling kinetics.

The results in yeast may have implications on the general role of formins in cells beyond yeast, such as the actin-cable network in plants [40–42]. Because changes to parameter values establish different distributions of actin and formins within yeast, many other eukaryotic cells may have also used this property to establish different patterns and structures. Future work will uncover the extent of universality in the mechanisms of formin function. Much remains to be established, for example, on the precise function of fission yeast formin Cdc12p in nucleating disperse actin meshworks and/or actin cables during the assembly of the cytokinetic contractile ring [43–46]. Hopefully, the modular structure of biological systems will allow us to proceed to a hierarchical understanding of the cell biological function of formin-mediated actin structures, starting from general features at a mesoscopic level of description, down to the full details of regulatory pathways that may differ across organisms.

Finally, we acknowledge some significant challenges. Attempts to measure cable elongation rates encounter technical difficulties – the end can be hard to locate and the dynamic nature of the marker complicates FRAP experiments. Models of actin cables become more complex as they include more elements, obscuring their interpretation. Also, the actin cables are only approximately a modular system, and incomplete knowledge of the systems with which they interact may limit understanding of the actin cables. However, we are optimistic that an increasing toolbox of quantitative methods will eventually help to overcome such obstacles.

Acknowledgements This work was supported by NIH Grant R21GM083928. We thank Nikola Ojkic, Matt Smith, and Jian-Qiu Wu for discussions.

References

1. Pollard TD, Cooper JA (2009) Actin, a central player in cell shape and movement. Science 326:1208–1212
2. Shaevitz JW, Gitai Z (2010) The structure and function of bacterial actin homologs. Cold Spring Harb Perspect Biol 2(9):a000364
3. Moseley JB, Goode BL (2006) The yeast actin cytoskeleton: from cellular function to biochemical mechanism. Microbiol Mol Biol Rev 70:605–645
4. Kovar DR, Sirotkin V, Lord M (2011) Three's company: the fission yeast actin cytoskeleton. Trends Cell Biol 21:177–187
5. Drake T, Vavylonis D (2010) Cytoskeletal dynamics in fission yeast: a review of models for polarization and division. HFSP J 4:122–130
6. Kovar DR, Pollard TD (2004) Insertional assembly of actin filament barbed ends in association with formins produces piconewton forces. Proc Natl Acad Sci USA 101:14725–14730
7. Vavylonis D, Kovar DR, O'Shaughnessy B, Pollard TD (2006) Model of formin-associated actin filament elongation. Mol Cell 21:455–466
8. Mizuno H, Higashida C, Yuan Y, Ishizaki T, Narumiya S, Watanabe N (2011) Rotational movement of the formin mDia1 along the double helical strand of an actin filament. Science 331:80–83
9. Wu JQ, Pollard TD (2005) Counting cytokinesis proteins globally and locally in fission yeast. Science 310:310–314
10. Joglekar AP, Bouck DC, Molk JN, Bloom KS, Salmon ED (2006) Molecular architecture of a kinetochore-microtubule attachment site. Nat Cell Biol 8:581–585
11. Sauer U, Heinemann M, Zamboni N (2007) Getting closer to the whole picture. Science 316:550–551
12. Amberg DC (1998) Three-dimensional imaging of the yeast actin cytoskeleton through the budding cell cycle. Mol Biol Cell 9:3259–3262
13. Buttery SM, Yoshida S, Pellman D (2007) Yeast Formins Bni1 and Bnr1 utilize different modes of cortical interaction during the assembly of actin cables. Mol Biol Cell 18:1826–1838
14. Martin SG, Chang F (2006) Dynamics of the formin for3p in actin cable assembly. Curr Biol 16:1161–1170
15. Yang H, Pon LA (2002) Actin cable dynamics in budding yeast. Proc Natl Acad Sci USA 99:751–756
16. Yu JH, Crevenna AH, Bettenbhl M, Freisinger T, Wedlich-Söldner R (2011) Cortical actin dynamics driven by formins and myosin V. J Cell Sci 124:1533–1541
17. Okada K, Ravi H, Smith EM, Goode BL (2003) Aip1 and cofilin promote rapid turnover of yeast actin patches and cables: a coordinated mechanism for severing and capping filaments. Mol Biol Cell 17:2855–2868

18. Wang H, Vavylonis D (2008) Model of For3p-mediated actin cable assembly in fission yeast. PLoS ONE 3:e4078
19. Kovar DR, Harris ES, Mahaffy R, Higgs HN, Pollard TD (2006) Control of the assembly of ATP- and ADP-actin by formins and profilin. Cell 124:423–435
20. Paul A, Pollard T (2008) The role of the FH1 domain and profilin in formin-mediated actin-filament elongation and nucleation. Curr Biol 18:9–19
21. Gao L, Bretscher A (2008) Analysis of unregulated formin activity reveals how yeast can balance F-actin assembly between different microfilament-based organizations. Mol Biol Cell 19:1474–1484
22. Evangelista M, Pruyne D, Amberg DC, Boone C, Bretscher A (2002) Formins direct Arp2/3-independent actin filament assembly to polarize cell growth in yeast. Nat Cell Biol 4: 260–269
23. Sagot I, Klee SK, Pellman D (2002) Yeast formins regulate cell polarity by controlling the assembly of actin cables. Nat Cell Biol 4:42–50
24. Huckaba TM, Lipkin T, Pon LA (2006) Roles of type II myosin and a tropomyosin isoform in retrograde actin flow in budding yeast. J Cell Biol 175:957–969
25. Skau CT, Kovar DR (2010) Fimbrin and tropomyosin competition regulates endocytosis and cytokinesis kinetics in fission yeast. Curr Biol 20:1415–1422
26. Toshima JY, Toshima J, Kaksonen M, Martin AC, King DS, Drubin DG (2006) Spatial dynamics of receptor-mediated endocytic trafficking in budding yeast revealed by using fluorescent alpha-factor derivatives. Proc Natl Acad Sci USA 103:5793–5798
27. Li H, Shen T, Smith MB, Fujiwara I, Vavylonis D, X Huang C (2009) Automated actin filament segmentation, tracking, and tip elongation measurements based on open active contour models. In: ISBI'09: Proceedings of the Sixth IEEE international conference on symposium on biomedical imaging, 1302–1305
28. Smith MB, Li H, Shen T, Huang X, Yusuf E, Vavylonis D (2010) Segmentation and tracking of cytoskeletal filaments using open active contours. Cytoskeleton 67:693–705
29. Xu T, Li H, Shen T, Ojkic N, Vavylonis D, X Huang C (2011) Extraction and analysis of actin networks based on open active contour models. In: ISBI'11: Proceedings of the sixth IEEE international conference on symposium on biomedical imaging, 1334–1340
30. Riedl J, Crevenna AH, Kessenbrock K, Yu JH, Neukirchen D, Bista M, Bradke F, Jenne D, Holak TA, Werb Z, Sixt M, Wedlich-Söldner R (2008) Lifeact: a versatile marker to visualize F-actin. Nat Methods 5:605–607
31. Bicek AD, Tuzel E, Demtchouk A, Uppalapati M, Hancock WO, Kroll DM, Odde DJ (2009) Anterograde microtubule transport drives microtubule bending in LLC-PK1 epithelial cells. Moll Cell Biol 20:2943–2953
32. Brangwynne CP, Koenderink GH, MacKintosh FC, Weitz DA (2008) Nonequilibrium microtubule fluctuations in a model cytoskeleton. Phys Rev Lett 100:118104
33. Huckaba TM, Gay AC, Pantalena LF, Yang HC, Pon LA (2004) Live cell imaging of the assembly, disassembly, and actin cable-dependent movement of endosomes and actin patches in the budding yeast, Saccharomyces cerevisiae. J Cell Biol 167:519–530
34. Brangwynne CP, MacKintosh FC, Weitz DA (2007) Force fluctuations and polymerization dynamics of intracellular microtubules. Proc Natl Acad Sci USA 104:16128–16133
35. Wagner F, Latanzi G, Frey E (2007) Conformations of confined biopolymers. Phys Rev E 75:050902(R)
36. Bathe M, Heussinger C, Claessens MMAE, Bausch AR, Frey E (2008) Cytoskeletal bundle mechanics. Biophys J 94:2955–2964
37. Wang B, Guan J, Anthony SM, Bae SC, Schweizer KS, Granick S (2010) Confining potential when a biopolymer filament reptates. Phys Rev Lett 104:118301
38. Eldar A, Dorfman R, Weiss D, Ashe H, Shilo BZ, Barkai N (2002) Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. Nature 419:304–308
39. Pollard TD, Wu JQ (2010) Understanding cytokinesis: lessons from fission yeast. Nat Rev Mol Cell Biol 11:149–155

40. Staiger CJ, Sheahan MB, Khurana P, Wang X, McCurdy DW, Blanchoin L (2009) Actin filament dynamics are dominated by rapid growth and severing activity in the Arabidopsis cortical array. J Cell Biol 184:269–280
41. Smertenko AP, Deeks MJ, Hussey PJ (2010) Strategies of actin reorganisation in plant cells. J Cell Sci 123:3019–3028
42. Vidali L, van Gisbergen PAC, Gurin C, Franco P, Li M, Burkart GM, Augustine RC, Blanchoin L, Bezanilla M (2009) Rapid formin-mediated actin-filament elongation is essential for polarized plant cell growth. Proc Natl Acad Sci USA 106:13341–13346
43. Coffman VC, Nile AH, Lee IJ, Liu H, Wu JQ (2009) Roles of Formin Nodes and Myosin Motor Activity in Mid1p-dependent Contractile-Ring Assembly during Fission Yeast Cytokinesis. Mol Biol Cell 20:5195–5210
44. Yonetani A, Chang F (2010) Regulation of cytokinesis by the formin cdc12p. Curr Biol 20:561–566
45. Kamasaki T, Osumi M, Mabuchi I (2007) Three-dimensional arrangement of F-actin in the contractile ring of fission yeast. J Cell Biol 178:765–771
46. Vavylonis D, Wu JQ, Hao S, O'Shaughnessy B, Pollard TD (2008) Assembly mechanism of the contractile ring for cytokinesis by fission yeast. Science 319:97–100

Chapter 20
Modeling Morphodynamic Phenotypes and Dynamic Regimes of Cell Motion

Mihaela Enculescu and Martin Falcke

Abstract Many cellular processes and signaling pathways converge onto cell morphology and cell motion, which share important components. The mechanisms used for propulsion could also be responsible for shape changes, if they are capable of generating the rich observed variety of dynamic regimes. Additionally, the analysis of cell shape changes in space and time promises insight into the state of the cytoskeleton and signaling pathways controlling it. While this has been obvious for some time by now, little effort has been made to systematically and quantitatively explore this source of information. First pioneering experimental work revealed morphodynamic phenotypes which can be associated with dynamic regimes like oscillations and excitability. Here, we review the current state of modeling of morphodynamic phenotypes, the experimental results and discuss the ideas on the mechanisms driving shape changes which are suggested by modeling.

1 Introduction

Cell motility plays a key role in tumor cell migration and enables the directed movement of embryonic cells to the appropriate locations in the body [141]. Understanding the mechanisms of cell motility might be a basic tool to inhibit cancer spread or prevent cardiac malfunctions [48].

M. Enculescu
Institute for Theoretical Physics, Technische Universität Berlin, Hardenbergstr. 36, 10623 Berlin, Germany
e-mail: mihaela.enculescu@tu-berlin.de

M. Falcke (✉)
Mathematical Cell Physiology, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
e-mail: martin.falcke@mdc-berlin.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_20,

The goal of this review is to critically discuss and classify mathematical models which capture the essential biological dependencies found experimentally. Such models show by reduction the most important interactions of a very complex biomechanical system. Also, models can predict how the migration pattern changes by perturbing different mechanisms, and offer therefore further insight into the biological phenomena.

Mathematical models for cell motility cover several levels of description – from single actin filaments to cell fragments, whole cells, and tissues [26]. They also focus on the description of different aspects of cell motility: initiation of actin-assembly and pre-merging conditions, perpetuation mechanisms after movement has started, adhesion and the interaction with the extra-cellular matrix, morphodynamics, or cell-to-cell communication and group dynamics of migrating cells. Here, we concentrate on the morphodynamics of single crawling cells.

The cell shape is mainly determined by the cell cytoskeleton, which is one of the main players in cell crawling. Hence, understanding the external cell shape deformations can indirectly provide information about the state of the motility machinery of the cell. Cell crawling occurs by the interplay of leading edge protrusion, adhesion of the front, deadhesion of the back, and contraction of the cell body [3, 57]. Membrane protrusion, that also determines the cell shape dynamics, occurs by the extension of a thin flat cytoskeletal structure, the lamellipodium, in the direction of motion. Inside the lamellipodium, a network of cross-linked actin filaments grows by polymerization in the direction of movement. This actin network, attached to the extra-cellular matrix and to the rest of the cell body, can be viewed as the motor of the cell. The main mechanisms that drive it are briefly reviewed in the following.

2 Basic Ideas on the Motile Machinery of Cells

The cytoskeleton of the cell contains several biopolymers that differ in stiffness and polarity. They can grow and shrink, rearrange, cross-link, and form bundles. This determines the form of the cell and can also generate movement. The force of protrusion in the lamellipodium is believed to arise from the polymerization of actin [92, 112]. Actin polymers are found in bundles in the interior of the lamellipodium, where myosin motor molecules can move along them to create contractions. Toward the leading edge membrane, actin forms a polar network with the fast polymerizing ends directed toward the membrane. At the opposite end, filaments depolymerize, actin monomers are recycled and diffuse to the front, where they are consumed by the growing tips [110]. This process of treadmilling is regulated by a number of proteins [14, 34, 55, 66, 87, 128]. Arp2/3 (actin related protein 2/3) binds to an existing actin filament and nucleates a new branch. Arp2/3 is activated by regulatory proteins, like the membrane associated WASP. Capping proteins bind to the end of a filament and prevent polymerization and depolymerization. Cofilin binds to filaments, enhances depolymerization, and severs them. Profilin binds to actin

monomers and favors the recycling of actin monomers into filaments. Thymosin β_4 binds actin monomers preventing their polymerization and acts as a buffer for monomeric actin. Different kinds of cross-linking proteins connect filaments and provide mechanical stability to the network. Other proteins are believed to bind actin polymers to the membrane.

Several studies have found differences in the actin network region just behind the leading edge and the network further in the bulk of the lamellipodium [111, 121, 133]. This leads to a picture where two different actin arrays – the wide lamella and a narrow lamellipodium in front or on top of the lamella – are pushing the membrane. While the lamellipodium is rich in branched polymerizing and depolymerizing actin filament ends, the lamella consists of more strongly cross-linked or bundled filaments. Earlier studies suggested the lamellipodium network to be highly branched and cross-linked very close to the leading edge membrane already [131, 139]. More recent studies showed that the branch point density in the lamellipodium may be rather low and the lamellipodium-like structures may extend several hundred nanometers into the cell [137]. The studies also differ in their results on filament length. While some conclude that filaments in the lamellipodium have a length of a few hundred nanometers [131, 139], others find a few micrometers [75,122,123,137]. In the dual picture, protrusion and retraction of the leading edge is due to the lamellipodium, while the lamella plays the main role in cell translocation, by integrating contractions due to myosin motors with adhesions to the substrate [13, 29, 30, 75]. Other studies however question the existence of two different actin networks in migrating cells [123] and a lamella beneath the lamellipodium [138].

3 A Short Review of the Scientific Discussion on Actin Filament Attachment to the Leading Edge Membrane of Lamellipodia and the Evidence for the Presence and Functioning of F-Actin-Membrane-Linking Proteins There

Attachment of filaments to the surface of the object which is moved by actin polymerization is found in many reconstituted systems and biomimetic systems. That observation led to the formulation of the tethered ratchet model [98]. Attached filaments may fundamentally change the force balance at the obstacle surface since they can exert pulling forces. Indeed, pushing and pulling forces exerted by attached and polymerizing filaments respectively, may both be much stronger than the resulting difference, which is then equal to the force actually moving the object [43, 98]. Hence, it is worth discussing whether filaments also attach to the leading edge membrane of lamellipodia and whether models should take that into account.

While there is no direct proof of attachment of filaments to the lamellipodium leading edge membrane, Carlier and Pantaloni state “Biomimetic assays of propulsion of N-WASP-functionalized microspheres or vesicles have demonstrated that the actin tail is attached to the particle surface…, suggesting that similar bonds

exist between the filaments and the membrane during protrusion" [14]. Attachment was also observed with oil droplets used as biomimetic system [135]. Binding of filaments to leading edge membrane is discussed as a possibility, suggestion, or even necessity for directed motion by several groups and labs [14, 15, 27, 75]. Keren and Theriot also point out "the high concentration of protein complexes at the leading edge and their extensive connections to the actin cytoskeleton" [72]. Co et al. demonstrate that actin filaments can bind to WH2 domains also independently of the branching process [27, 130]. Hence, even without considering the major F-actin-membrane linker ezrin, radixin, and moesin (ERM proteins), membrane binding of F-actin is suggested.

Only activated ERM proteins link F-actin to membrane proteins. They are activated by first binding PIP2 and subsequent phosphorylation at a threonine residue (T576 ezrin, T558 moesin, and T564 radixin) [46, 47]. ERMs are phosphorylated by myotonic dystrophy kinase-related Cdc42-binding kinase in filopodia [100], protein kinase Cα in membrane protrusions [101], and the Rho-associated kinase (ROCK) in microvilli [104], although this latter finding is controversial [93]. G protein-coupled receptor kinase 2 phosphorylates radixin in epithelial cells [71]. The Nck-interacting kinase NIK phosphorylates ERM proteins in rat mammary epithelial cells and in CCL39 fibroblasts [7]. ERM proteins are phosphorylated in response to stimuli linked to motility and morphodynamics.

Active ERM proteins and their binding partners are located at the leading edge. Phosphorylated ezrin is localized in ruffles and at the leading edge of pseudopodia of fibroblasts [85]. Similarly, phosphorylated ezrin and NIK were found at the distal margins of lamellipodia in mammary epithelial cells [7]. Baumgartner et al. mention the interesting idea that localization of kinases may sharpen the localization of pERM at the distal margins of lamellipodia beyond the localization of ERM, which is already restricted to lamellipodia [7]. The Na^+–H^+-exchanger NHE1 is one of the ERM binding partners in the plasma membrane [36]. NHE1 is enriched in lamellipodia and membrane tufts of fibroblasts [36, 60, 114] (and other cell types [74, 83]) and the membrane pool of ezrin is predominantly bound to NHE1 [36]. NHE1 can also be found along the smooth edge of the cell [36]. Ezrin localization showed a striking overlap with NHE1, but radixin was only found in lamellipodia and membrane tufts [36]. It is interesting to note in this context that radixin was originally identified as a barbed end capping protein [136].

Activation of ERM proteins may cause lamellipodium formation and ezrin–NHE1 binding is required for normal lamellipodium shape. Radixin is involved in lamellipodia stability of nerve growth cones [25]. ERM are also involved in lamellipodium formation. Phosphorylation at T567 causes formation of lamellipodia in LLC-PK cells [50]. F-actin networks extended to the peripheral edge of membrane protrusions in fibroblasts expressing NHE1, which was able to bind ezrin, but not in fibroblasts deficient of NHE1 or expressing NHE1 not able to bind ezrin [36]. Loss of NHE1-dependent cytoskeletal anchoring impairs directionality of cell migration [35]. Migrating fibroblasts expressing ezrin-binding NHE1 form a broad lamellipodium, by contrast with migrating cells expressing NHE1 unable to bind ezrin which form many small protrusions [35].

Another actin- and membrane-binding protein – myristoylated alanine-rich C kinase substrate (MARCKS) – is involved in lamellipodia formation [117]. It translocates to the membrane upon dephosphorylation. MARCKS is phosphorylated at Ser 159 by Rho-kinase as well as PKC [67, 99, 132]. In SH-SY5Y cells, stimulation with insulin-like growth factor-I (IGF-I) causes dephosphorylation of MARCKS. PI3-K has been reported to be involved in the dephosphorylation via activation of the PI3-K/Akt pathway [106,118,140]. PI3-K inhibitors attenuated the IGF-I-induced dephosphorylation of MARCKS, MARCKS translocation to lipid rafts and lamellipodia formation. These results support the idea that the transient dephosphorylation of MARCKS induced by IGF-I triggers the translocation of MARCKS to lipid rafts and lamellipodia formation [144]. IGF-I stimulation of SH-SY5Y cells caused the translocation of MARCKS to lipid rafts in the edge of lamellipodia, where it forms a complex with PIP2 [144]. Knockdown of MARCKS with siRNA technology abolished lamellipodia and neurite formation induced by IGF-I [144]. Cells exhibited a small number of tiny lamellipodia-like structures at the cell edge instead but not widely spread F-actin structures. That is evocative of the small protrusions reported from migrating fibroblasts expressing NHE1 unable to bind ezrin [35].

IGF-I stimulation also transiently decreases RhoA–GTP content in SH-SY5Y cells [118]. The RhoA/Rho-kinase pathway is considered to be a major target of the PI3-K/Akt signaling pathway, and PI3-K negatively controls RhoA activity [106, 140]. Hence, a link from MARCKS to ERM proteins via RhoA might exist.

Gelsolin is an actin severing and barbed end capping protein [129, 145, 146]. Gelsolin can bind actin filaments and membrane at the same time [61, 94]. Gelsolin interacts with PIP2, which inhibits capping [68]. Whether PIP2 also uncaps filaments [38, 115] or not [79] is a matter of debate. Gelsolin can also bind polyphosphoinositide-free lipid vesicles and simultaneously to actin microfilaments [94]. CP (called CapZ in muscle) also caps F-actin barbed ends. It also interacts with PIP2 [62, 79]. It has also been suggested that CapZ can link F-actin and membrane independently of PIP2 [125]. Both gelsolin [124] and CapZ [28] are present in the lamellipodium. Hence, gelsolin and CP are further potential F-actin-membrane linkers.

Actin binding membrane proteins can stay at the leading edge despite the retrograde flow of the actin network. References [7,85] suggest pERM to be located directly at the leading edge. This is supported by another simple consideration. Actin binding proteins in the membrane are carried away by F-actin retrograde flow in the lamellipodium, if there is no counteracting force. Hence, actin binding proteins staying in the lamellipodium must either be anchored or transported retrogradely. Proteins in the leading edge membrane experience a force orthogonal to the membrane when they bind to actin in the lamellipodium. The force keeping them in the lipid bilayer provides the force counteracting retrograde transport and they are therefore not swept away by retrograde flow. Keren and Theriot remark on the observation that actin binding proteins at the leading edge do not flow rearward "The lack of lipid flow, together with the presence of a diffusion barrier at the

leading edge, imply that physical trapping may be sufficient for maintaining the localization of various essential membrane-bound components there" [72].

In summary, it has been shown that F-actin-membrane-binding is necessary for the formation of lamellipodia and that activated linker proteins are at the leading edge.

The effect of binding of F-actin to the membrane on the shape of lamellipodia favors larger coherent structures, as mentioned above [35, 144]. This suggests that the occurrence of pushing and pulling filaments at the leading edge does not strongly distort the membrane on the length scale of typical filament distances. This is supported by another estimate. We can obtain an idea about the scale on which cellular forces cause membrane distortion from an estimate of the critical radius for bleb formation. Blebbing occurs at patches of membrane not bound to the actin cortex. The pressure difference across the membrane drives blebbing. Membrane tension and resistance to bending counteract deformation and cause a minimal critical radius of the unattached membrane patch. Sheetz et al. estimated it to be about 470 nm [117]. Hence, the critical diameter is at least by a factor of nine larger than typical distances of filaments in lamellipodia, if calculated from filament density measurements (100/μm, lamellipodium height 200 nm). More recently, filament distance in lamellipodia was estimated to be even 30 nm only [137]. In summary, there are good reasons to assume that membrane distortion is negligible on the length scale of filament distances. Modeling methods for dealing with membrane shape on larger length scales have been published [44, 73].

Modeling has shown that transient binding is compatible with protrusion [43, 43, 98, 147]. Based on these considerations, we conclude that the experimental evidence strongly suggests inclusion of F-actin-membrane binding into lamellipodium leading edge models.

4 Dynamic Regimes of Actin-Based Motion

When placed on a substrate, cells spread and eventually start moving spontaneously or as a result of mechanical or chemical stimulation. Sometimes cells are found to be testing the substrate, the topology of which influence the behavior [109]. The movement of the cell boundary can occur continuously or in cycles of protrusion and retraction. Mouse embryonic fibroblasts spreading on a fibronectin-coated glass show phase transitions from a resting state to a state of fast and continuous spreading and further to periodic membrane retractions [39]. Lateral membrane waves with a lateral speed of about 100 nm/s have been observed in a variety of spreading cells, including mouse embryonic fibroblasts, T cells, as well as wing disk cells from fruit flies [40]. For keratocytes, the leading edge morphology seems to be coupled to the motile behavior – coherent, smooth cells migrate significantly faster than decoherent, rough cells [82]. Epithelial cells show three different protrusion phenotypes: A state where long cell edge sectors are synchronized in cycles of protrusion and retraction, a state where random bursts of protrusion initiate protrusion waves propagating transversally in both directions, and a state where

continuous protrusion is occasionally interrupted by self-propagating ruffles [91]. Cells switch between states depending on the Rac1 activation level and the PAK and Arp2/3 concentrations. Increased Rac1 levels lead to increased activation of Arp2/3 and inhibition of cofilin via PAK [41, 65]. Arp2/3 nucleates filaments on existing filaments [64,127], and cofilin severs filaments and promotes their depolymerization [6,15,84]. Changing the activities of Rac1, PAK, and Arp2/3 results therefore into a change of the most important parameters of the actin network – density, length, and growth velocity. Thus, experiments show that in principle, the structure and function of the actin network inside the cell can be mapped into the external shape dynamics, which can be observed without interfering directly with the cell.

Experiments on model systems, such as protein-coated beads or fluid droplets placed in a motility medium, are helpful in understanding the motile machinery inside a cell. The motion of protein-coated plastic beads can be smooth or saltatory, depending on the bead radius and the surface concentration of the protein [9, 105]. Also, deformable lipid vesicles show both regimes of motion, and can reach up to 10 μm/min, compared to 3–4 μm/min for beads. A comparative study comes to the conclusion that hard and fluid actin propelled objects rely on different mechanisms to establish and maintain directed movement: Stress relaxation within the actin gel prevents the accumulation of filaments at the front of moving beads, while segregation of nucleators reduces actin polymerization at the front of moving vesicles [33]. Similarly, oil droplets can show continuous or hopping motion in a motility assay [12, 135]. The probability for oscillatory movement is higher for smaller droplets, and the oscillatory mechanism seems to be based on diffusion and convection of the surface protein activating actin polymerization.

5 Modeling Concepts

We distinguish in the following between continuum and filament models. This classification is not based on the mathematical form of the model, but rather on the primary treatment of the actin cytoskeleton. Continuum models start from the theory for visco-elastic gels and the filament properties enter via constitutive equations and material constants. Filament models start from the properties of single filaments and investigate how a population or network composed of them behaves.

5.1 Continuum Models

Part of the theoretical work on cell motility has been done within the framework of continuum models. Such models treat the cytoskeleton as a continuum medium and do not consider the microscopic details of the force generation process. Existing continuum models are based on various physical theories and differ in the choice of the state variables used to describe the cytoskeleton.

Several approaches focus on the biochemical processes inside the cell. The dynamics of the cytoskeleton are thereby described by the concentration of

actin filaments as well as of the regulatory proteins controlling their growth, leading to coupled systems of differential equations. Mogilner et al. establish reaction–diffusion equations for the actin monomers in their different forms (ADP–G-actin–ADF/cofilin, ADP–G-actin–profilin, ATP–G-actin–profilin, and ATP–G-actin–thymosin β_4 complexes) and include growth by polymerization of the barbed ends of actin filaments, capping and depolymerization [96]. The study calculates stationary velocities as stationary solution of the set of reaction diffusion equations in dependence on concentrations of capping proteins, thymosin β, profilin, and other biochemical parameters. Its force balance at the leading edge includes pushing forces from polymerizing filaments and a constant force as membrane resistance. Force dependence of polymerization and the limitations by G-Actin flux toward the front lead to an optimal filament density for a given membrane resistance.

A study by Grimm et al. aims at predicting the shape of the leading edge [59]. It models the dynamics of the density of right and left oriented barbed ends by considering growth, branching, and capping but not retrograde flow or filament attachment. The resistance of the membrane to motion is a constant force. Consequently, leading edge velocity increases with filament density in that model. The feedback for the shape of the leading edge to the actin density increases densities at local protrusions. This positive feedback loop may cause shape instability at high capping rate. The model predicts well the leading edge shape of fish keratocytes at low capping rates. The theory was supplemented by G-actin consumption by growth and membrane tension in the stability analysis in [73, 82].

Dawes et al. [31] consider the spatial distribution of actin filaments and their barbed ends in a simplified 1D geometry. The model includes diffusion of the Arp2/3 complex, force-dependent polymerization, retrograde flow, spontaneous nucleation, tip and side branching as well as capping and depolymerization. As in many models of this type, the protrusion rate is proportional or equal to the polymerization rate. Increasing the rate of nucleation of filaments (by the actin related protein Arp2/3) or the rate of actin polymerization leads to faster cell speed, whereas increasing the rate of capping or the membrane resistance reduces cell speed in this study. A simple model [49] considers the densities of barbed and pointed ends, coupled to a reaction–diffusion equation for the concentration of actin monomers and allows for the description of the polarization of an initially symmetric cytoskeleton and the initiation of motion.

A very extensive model has been developed in [10, 11]. It provides a method to solve the complete nucleotide profile within filaments by considering the cycle of actin-assembly and disassembly, including many details such as ATP hydrolysis and the role of profilin in the nucleotide exchange.

Other models focus on the mechanics of the cytoskeleton, which is treated as viscous or visco-elastic fluid. References [2, 81] consider two dynamic components: the cytosol, treated as a Newtonian fluid, and the polymerized actin filaments, treated as an elastic medium. Adhesion kinetics is considered here through a frictional force on the filamentous phase. The idea of a two-phase network has been elaborated further in [103], where a nonlocal pressure term modeling long-range network compaction was included. A variety of models consider one-dimensional

visco-elastic strips as a model for a radial cross-section through the lamellipodium [58, 77, 86]. Gracheva and Othmer consider a one-dimensional visco-elastic cell in contact with a viscous substrate [58]. The inclusion of graded adhesion (strong at the front, weak at the rear) allows for reproduction of the bell-shaped dependence of the cell velocity on adhesion strength [37, 58].

In [51], the actin network around a bacteria is treated by an elastic approach. Filament growth on the bacterial surface produces here elastic stresses that propel the bacterium forward. The same idea is used in [70] for the study of the symmetry break at the formation of the actin tail around a propelled bead in a biomimetic assay. Reference [126] treats the cell as an incompressible, visco-elastic solid and uses classical mass balance and equilibrium equations to describe its motion. This model allows to make predictions about the traction patterns on the substrate.

Based on a generic theory for active polar gels [76, 77], a model for the lamellipodium motion was developed in [78]. Here, the cytoskeleton is treated as a viscous polar gel. Myosin contraction in the cytoskeleton is included through an additional intrinsic anisotropic stress.

A model coupling membrane elasticity with actin polymerization has been proposed in [119] to explain membrane waves driven by actin and myosin. The wave mechanism is based on the presence of freely diffusing membrane proteins, the curvature of which influences the morphodynamic pattern of the cell.

5.2 Filament Models

A first model aiming to explain how polymerization of actin filaments can produce the force of protrusion in migrating cells was proposed in [108]. This "Brownian ratchet" model considers the polymerization of a stiff filament against a barrier, upon which a load acts. The barrier is able to diffuse, and the ratchet mechanism is based on the intercalation of monomers between the barrier and polymer tip. This model has been extended to an "elastic Brownian ratchet" model in [97], by including the thermal motions of the polymerizing actin filaments. It was further extended in [98] by including transient attachment to the obstacle ("tethered ratchet").

The entropic force exerted by a grafted semiflexible polymer on a rigid obstacle has been calculated both analytically and by Monte Carlo simulations in [53]. Explicit scaling functions as well as analytical results for certain asymptotic regimes were found. These results were used in [43, 54] in a model for the actin-based propulsion of flat rigid obstacles. Polymerization, attachment to and detachment from the obstacle as well as cross-linking between filaments were considered. The model is used to find the dynamics of the length for attached and detached filaments, which is required for the computation of the total force on the obstacle. This approach has been extended to the propulsion of soft membranes under tension in [44] as well as of rigid spherical beads in [42]. The actin network is described here also by continuous state variables reflecting the densities and lengths of the actin filaments. However, in contrast to the models discussed in the previous section, the microscopic form of the force exerted by single actin filaments is taken into account here.

Several studies model the microscopic growth of the actin network explicitly. In [16], a stochastic simulation frame for an actin network growing against an obstacle is proposed, where single filaments and single subunits in each filament are considered. Growth, depolymerization, capping, and branching are included, allowing the prediction of the network growth velocity. Based on this model, the structure of branched actin networks has been analyzed in [18]. In [4], actin filaments are treated as rigid rods under volume exclusion. Polymerization, depolymerization, branching, and capping are simulated using a continuous-time Markov algorithm, allowing for the prediction of the angular distribution of the filaments with respect to the leading edge.

A mesoscopic network approach to the cross-linked actin network has been proposed in [32]. Here, an Accumulative Particle-Spring model that builds on the elastic gel model [51] is used. Network links have no direct correspondence to actin filaments, but the bulk visco-elastic properties of the chains of nodes and springs are intended to capture the bulk visco-elastic properties of the actin network.

5.3 Coupling of Membrane and Cytoskeletal Dynamics

The common goal of most modeling approaches is finding the dynamics of the considered obstacles, e.g., the regime of motion of a bead or bacteria, or morphology of the leading edge. To this end, the dynamics of the cytoskeleton has to be coupled to the mechanics of the membrane. Most models do not include directly this interaction, but assume that the membrane moves at the growth velocity of the network. A model focusing on the membrane–cytoskeleton coupling has been proposed in [147]. It combines a filament model [54] for the filament tips that reach to the leading edge with and a continuum description of the cross-linked part of the actin network farther in terms of the active polar gel model [78]. Thereby, the filament model provides the force boundary condition for the visco-elastic part of the network. In return, the flow of this network provides the grafting points of the filament tips described by the filament model. This allows for the calculation of the total force exerted on the membrane that is used to find its dynamics.

6 Mechanisms Suggested by Models

6.1 Comparison of Model Assumptions

One of the main differences between continuum and filament models lies in the way the interaction between the actin network inside the cell and the cell membrane is included. For continuum models, the interaction force is assumed to be either a given constant [78] or to depend mainly on the membrane geometry, e.g., on the curvature [2, 119, 120]. Filament models include often the length, position, and orientation of actin filaments, which allow for a more accurate calculation of the entropic force exerted on the membrane [4, 5, 16, 17, 43, 44, 53, 54, 147].

Some continuum models require the knowledge of force boundary conditions, that have to be included artificially, by assuming for example a constant external force. In [78], such a boundary condition is needed to determine the force profile in a gel strip across the leading edge and critically influences the resulting leading edge velocity. Different other models calculate the interaction force in various ways. In the two-phase flow model [2], the pressure exerted by both filamentous phase and solvent phase are included explicitly. The analysis of *Listeria* propulsion in [51] and the study of symmetry breaking leading to actin tail formation [70] assume the interaction between the actin network and the obstacle relies on the formation of an elastic stress at the obstacles surface, due to the creation of a new layers of gel through polymerization. In [119, 120], membrane waves based on the competition between protrusive forces due to actin, and contractile forces due to myosin are studied. Here, the protrusion force is assumed to be proportional to the local concentration of membrane proteins driving actin polymerization. Additionally, membrane tension and elastic force are included. A common feature of these continuum models is that the assumed interaction between the actin network and membrane/obstacle involves almost exclusively properties of the membrane or obstacle, and not of the actin filaments, that are not modeled explicitly. However, it is known that the entropic force exerted by single actin filaments on an obstacle depends strongly on their fluctuating length and their position and orientation with respect to the obstacle being pushed [53, 97]. Additionally, in most continuum models and many filament models describing filaments as stiff rods, the sum of protrusion velocity and retrograde flow velocity equals the (effective, projected) polymerization velocity. However, experiments showed that this is not always the case [69, 89]. In order to include these observations, the properties of the filaments close to the leading edge have to be modeled explicitly like, e.g., in [43, 44, 147].

The growth of an actin network against an obstacle has been simulated in [16]. The approach includes the position of single filament tips, which is considered for the calculation of the total force on the obstacle. The dependence of this force on the filament orientation has been included in [17]. The response of filaments to force depends sensitively on the freely fluctuating length between the graft point closest to the leading edge and the filament tip experiencing the force [53]. That dependence is crucial in understanding different dynamic regimes of cell motion. Long free lengths yield slow edge velocities because filaments are too floppy to exert a strong pushing force and cell motion may even pause or stop if filaments become too long and floppy [43, 75]. Short free lengths yield slow velocities due to the polymerization rate limitation by strong force [8, 43, 75].

Explicit consideration of the length dynamics of actin filaments [43, 44, 54] allows to include the dependence of the interaction force on the free fluctuation length of the filaments. Initially, models made simplifying assumptions on the dynamics of the cross-linking points of the filaments that are critical ingredients in determining the force. Recently, a model combining a gel description of the actin network, a cross-linking dynamics accounting for diffusion of free cross-linkers and a filament description of the boundary has been studied [147]. The filament model provides here an accurate force boundary condition for the gel model, that, in turn, allows for the proper calculation of the filament position and length dynamics.

Another controversy between several models concerns the relation between local membrane velocity and local growth velocity of the actin network. Several models assume for simplicity that the membrane moves with the mean polymerization velocity of the actin network [31, 78, 113, 119, 120]. This is a strong constraint, meaning that the relative position of the filament tips with respect to the membrane is assumed to be constant. Such a constraint is realistic only during steady motion of a cell, for example, a crawling fish keratocyte. Experimentally, it has been shown that time shifts up to 20 s between the maxima of protrusion and polymerization velocity at the leading edge are possible in dynamic regimes with oscillatory motion [69].

Most models do not reproduce this phase shift between polymerization and protrusion, and that has to be seen in connection with the force balance at the leading edge, the way retrograde flow is included, the force–velocity relation and the relation between polymerization and leading edge motion. If the leading edge velocity is equated with the polymerization velocity (in some models subtracting a constant retrograde flow), there is no phase shift between protrusion and polymerization and the force–velocity relation will reflect the force dependence of the polymerization rate. However, measurements with fish keratocytes showed that the force–velocity relation is different from the force dependence of polymerization [63, 112].

Several models do not couple actin network and boundary motion by the same velocity, but by the same interaction force, according to Newton's third law. The processes contributing to the force balance and the relation between force and gel flow as well as force and membrane velocity then decide whether the measured force–velocity relation for the whole cell and the measured phase behavior are explained by the model. Many studies assume a linear relation between the total force exerted on the membrane and the resulting membrane velocity [16, 43, 44, 53, 119, 120, 147]. That relation results from the assumption of a viscous drag to over-damp membrane velocity dynamics. This drag comprises viscous drag from the external medium and the transport of membrane to the protruding parts of the cell. As mentioned above, some models describe membrane resistance as a constant force.

The force driving protrusion is due to polymerizing filaments in lamellipodial motion [1, 14, 73, 110]. The force with which these filaments push against the membrane determines the polymerization rate [53, 98]. Models for the motion of protein-coated beads include also the force exerted by attached filaments on the obstacle surface [9, 42, 98]. Groswasser et al. derive a bi-phasic friction force–velocity relation from this transient attachment of filaments [9]. It causes an additional friction force proportional to the velocity at small velocities. At high velocities, this additional friction force vanishes, since the time during which filaments are attached drops at a certain velocity. Thus, for high velocities, the proportionality constant between force and velocity is reduced. This bi-phasic friction may lead to bead velocity oscillations [9]. Interestingly, velocity oscillations are possible also when the friction force–velocity relation is assumed to be linear, but attachment to the obstacle is considered in the computation of the total force on the obstacle by separating the dynamics of attached and detached filaments and tracking their mean length [43, 54]. This explicit consideration of attached filaments

comprises contributions of each of them to the force balance at the leading edge, which again may change the phase relation between protrusion and polymerization.

If there are only viscous or constant forces resisting motion, each change in the polymerization force necessarily entails immediately a change of velocity without phase shift. Other forces resisting motion – like the one from attached filaments – can modulate this temporal relation such that an increase in polymerization may first increase retrograde flow or compress filaments close to the leading edge and only slowly or later protrusion. That can be investigated by models like introduced in [147].

Mathematical models for cell motility vary further in the complexity of the biomechanical processes considered. Creation of new filaments by nucleation has been explicitly included in some filament models, either by assuming creation on existing filaments (autocatalytic model) or free creation on subsequent attachment to existing filaments (nucleation model) [4, 17–19, 22, 24, 90]. Similarly, filament capping and severing has been included explicitly in [21,23,90,95,96]. By contrast, other filament models assume implicitly that the processes involving creation and severing of filaments are balanced, such that a steady state with a constant number of active filaments is reached [43,44,53,97,147].

Several models include contraction of myosin motors explicitly, e.g., [20, 78, 86, 102, 113, 119, 120, 147]. Similarly, adhesion to the substrate might be explicitly included [78,81,102]. Other models neglect these processes, under the tacit assumption that protrusion at the leading edge is decoupled from attachment/detachment to the substrate and contraction of the cell body. Attachment to the membrane or the surface of propelled artificial objects is explicitly included in several models [44,53,90,98], motivated by different attachment mechanisms found experimentally [27, 52, 80]. The whole actin cycle including the major regulatory mechanisms has been modeled in [10, 11].

6.2 Comparison of Sets of Experiments Explained by Models

The morphodynamics of crawling cells has been analyzed in several experimental conditions and with different cell types. Using various analysis techniques and computational tools, high-resolution membrane velocity maps along the leading edge can be obtained from processing experimental images.

Velocity maps of crawling cells show distinguishable morphodynamic patterns, some of which seem to be characteristic to the cell type under the given experimental conditions. Experiments with spreading cells show lateral membrane waves [40], periodic lamellipodial contractions [56, 143], and phase transitions between different morphodynamic pattern during the spreading process [39]. Observation of different types of migrating cells has shown traveling waves with different profiles, like protrusions spreading laterally from one point of the membrane in both directions, traveling retractions, and slightly spatially modulated velocity oscillations [91].

Early models for cell migration aim to explain the global movement characteristics of the cell, e.g., traveling speed and steady height profile [31, 78, 86, 96, 97], shape determination of motile cells [73, 113] or the structure of the actin network [4, 16]. Later models include the morphology of the leading edge. The existence of periodic traveling waves along the lamellipodium can be explained by combining protrusion forces due to polymerization and contraction forces due to the presence of myosin motors [119, 120]. The wave mechanism described by Shlomovitz et al. requires molecules inducing lateral curvature of the leading edge and myosin activity [119]. However, at least some types of waves do not depend on myosin activity [40, 91]. An alternative wave formation mechanism, based on the competition between protrusive forces due to detached, polymerizing filaments and pulling forces due to attached filaments has been proposed in [44]. It reproduces the laterally traveling protrusions, the modulated velocity oscillations and the Rac-induced transition between both patterns. In agreement with experiments, the lateral velocity of protrusions is independent from cell velocity and both patterns do not depend on myosin activity [44].

The velocity oscillations of *Listeria* bacteria have been modeled with the elastic gel theory by Gerbal et al. [51]. This theory offers an explanation for oscillations due to a competition between actin gel growth from the sides and growth from the back of the bacterium, with different velocities and strengths for each. While the simulated period agrees well with experiments, velocity amplitudes are about one order of magnitude larger than measured values [51]. The filament model by Gholami et al. including dynamics of free filament length and filament attachment to the bacterium reproduces *Listeria* velocity oscillations quantitatively with respect to periods and amplitudes [54].

The validity of model predictions can be further tested with the help of biomimetic systems, where various parameters can be changed, in contrast to migrating or spreading cells. Experiments on protein-coated spherical beads [9, 33] and oil droplets [12, 33] have revealed different regimes of motion – continuous and oscillatory, and identified parameters that might induce transitions from one state of motion to the others. The same model as used by Gholami et al. for *Listeria* with slightly changed parameter values also reproduces the velocity oscillations observed with oil droplets including the onset of oscillations due to weakening of filament attachment by VASP [43, 134]. The mechanism has periodic attachment and detachment of filaments as central processes [43, 134]. There are several theoretical studies on bead motion characteristics [9, 42, 51, 90, 98]. Mogilner and Oster demonstrated the compatibility of attachment and propulsion by polymerization by the ground-breaking tethered ratchet model [98]. Elastic gel theory explains velocity oscillations of protein-coated beads [9] by a mechanism called the "soap effect", "because it recalls the rapid motion of a wet bar of soap slipping away as it is slowly squeezed by hand". [51]. This is mainly justified by scaling arguments for maximal velocities of the oscillations and the threshold for the onset of oscillations in dependence on the bead radius [9, 51]. The oscillation mechanism relies on a curved obstacle surface and the bi-phasic dependence of the friction force on bead velocity mentioned in the previous section. Bead motion has also been investigated

by the model used for *Listeria* motion, oil droplets, and morphodynamic phenotypes [42–44,54]. The oscillation mechanism is essentially the same as with oil droplets or *Listeria*. That model is able to simulate the velocity oscillations quantitatively with respect to periods and amplitudes except a small shift in the average velocity [42]. It also reproduces the dependence of the onset of oscillations on the bead radius and protein density on the bead surface.

While there are many models explaining very well certain aspects or systems of cell motility or morphodynamics, the appeal of the modeling concept including the dynamics of free fluctuating length of filaments, filament binding dynamics at the obstacle surface, force-dependent polymerization and – if required – nucleation and capping is for us the reproduction of experimental results with a variety of systems in a very intuitive way [42, 43, 45, 53, 147]. It also offers a natural explanation for the variety of dynamic regimes observed in cell motility, morphodynamics, and biomimetic systems.

7 Open Problems

According to our view on the field, there are three conceptually highly relevant problems the solution of which could advance the field: (1) Despite the molecular similarities between the variety of biological and biomimetic systems, there is not a unifying theory or model. (2) There is no satisfying theoretical explanation of the force–velocity relation of fish keratocytes. (3) The phase shift between protrusion and polymerization is unexplained and the function of the two functionally and structurally different regions of the lamellipodium – often described as lamellum and lamellipodium – has not been investigated theoretically.

The force–velocity relation of fish keratocytes must be shaped by the intracellular force generation mechanism. The compatibility of the ideas on force generation by actin polymerization with measured force–velocity relations has not been shown in a mathematical model yet. But this if of course required for a consistent theory on cell motility. The force–velocity relation exhibits a dramatic velocity drop upon first contact with an AFM cantilever or glass fiber followed by a concave-down relation in the slow-velocity regime [63, 112]. The discussion around it has focused on an explanation for the concave down part since this shape is in contradiction to the convex shape of the force exponential dependence of the actin polymerization rate [53, 97]. Most theoretical studies essentially neglected the initial velocity drop. Simulations of branched actin networks made of rigid rod-like filaments with excluded volume effects taken into account [116] produce a concave-down force–velocity relation. However, they predict stall forces by a factor of 20–50 too large. Brownian dynamics simulations of stiff actin filaments in a branched network [88] also give rise to a concave shaped force–velocity curve but velocities at half stall force are orders of magnitude faster than in experiments with fish keratocytes and stall forces are by a factor 200–400 too small. No retrograde flow is found in actin networks growing under an AFM cantilever [107]. The shape

of the force–velocity curve of those systems resembles that of keratocytes, though on much different scales – it takes more than 100 min to reach the stall force which is in the order of 300 nN. Weichsel and Schwarz suggested to explain this behavior by a configurational bistability of the actin network [142]. However, that bistability has recently been excluded as the mechanism of the force–velocity relation in fish keratocytes by Heinemann et al. [63]. They repeated the measurement with the same cell and a time lag of 30–40 s. The second measurement should have provided results different from the first one, if the actin network exhibited configurational bistability and a state transition shaped the force–velocity relation. However, two repetitions showed the same outcome as the first measurement. Heinemann et al. excluded the autocatalytic branching model by the same reasoning. That model explains the plateau after the initial drop by growing filament density [16]. Such an increase in density should also affect the second and third measurement, according to [63], what was not observed.

We subsume the phase shift between polymerization and protrusion and the structure of the lamellipodium under one problem, because it is likely that the phase shift depends on structural elements close to the leading edge membrane showing dynamics which has not been accounted for by mathematical models yet. Several processes may contribute to such a phase shift: Polymerization drives not only protrusion but also retrograde flow, the region close to the leading edge might be much softer than modeled until now, protrusion might not only depend on polymerization forces but also on the binding of filaments to the membrane pulling it back. Zimmermann et al. have recently suggested a model including these processes but it has not been applied to the problem yet [147].

The vision of modeling of morphodynamics is to infer at least in part the state of signaling pathways and the cytoskeleton from observing the changes of cell shape and velocity. The starting point can be a biomechanical model in terms of the elemental processes and rates like polymerization, depolymerization, capping and nucleation, elastic responses, retrograde flow, membrane tension, etc. The above-mentioned problems show that this still has to be established. Modeling of the control of the parameters of such a model by signaling pathways can then lead to a comprehensive understanding of morphodynamics and motility.

References

1. Abraham V, Krishnamurthi D, Taylor D, Lanni F (1999) The actin-based nanomachine at the leading edge of migrating cells. Biophys J 77:1721–1732
2. Alt W, Dembo M (1999) Cytoplasm dynamics and cell motion: two phase flow models. MathBiosci 156:207–228
3. Ananthakrishnan R, Ehrlicher A (2007) The forces behind cell movement. Int J Biol Sci 3(5):303–317
4. Atilgan E, Wirtz D, Sun S (2005) Morphology of the lamellipodium and organization of actin filaments at the leading edge of crawling cells. Biophys J 89:3589–3602
5. Atilgan E, Wirtz D, Sun S (2006) Mechanics and dynamics of actin-driven thin membrane protrusions. Biophys J 90:65–76

6. Bamburg J (1999) Proteins of the adf/cofilin family: essential regulators of actin dynamics. Ann Rev Cell Dev Biol 15:185–230
7. Baumgartner M, Sillman A, Blackwood E, Srivastava J, Madson N, Schilling J, Wright J, Barber D (2006) The nck-interacting kinase phosphorylates erm proteins for formation of lamellipodium by growth factors. PNAS 103:13391–13396
8. Bear JE, Svitkina TM, Krause M, Schafer DA, Loureiro JJ, Strasser GA, Maly IV, Chaga OY, Cooper JA, Borisy GG, Gertler FB (2002) Antagonism between ena/vasp proteins and actin filament capping regulates fibroblast motility. Cell 109(4):509–521
9. Bernheim-Groswasser A, Prost J, Sykes C (2005) Mechanism of actin-based motility: a dynamic state diagram. Biophys J 89:1411–1419
10. Bindschadler M, McGrath J (2007) Relationships betwenn actin regulatory mechanisms and measurable state variables. Annals Biomed Eng 35:995–1011
11. Bindschadler M, Osborn E, Dewey C, McGrath J (2004) A mechanistic model of the actin cycle. Biophys J 86:2720–2739
12. Boukellal H, Campas O, Joanny J, Prost J, Sykes C (2004) Soft listeria: actin-based propulsion of liquid drops. Phys Rev E 69:061906
13. Bugyi B, Didry D, Carlier M (2010) How tropomyosin regulates lamellipodial actin-based motility: a combined biochemical and reconstituted motility approach. EMBO J 29:14–26
14. Carlier M, Pantaloni D (2007) Control of actin assembly dynamics in cell motility. J Biol Chem 282(32):23005–23009
15. Carlier M, Ressad F, Pantaloni D (1999) Control of actin dynamics in cell motility – role of adf/cofilin. J Biol Chem 274:33827–33830
16. Carlsson A (2001) Growth of branched actin networks against obstacles. Biophys J 81: 1907–1923
17. Carlsson A (2003) Growth velocities of branched actin networks. Biophys J 84:2907–2918
18. Carlsson A (2004) Structure of autocatalitically branched actin solutions. Phys Rev Lett 92(23):239102
19. Carlsson A (2005) The effect of branching on the critical concentration and average filament length of actin. Biophys J 89(1):130–140
20. Carlsson A (2006) Contractile stress generation by actomyosin gels. Phys Rev E 74:051912
21. Carlsson A (2006) Stimulation of actin polymerization by filament severing. Biophys J 90:413–422
22. Carlsson A (2010) Dendritic actin filament nucleation causes traveling waves and patches. Phys Rev Lett 104:228102
23. Carlsson A, Sept D (2008) Mathematical modeling of cell migration. Biophys Tools Biol: Vol 1 In Vitro Techniques 84:911
24. Carlsson A, Wear M, Cooper J (2004) End versus side branching by arp2/3 complex. Biophys J 86:1074–1081
25. Casteo L, Jay D (1999) Radixin is involved in lamellipodial stability during nerve growth cone motility. Mol Biol Cell 5:1511–1520
26. Chauvière A, Preziosi L, Verdier C (eds) (2010) Cell mechanics. From single scale-based models to multiscale modeling. Taylor & Francis Group, Chapman & Hall/CRC Mathematical and Computational Biology Series, Boca Raton
27. Co C, Wong D, Gierke S, Chang V, Taunton J (2007) Mechanism of actin network attachment to moving membranes: barbed end capture by n-wasp wh2 domains. Cell 128:901–913
28. Cooper JA, Sept D (2008) New insights into mechanism and regulation of actin capping protein. Int Rev Cell Mol Biol, 267:183–206
29. Danuser G (2005) Coupling the dynamics of two actin networks – new views on the mechanics of cell protrusion. Biochem Soc Trans 33:1250–1253
30. Danuser G (2009) Testing the lamella hypothesis: the next steps on the agenda. J Cell Sci 122:1950–1962
31. Dawes A, Ermentrout G, Cytrynbaum E, Edelstein-Keshet L (2006) Actin filament branching and protrusion velocity in a simple 1d model of a motile cell. J Theor Biol 242:265–279

32. Dayel M, Akin O, Landeryou M, Risca V, Mogilner A, Mullins R (2009) In silico reconstitution of actin-based symmetry breaking and motility. PLoS Biol 7:e1000201
33. Delatour V, Shekhar S, Reymann AC, Didry D, Lê KHD, Romet-Lemonne G, Helfer E, Carlier MF (2008) Actin-based propulsion of functionalized hard versus fluid spherical objects. New J Phys 10(2):025001
34. Delorme V, Machacek M, DerMardirossian C, Anderson K, Wittmann T, Hanein D, Waterman-Storer C, Danuser G, Bokoch G (2007) Cofilin activity downstream of pak1 regulates cell protrusion efficiency by organizing lamellipodium and lamella actin networks. Dev Cell 13:646–662
35. Denker S, Barber D (2002) Cell migration requires both ion translocation and cytoskeletal anchoring by the Na–H exchanger NHE1. J Cell Biol 159:1087–1096
36. Denker S, Huang D, Orlowski J, Furthmayr H, Barber D (2000) Direct Binding of the Na H Exchanger NHE1 to ERM Proteins Regulates the Cortical Cytoskeleton and Cell Shape Independently of H^+ Translocation. Mol Cell 6:1425–1436
37. DiMilla PA, Barbee K, Lauffenburger DA (1991) Mathematical model for the effects of adhesion and mechanics on cell migration speed. Biophys J 60:15–37
38. DiNubile MJ, Huang S (1997) High concentrations of phosphatidylinositol-4,5-bisphosphate may promote actin filament growth by three potential mechanisms: inhibiting capping by neutrophil lysates, severing actin filaments and removing capping protein-[beta]2 from barbed ends. Biochim Biophys Acta (BBA) – Mol Cell Res 1358(3):261–278
39. Doebereiner HG, Dubin-Thaler B, Giannone G, Xenias H, Sheetz M (2004) Dynamic phase transitions in cell spreading. Phys Rev Lett 93(10):108105
40. Doebereiner HG, Dubin-Thaler B, Hofman J, Xenias H, Sims T, Giannone G, Dustin M, Wiggins C, Sheetz M (2006) Lateral membrane waves constitute a universal dynamic pattern of motile cells. Phys Rev Lett 97(3):038102
41. Eden S, Rohatgi R, Pdtelejnikov A, Mann M, Kirschner M (2002) Mechaniscm of regulation of wave1-induced actin nucleation by rac1 and nck. Nature 418:790–793
42. Enculescu M, Falcke M (2011) Actin-based propusion of spatially extended objects. New J Phys 13:053040
43. Enculescu M, Gholami A, Falcke M (2008) Dynamic regimes and bifurcation in a model of actin-based motility. Phys Rev E 78:031915
44. Enculescu M, Sabouri-Ghomi M, Danuser G, Falcke M (2010) Modeling of protrusion phenotypes driven by the actin-membrane interaction. Biophys J 98:1–11
45. Faber, M, Enculescu, M, Falcke, M (2010) Filament capping and nucleation in actin-based motility. Eur Phys J Special Topics 191:147–158, DOI 10.1140/epjst/e2010-01347-3, URL http://dx.doi.org/10.1140/epjst/e2010-01347-3
46. Fievet B, Gautreau A, Roy C, Del Maestro L, Mangeat P, Louvard D, Arpin M (2004) Phosphoionositide binding and phosphorylation act sequentially in the activation mechanism of ezrin. J Cell Biol 164:653–659
47. Fievet B, Louvard D, Arpin M (2007) Erm proteins in epithelial cell organization and functions. BBA – Mol Cell Res 1773:653–660
48. Friedel P, Hegerfeld Y, Tusch M (2004) Collective cell migration in morphogenesis and cancer. Int J Dev Biol 48:441–449
49. Fuhrmann J, Ks J, Stevens A (2007) Initiation of cytoskeletal asymmetry for cell polymerization and movement. J Theor Biol 249:278–288
50. Gautreau A, Louvard D, Arpin M (2000) Morphogenic effects of ezrin require a phosphorylation-induced transition from oligomers to monomers at the plasma membrane. J Cell Biol 150:193–203
51. Gerbal F, Chaikin P, Rabin Y, Prost J (2000) An elastic analysis of listeria monocytogenes propulsion. Biophys J 79:2259–2275
52. Gerbal F, Laurent V, Ott A, Carlier M, Chaikin P, Prost J (2000) Measurement of the elasticity of the actin tail of listeria monocytogenes. Eur Biophys J with Biophys Lett 29:134–140
53. Gholami A, Wilhelm J, Frey E (2006) Entropic forces generated by grafted semiflexible polymers. Phys Rev E 74(4):041803

54. Gholami A, Falcke M, Frey E (2008) Velocity oscillations in actin-based motility. New J of Phys 10:033022
55. Ghosh M, Sonx X, Mouneimne G, Sidani M, Lawrence D, Condeelis J (2004) Cofilin promotes actin polymerization and defines the direction of cell motility. Science 304: 743–746
56. Giannone G, Dubin-Thaler J, Doebereiner HG, Kieffer N, Bresnick A, Sheetz M (2004) Periodic lamellipodial contractions correlate with rearward actin waves. Cell 116: 431–443
57. Giannone G, Dubin-Thaler B, Rossier O, Cai Y, Chaga O, Jiang G, Beaver W, Dobereiner H, Freund Y, Borisy G, Sheetz M (2007) Lamellipodial actin mechanically links myosin activity with adhesion-site formation. Cell 128:561–575
58. Gracheva MA, Othmer HG (2004) A continuum model of motility in ameboid cells. Bull Math Biol 66:167–193
59. Grimm H, Verkhovsky A, Mogilner A, Meister JJ (2003) Analysis of actin dynamics at the leading edge of crawling cells: implications for the shape of keratocyte lamellipodia. Eur Biophys J 32:563–577
60. Grinstien S, Woodside M, Waddell T, Downey G, Orlowski J, Pouyssegur J, Wong D, Foskett J (1993) Focal localozation of the nhe-1 isoform of the na1/h1 antiport: assesment of effects on intracellular ph. EMBO J 12:5209–5218
61. Hartwig JH, Chambers KA, Stossel TP (1989) Association of gelsolin with actin filaments and cell membranes of macrophages and platelets. J Cell Biol 108(2):467–479
62. Hartwig JH, Bokoch GM, Carpenter CL, Janmey PA, Taylor LA, Toker A, Stossel TP (1995) Thrombin receptor ligation and activated Rac uncap actin filament barbed ends through phosphoinositide synthesis in permeabilized human platelets. Cell 82(4):643–653
63. Heinemann F, Doschke H, Radmacher M (2011) Keratocyte lamellipodial protrusion is characterized by a concave force–velocity relation. Biophys J 100(6):1420–1427
64. Higgs H, Pollard T (2001) Regulation of actin filament network formation through arp2/3 complex: activation by a diverse array of proteins. Ann Rev Biochem 70:649–676
65. Huang T, DerMardirossian C, Bokoch G (2006) Cofilin phosphytases and regulation of actin dynamics. Curr Op Cell Biol 18:26–31
66. Ichetovkin I, Grant W, Condeelis J (2002) Cofilin produces newly polymerized actin filaments that are preferred for dendritic nucleation by the arp2/3 complex. Curr Biol 12:79–84
67. Ikenoya M, Hidaka H, Hosoya T, Suzuki M, Yamamoto N, Sasaki Y (2002) Inhibition of rhokinase-induced myristoylated alanine-rich c kinase substrate (marcks) phosphorylation in human neuronal cells by h-1152, a novel and specific rho-kinase inhibitor. J Neurochem 81: 9–16
68. Janmey PA, Stossel TPS (1987) Modulation of gelsolin function by phosphatidylinositol 4,5-bisphosphate. Nature 325:362–364
69. Ji L, Lim J, Danuser G (2008) Fluctuations of intracellular forces during cell protrusion. Nat Cell Biol 10(12):1393–1400
70. John K, Peyla P, Kassner K, Prost J, Misbah C (2008) Nonlinear study of symmetry breaking in actin gels: implications for cellular motility. Phys Rev Lett 100(6):068101
71. Kahsai A, Zhu S, Fenteany G (2010) G protein-coupled receptor kinase 2 activates radixin, regulating membrane protrusion and motility in epithelial cells. BBA – Mol Cell Res 1803:300–310
72. Keren K, Theriot J (2008) Biophysical Aspects of actin-based cell motility in fish epithelial keratocytes. In: Cell motility, Springer New York, pp 31–58
73. Keren K, Pincus Z, Allen G, Barnhart E, Marriott G, Mogilner A, Theriot A (2008) Mechanism of shape determination in motile cells. Nature 453:485–U1
74. Klein M, Seeger P, Schuricht B, Alper S, Schwab A (2000) Polarization of Na^+/H^+ and Cl^-/HCO_3^- exchangers in migrating renal epithelial cells. J Gen Physiol 115:599–608
75. Koester S, Auinger S, Vinzenz M, Rottner K, Small J (2008) Differentially oriented populations of actin filaments generated in lamellipodia collaborate in pushing and pausing at the cell front. Nat Cell Biol 10:306–313

76. Kruse K, Joanny J, Jlicher F, Prost J, Sekimoto K (2004) Asters, vortices, and rotating spirals in actove gels of polar filaments. Phys Rev Lett 92(7):078101
77. Kruse K, Joanny J, Jlicher F, Prost J, Sekimoto K (2005) Generic theory of active polar gels: a paradigm for cytoskeletal dynamics. Eur Phys J E 16:5–16
78. Kruse K, Joanny J, Jlicher F, Prost J (2006) Contractility and retrograde flow in lamellipodium motion. Phys Biol 3:130–137
79. Kuhn JR, Pollard TD (2007) Single molecule kinetic analysis of actin filament capping. J Biol Chem 282(38):28014–28024
80. Kuo S, McGrath J (2000) Steps and fluctuations of listeria monocytogenes during actin-based motility. Nature 407:1026–1029
81. Kuusela E, Alt W (2009) Continuum model of cell adhesion and migration. J Math Biol 58:135–161
82. Lacayo C, Pincus Z, VanDujin M, Wilson C, Fletcher D, Gertler F, Mogilner A, Theriot J (2007) Emergence of large-scale cell morphology and movement from local actin filament growth dynamics. PLoS Biol 5(9):2035–2052
83. Lagana A, Vadnais J, Le P, Nguyen T, Laprade R, Nabi I, Noel J (2000) Regulation of the formation of tumor cell pseudopodia by the Na^+/H^+ exchanger NHE1. J Cell Sci 113: 3649–3662
84. Lai F, Bosse T, Szczodrak M, Benesch S, Auinger S, Faix J, Small J, Stradel T, Rottner K (2008) Arp2/3-complex regulation in motility and host–pathogen interaction. FEBS J 275:39
85. Lamb R, Ozanne B, Roy C, McGarry L, Stipp C, Mangeat P, Jay D (1997) Essential functions of ezrin in maintenance of cell shape and lamellipodial extension in normal and transformed fibroblasts. Curr Biol 7:682–688
86. Larripa K, Mogilner A (2006) Transport of a 1d viscoelastic actin–myosin strip of gel as a model of a crawling cell. Phys A 372:113–123
87. Le Clainche C, Carlier M (2008) Regulation of actin assembly associated with protrusion and adhesion in cell migration. Physiol Rev 88:89–513
88. Lee KC, Liu AJ (2009) Force–velocity relation for actin-polymerization-driven motility from brownian dynamics simulations. Biophys J 97(5):1295–1304
89. Lim JI, Sabouri-Ghomi M, Machacek M, Waterman CM, Danuser G (2010) Protrusion and actin assembly are coupled to the organization of lamellar contractile structures. Exp Cell Res 316(13):2027–2041
90. Lin Y (2009) Mechanics model for actin-based motility. Phys Rev E 79:021916
91. Machacek M, Danuser G (2006) Morphodynamic profiling of protrusion phenotypes. Biophys J 90:1439–1442
92. Marcy Y, Prost J, Carlier MF, Sykes C (2004) Forces generated during actin-based propulsion: a direct measurement by micromanipulation. PNAS 101(16):5992–5997
93. Matsui T, Yonemura S, Tsukita S, Tsukita S (1999) Activation of erm proteins in vivo by rho involves phosphatidyl-inositol 4-phosphate 5-kinase and not rock kinases. Curr Biol 9: 1259–1262
94. Méré J, Chahinian A, Maciver SK, Fattoum A, Bettache N, Benyamin Y, Roustan C (2005) Gelsolin binds to polyphosphoinositide-free lipid vesicles and simultaneously to actin microfilaments. Biochem J 386:47–56
95. Michalsky P, Carlsson A (2010) The effects of filament aging and annealing on a model lamellipodium undergoing disassembly by severing. Phys Biol 7:026004
96. Mogilner A, Edelstein-Keshet L (2002) Regulation of actin dynamics in rapidly moving cells: a quantitative analysis. Biophys J 83:1237–1258
97. Mogilner A, Oster G (1996) Cell motility drivern by actin polymerization. Biophys J 71:3030–3045
98. Mogilner A, Oster G (2003) Force generation by actin polymerization ii: the elastic ratchet and tethered filaments. Biophys J 84:1591–1605
99. Nagumo H, Ikenoya M, Sakurada K, Furuya K, Ikuhara T, Hiraoka H, Sasaki Y (2001) Rho-associated kinase phosphorylates marcks in human neuronal cells. Biochem Biophys Res Commun 280:605–609

100. Nakamura N, Pshiro N, Fukata Y, Amoano M, Fukata M, Kuroda S, Matsuura Y, Leung T, Lim K, Kaibuchi K (2000) Phosphorylation of erm proteins at filopodia induced by cdc42. Genes Cells 5:571–581
101. Ng T, Parsons M, Hughes W, Monypenny J, Zicha D, Gautreau A, Arpin M, Gschmeissner S, Verveer P, Bastiaens P, Parker P (2001) Ezrin is a downstream effector of trafficking pkc-integrin complexes involved in the control of cell motility. EMBO J 20:2723–2741
102. Novak I, Slepchenko B, Mogilner A, Loew L (2004) Cooperativity between cell contractility and adhesion. Phys Rev Lett 93:268109
103. Oliver J, King J, Mckinlay K, Brown P, Grant D, Scotchford C, Wood J (2005) Thin-film theories for two-phase reactive flow models of active cell motion. Math Med Biol 22:53–98
104. Oshiro N, Fukata Y, Kaibuchi K (1998) Phosphorylation of moesin by rho-associated kinase (rho-kinase) plays a crucial role in the formation of microvilli-like structures. J Biol Chem 273:34663–34666
105. Paluch E, van der Gucht J, Joanny JF, Sykes C (2006) Deformations in actin comets from rocketing beads. Biophys J 91(8):3113–3122
106. Papakonstanti E, Ridley A, Vanhaesebroeck B (2007) The p110d isoform of pi 3-kinase negatively controls rhoa and pten. EMBO J 26:3050–3061
107. Parekh SH, Chaudhuri O, Theriot JA, Fletcher DA (2005) Loading history determines the velocity of actin-network growth. Nat Cell Biol 7(12):1219–1223
108. Paskin C, Odell G, Oster G (1993) Cellular motions and thermal fluctuations: the Brownian ratchet. Biophs J 65:316–324
109. Pierres A, Benoliel A, Touchard D, Bongard P (2008) How cells tiptoe on adhesive surfaces before sticking. Biophys J 94:4114–4122
110. Pollard T (2003) The cytoskeleton, cellular motility and the reductionist agenda. Nature 422:741–745
111. Ponti A, Machacek M, Gupton S, Waterman-Storer C, Danuser G (2004) Two distinct actin networks drive the protrusion of migrating cells. Science 207:1782–1786
112. Prass M, Jacobson K, Mogilner A, Radmacher M (2006) Direct measurement of the lamellipodial protrusive force in a migrating cell. J Cell Biol 174:767–772
113. Rubinstein B, Jacobson K, Moginer A (2005) Multiscale two-dimensional modelling of a motile simple-shaped cell. Multiscale Model Simul 3(2):413–439
114. Sardet C, Counillon L, Franchi A, Pouyssegur J (1993) Growth factors induce phosphorylation of the na1/h1 antiporter, glycoprotein of 110 kd. Science 247:723–726
115. Schafer DA, Jennings PB, Cooper JA (1996) Dynamics of capping protein and actin assembly in vitro: uncapping barbed ends by polyphosphoinositides. J Cell Biol 135(1):169–179
116. Schreiber CH, Stewart M, Duke T (2010) Simulation of cell motility that reproduces the force–velocity relationship. Proc Natl Acad Sci 107(20):9141–9146
117. Sheetz M, Sable J, Dobereiner HG (2006) Continous membrane-cytoskeleton adhesion requires continous accomodation to lipid and cytoskeleton dynamics. Annu Rev Biophys Biomol Struct 35:417–434
118. Shiraishi M, Tanabe A, Saito N, Sasaki Y (2006) Unphosphorylated marcks is involved in neurite initiation induced by insulin-like growth factor-l in sh-sy5y cells. J Cell Physiol 209:1029–1038
119. Shlomovitz R, Gov N (2007) Membrane waves driven by actin and myosin. Phys Rev Lett 98:168103
120. Shlomovitz R, Gov N (2008) Exciting cytoskeleton-membrane waves. Phys Rev E 78:041911
121. Small J, Resch G (2005) The comings and goings of actin: coupling protrusion and retraction in cell motility. Curr Op Cell Biol 17:517–523
122. Small J, Herzog M, Anderson K (1995) Actin filament organization in the fish keratocyte lamellipodium. J Cell Biol 129(5):1275–1286
123. Small J, Auinger S, Nemethova M, Koestler S, Goldie K, Hoenger A, Resch G (2008) Unravelling the structure of the lamellipodium. J Microsc-Oxford 231:479–485
124. Small JV, Stradal T, Vignal E, Rottner K (2002) The lamellipodium: where motility begins. Trends Cell Biol 12(3):112–120

125. Smith J, Diez G, Klemm A, Schewkunow V, Goldmann W (2006) Capz-lipid membrane interactions: a computer analysis. Theor Biol Med Model 3(1):30
126. Stolarska M, Y K, Othmer H (2009) Multi-scale models of cell and tissue dynamics. Phyl Trans Roy Soc A 367:3525–3553
127. Stradal T, Scita G (2006) Protein complexes regulation arp2/3-mediated actin assembly. Curr Op Cell Biol 18:4–10
128. Stradal T, Rottner K, Disanza A, Confalonieri S, Innocenti M, Scita G (2004) Regulation of actin dynamics by wasp and wave family proteins. Trends Cell Biol 14:303–311
129. Sun HQ, Yamamoto M, Mejillano M, Yin HL (1999) Gelsolin, a multifunctional actin regulatory protein. J Biol Chem 274(47):33179–33182
130. Svitkina T (2007) N-wasp generates a buzz at membranes on the move. Cell 128:828–830
131. Svitkina TM, Verkhovsky AB, McQuade KM, Borisy GG (1997) Analysis of the actin-myosin II system in fish epidermal keratocytes: mechanism of cell body translocation. J Cell Biol 139:397–415
132. Tatsumi S, Mabuchi T, Katano T, Matsumura S, Abe T, Hidaka H, Suzuki M, Sasaki Y, Minami T, Ito S (2005) Involvement of rho-kinase in inflammatory and neuropathic pain through phosphorylation of myristoylated alanine-rich c-kinase substrate (marcks). Neurosci 131:491–498
133. Timpson P, Daly R (2005) Distinction at the leading edge of the cell. Bioessays 27:349–352
134. Trichet L, Campàs O, Sykes C (2007) Vasp governs actin dynamics by modulating filament anchoring. Biophys J 92:1081–1089
135. Trichet L, Campas O, Sykes C, Palastino J (2007) Vasp governs actin dynamics by modulating filament anchoring. Biophys J 92:1081–1089
136. Tsukita S, Hieda Y, Tsukita S (1989) A new 82-kd barbed end-capping protein (radixin) localized in the cell-to-cell adherens junction: purification and characterization. J Cell Biol 108:2369–2382
137. Urban W, Jacob S, Nemethova M, Resch G, Small J (2010) Electron tomography reveals unbranched networks of actin filaments in lamellipodia. Nat Cell Biol 12:429–435
138. Vallotton P, Small JV (2009) Shifting views on the leading role of the lamellipodium in cell migration: speckle tracking revisited. J Cell Sci 122(12):1955–1958
139. Verkhovsky A, Chaga O, Schaub S, Svitkina T, J-J M, Borisy G (2003) Orientational order of the lamellipodial actin network as demonstrated in living motile cells. Mol Biol Cell 14: 4667–4675
140. Watt S, Kular G, Fleming I, Downes C, Lucocq J (2002) Subcellular localozation of phosphytidylinositol 4,5-biphosphate using the pleckstrin homology domain of phospholipase cd1. Biochem J 363:657–666
141. Wedlich D (ed) (2004) Cell Migration in Development and Disease. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
142. Weichsel J, Schwarz US (2010) Two competing orientation patterns explain experimentally observed anomalies in growing actin networks. Proc Natl Acad Sci 107(14):6304–6309
143. Wolgemuth C (2005) Lamellipodial contractions during crawling and spreading. Biophys J 89:1643–1649
144. Yamaguchi H, Shiraishi M, Fukami K, Tanabe A, Ikeda-Matsuo Y, Naito Y, Sasaki Y (2009) MARCKS regulates lamellipodia formation induced by IGF-I via association with PIP2 and β-actin at membrane microdomains. J Cell Physiol 220:748–755
145. Yin HL, Stossel TP (1979) Control of cytoplasmic actin gel-sol transformation by gelsolin, a calcium-dependent regulatory protein. Nature 281:583–586
146. Yin HL, Albrecht JH, Fattoum A (1981) Identification of gelsolin, a Ca^{2+}-dependent regulatory protein of actin gel–sol transformation, and its intracellular distribution in a variety of cells and tissues. J Cell Biol 91(3):901–906
147. Zimmermann J, Enculescu M, Falcke M (2010) Leading edge - gel coupling in lamellipodium motion. Phys Rev E 82(5):051925

Chapter 21
Time-Structure of the Yeast Metabolism In vivo

Kalesh Sasidharan, Masaru Tomita, Miguel Aon, David Lloyd, and Douglas B. Murray

Abstract All previous studies on the yeast metabolome have yielded a plethora of information on the components, function and organisation of low molecular mass and macromolecular components involved in the cellular metabolic network. Here we emphasise that an understanding of the global dynamics of the metabolome in vivo requires elucidation of the temporal dynamics of metabolic processes on many time-scales. We illustrate this using the 40 min oscillation in respiratory activity displayed in auto-synchronous continuously grown cultures of *Saccharomyces cerevisiae*, where respiration cycles between a phase of increased respiration (oxidative phase) and decreased respiration (reductive phase). Thereby an ultradian clock, i.e. a timekeeping device that runs through many cycles during one day, is involved in the co-ordination of the vast majority of events and processes in yeast. Through continuous online measurements, we first show that mitochondrial and redox physiology are intertwined to produce the temporal landscape on which cellular events occur. Next we look at the higher order processes of DNA duplication and mitochondrial structure to reveal that both events are choreographed during the respiratory cycles. Furthermore, spectral analysis using the discrete Fourier transformation of high-resolution (10 Hz) time-series of NAD(P)H confirms the

K. Sasidharan • M. Tomita • D.B. Murray (✉)
Institute for Advanced Biosciences, Keio University, Nipponkoku 403–1, Daihouji, Tsuruoka City, Yamagata 997–0017, Japan
e-mail: cskalesh@sfc.keio.ac.jp; mt@sfc.keio.ac.jp; dougie@ttck.keio.ac.jp

M. Aon
Johns Hopkins University, School of Medicine, 720 Rutland Avenue, 1059 Ross Building, Baltimore, MD 21205, USA
e-mail: maon1@jhmi.edu

D. Lloyd (✉)
Microbiology, School of Biosciences, Cardiff University, Main Building, P.O. Box 915 Cardiff CF10 3AT, Wales, UK
e-mail: LloydD@cardiff.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_21,

existence of higher frequency components of biological origin and that these follow a scale-free architecture even in stable oscillating modes. A different signal-processing approach using discrete wavelet transformations (DWT) indicates that there is a significant contribution to the overall signal from ∼5, ∼10 and ∼20-minutes cycles and the amplitudes of these cycles are phase-dependent. Further investigation (derivative of Gaussian continuous wavelet transformation) reveals that the observed 20-minutes cycles are actually confined to the reductive phase and consist of two ∼15-minutes cycles. Moreover, the 5 and 10-minutes cycles are restricted to the oxidative phase of the cycle. The mitochondrial origin of these signals was confirmed by pulse-injection of the cytochrome c oxidase inhibitor H_2S. We next discuss how these multi-oscillatory states can impinge on the apparently complex reactome (represented as a phase diagram of 1,650 chemical species that show oscillatory behaviour). We conclude that biological processes can be considerably more comprehensible when dynamic *in vivo* time-structure is taken into account.

Abbreviation List

CWT	Continuous wavelet transformation
DFT	Discrete Fourier transformation
DOG	Derivative of Gaussian
DWT	Discrete wavelet transformation
FACS	Fluorescence activated cell sorting
FAD	Flavin adenine dinucleotide
FMN	Flavin mononucleotide
rfu	Relative fluorescence units

1 Introduction

The metabolic network is a highly dynamic system from millisecond interactions at the cell membranes, seconds of metabolic reactions, minutes involved in transcription/translation to cell-division cycles in the hour domain. However, rather few studies have addressed how these domains interlink to create the "temporal architecture *in vivo*" [1–4]. This is chiefly because detailed comprehension requires time-resolutions not easily achieved for any one type of analysis. For example, high-throughput analyses are costly and are mostly applicable to populations of cells, whereas single cells analyses can only be conducted a set number of times before the monitoring system perturbs and/or irrevocably damages the individual [5]. In single cells, despite the astonishing advances in fast imaging techniques, and the growing availability of specific fluorescent probes, limitations still exist (e.g., for the quantification of molecular participants during the rapid observation

of their kinetic interactions and for elucidation of their long-term fates). Use of synchronous cultures, from which information on individual cells or organisms can be inferred from populations, often suffers from the effect of perturbations given by the experimental procedures employed to obtain the initial synchrony [6].

Early workers seeking to describe the cell-division cycles or organisms or cells as ordered sequences of events or processes achieved considerable success for the major steps (e.g., DNA synthesis (S-phase), nuclear division, organelle dynamics), but description of metabolic changes was limited [7–12]. Observations of the oscillatory behaviour in enzyme activities [13], concentrations of metabolites and key co-enzymes (NAD(P)H and adenine nucleotides) necessitated the postulation of the operation of a short period (τ) biological clock (for various species of yeast $\tau = 30$ min to 8 h) [14, 15]. Auto-synchrony during continuous aerobic cultures of brewer's yeast (*Saccharomyces cerevisiae*) can be maintained in an auto-dynamic oscillatory state over extended times (up to months). This has been known for over 40 years [16–20]; however, it is only recently that with developments in high-throughput and continuous online monitoring (coupled to rapid acquisition systems) the full extent of the oscillatory system has been revealed [3, 21–26]. The burgeoning power of mass-spectrometric analysis of metabolites [25, 27], high-throughput dissection of the cell-division cycle events [18, 21], transcriptomics [21, 26], continuous in-line monitoring [23, 28] and computational modelling [29] have now made possible a detailed study of the intricate operation of the metabolic network of synchronised yeast on multiple time-scales. It should be stressed that the extensive evolutionary conservation of central metabolic reactions and pathways imply that our description of the yeast metabolome is an archetype and is, therefore, of general applicability to eukaryotic microbes, animals and plants [30].

In this chapter, we first illustrate the global redox cycle underpinning auto-synchronous respiratory oscillations using online measurements and conclude that nucleo-mitochondrial activity is critical. We next utilise flow cytometry to further explore the relationship between the redox cycle, mitochondrial mass and cell-division cycle progression. We then use high frequency sampled NAD(P)H data (analysed using Fourier, discrete and continuous wavelets) to illustrate that the underlying architecture of the metabolome is multi-oscillatory, i.e., have a statistically self-similar multiple frequency output. Inhibiting mitochondrial respiration at the level of cytochrome oxidase with H_2S abates all oscillatory frequencies including the 40 min period ultradian clock. Therefore, providing proof-of-principle that multi-scale timekeeping is an emergent property of the overall network involved in metabolism, growth, and proliferation in yeast, as the overall impact of the perturbation, was shown over all temporal scales. This illustrates that the architecture of the cellular network, conventionally obtained from infrequent sampling, is merely a snapshot of a highly dynamic system. These data are then discussed with respect to amino acid regulation and cellular energetics. We next reconstruct the available data to produce a polar plot or "clock-face" of phenotypic and molecular events occurring during each redox cycle.

2 Methods

2.1 Strain and Culture Conditions

Unless otherwise stated, all chemicals were supplied by Wako Chemicals, Japan. The *Saccharomyces cerevisiae* strain used in this study was the diploid strain IFO 0233. The medium consisted of glucose monohydrate (20 g dm^{-3}), $(NH_4)_2SO_4$ (5 g dm^{-3}), KH_2PO_4 (2 g dm^{-3}), $MgSO_4 \cdot 7H_2O$ (0.5 g dm^{-3}), $CaCl_2 \cdot 2H_2O$ (0.1 g dm^{-3}), $FeSO_4 \cdot 7H_2O$ (0.02 g dm^{-3}), $ZnSO_4 \cdot 7H_2O$ (0.01 g dm^{-3}), $CuSO_4 \cdot 5H_2O$ (0.005 g dm^{-3}), $MnCl_2 \cdot 4H_2O$ (0.001 g dm^{-3}), yeast extract (Difco; 1 g dm^{-3}) and 70% (v/v) H_2SO_4 (1 cm^3 dm^{-3}). Sigma-Aldrich antifoam 204 was used at 0.2 cm^3 dm^{-3}. Continuous culture was carried out as described previously using an Eyela bioreactor (MBF-250ME, Eyela, Japan; working volume 0.65 dm^{-3}). The fermentor was stirred at 750 rpm utilising high-precision stirrers (Eurostar Power Control Visc, IKA, Japan) and aerated at 0.15 dm^3 min^{-1}, the aeration rate being controlled by a mass-flow controller (GFC, Aalborg, Japan). The feed pump (Perista, AC-2120, Atto, Japan) was automatically calibrated to deliver 55 cm^3 h^{-1} by monitoring the weight drop on a high-capacity balance (PM160001, Mettler Toledo, Japan). Pulses were minimised by using small bore tubing (ID 1 mm, Pharmed, Cole-Parmer, Japan) and by the planetary design of the pump. The pH was controlled at 3.4 by monitoring with an immersed electrode (InPro3030, Mettler Toledo, Japan) and controlled addition of 2.5 M NaOH solution by a Labo controller (B.E. Marubishi, Japan). An independent circulating water-bath (FE-25, Julabo, Japan) maintained the culture at 30°C using an external thermometer. The dissolved oxygen concentration was monitored using an immersed polarographic electrode (InPro6800, Mettler Toledo, Japan).

2.2 Electron Microscopy and Flow Cytometry

Samples were prepared and visualised by electron microscopy as previously described [31]. Culture (1 mL) was removed from the reactor, pelleted at $12k \times g$(10 s) and aspirated. The staining solution (1 mL) containing 0.25 μg of Hoechst 33342 (Wako Chemicals, Japan), 0.1 μg 2-[3-[5,6-dihydro-2*H*-benzimidazol-2-ylidene]-1-propen-1-yl]-3-methyl-benzoxazoliumchloride (Mitotracker Green FM, Molecular probes, Japan) and phosphate buffered saline at 4°C was used to re-suspend the pellet. Samples were then analysed on a FACScalibur flow cytometer. The time between sampling and analysis was ∼15 min. Data from the analysis were visualised and analysed using the Bioconductor package FlowClust [32]. The resulting cluster centre means and distributions were calculated for three independent measurements and are represented in relative fluorescence units (rfu).

2.3 Monitoring and Calculations

All instruments were supervised and their outputs monitored and logged using custom designed software.

Heat production rate Q in the culture was estimated using Fourier's law of heat conduction (21.1):

$$Q = -2k\pi rl\frac{\mathrm{d}T}{x} \tag{21.1}$$

where $\mathrm{d}T$, temperature difference, was calculated by subtracting the fermentor temperature from the water-bath temperature, r is the internal radius of the reactor, l is the internal height of the reactor's liquid phase, and k is the conductivity of glass and x was the thickness of the glass (2 mm).

Continuous partial pressure of oxygen $\left(P_{\mathrm{O}_{2(\mathrm{o})}}\right)$ and partial pressure of carbon dioxide $\left(P_{\mathrm{CO}_{2(\mathrm{o})}}\right)$ off-gas measurements were carried out using an Enoki-III (Figaro engineering, Japan) analyser. The partial pressure of hydrogen sulphide $\left(P_{\mathrm{H_2S}_{(\mathrm{o})}}\right)$ in the off-gas was measured continuously using an electrode based gas monitor (HSC-1050HL, GASTEC, Japan). Instruments were calibrated as per manufacturer's instruction. The input air partial pressures of oxygen $\left(P_{\mathrm{O}_{2(\mathrm{i})}}\right)$ was 0.20947 atm, carbon dioxide $\left(P_{\mathrm{CO}_{2(\mathrm{i})}}\right)$ was 0.0004 atm and hydrogen sulphide $\left(P_{\mathrm{H_2S}_{(\mathrm{o})}}\right)$ was 0 atm. O_2 uptake rates (q_{O_2}), CO_2 production rates (q_{CO_2}) and H_2S production rates $(q_{\mathrm{H_2S}})$ were derived from the following equations:

$$\begin{aligned}
I_{\mathrm{i}} &= 1 - \left(P_{\mathrm{O}_{2(\mathrm{i})}} + P_{\mathrm{CO}_{2(\mathrm{i})}} + P_{\mathrm{H_2S}_{(\mathrm{i})}}\right) = 0.79013 \\
I_{\mathrm{o}} &= 1 - \left(P_{\mathrm{O}_{2(\mathrm{o})}} + P_{\mathrm{CO}_{2(\mathrm{o})}} + P_{\mathrm{H_2S}_{(\mathrm{o})}}\right) \\
q_{\mathrm{O}_2} &= \frac{F}{RTV}\left(P_{\mathrm{O}_{2(\mathrm{i})}} - P_{\mathrm{O}_{2(\mathrm{o})}}\frac{I_{\mathrm{i}}}{I_{\mathrm{o}}}\right) \\
q_{\mathrm{CO}_2} &= \frac{F}{RTV}\left(P_{\mathrm{CO}_{2(\mathrm{o})}}\frac{I_{\mathrm{i}}}{I_{\mathrm{o}}} - P_{\mathrm{CO}_{2(\mathrm{i})}}\right) \\
q_{\mathrm{H_2S}} &= \frac{F}{RTV}\left(P_{\mathrm{H_2S}_{(\mathrm{o})}}\frac{I_{\mathrm{i}}}{I_{\mathrm{o}}} - P_{\mathrm{H_2S}_{(\mathrm{i})}}\right)
\end{aligned} \tag{21.2}$$

where F was the gas flow into the system, R was the universal gas constant (0.0820575 L atm mol^{-1} K^{-1}) and V was the volume of the reactor (0.65 dm^{-3}).

NAD(P)H and oxidised flavins were measured by a fluorimeter. Briefly a Hg source (Nikon, Japan) was coupled to a fibre guide through a 360 nm ± 10 nm filter for NAD(P)H excitation or 460 nm ± 10 nm filter for flavin (Optoscience, Japan). The fibre optic guide end was chemically sterilised with ethanol and the tip was immersed in the culture. The fibre guide was split so that half the fibres (15 fibres) was used for excitation and the other half was used for emission detection. Detection was carried out using a photomultiplier tube with 460 nm ± 10 nm filter

for NAD(P)H or 530 nm ± 10 nm filter for flavins. The scatter fluorescence and lamp fluctuations were measured by monitoring the excitation wavelength using filtered photomultiplier tube or photocell, respectively, and the excitation signal was compensated for any variation in these signals [33]. The system was supplied by Norman Graham (University of Pennsylvania, USA) [34]. Data were acquired from the PMT amplifier using a fast USB data acquisition module (PMD-1608FS, Measurement computing, Japan) at a rate of 10 Hz. Custom designed software (VirtualChartRecorder) was then used to visualise and log the data stream.

2.4 Data and Signal Processing

All data and signal processing were carried out in R using custom scripts available on request. Wavelet analyses were carried out using Rwave R package [35] and MassSpecWavelet Bioconductor package [36]. Models were constructed in CellDesigner [37, 38].

The data from numerous studies [3, 4, 21, 23, 31, 39–49] were compiled together with their reference dissolved oxygen concentrations. Where the raw data were not available, the plots were first digitised (Plot Digitiser, J.A. Huwaldt, http://plotdigitizer.sourceforge.net/). The phase (θ) of each sample (k) was calculated for each cycle (m) for each dataset (21.3).

$$\theta_{\mathrm{k}} = 360^{\circ}\left(\frac{t_{\mathrm{k}} - t^{m}_{(\mathrm{d}[\mathrm{O}_2]/\mathrm{d}t)_{\min}}}{t^{m+1}_{(\mathrm{d}[\mathrm{O}_2]/\mathrm{d}t)_{\min}} - t^{m}_{(\mathrm{d}[\mathrm{O}_2]/\mathrm{d}t)_{\min}}}\right) \tag{21.3}$$

where t was the sample time and the start point for each cycle was the minimum first derivative ($\mathrm{d}[O_2]/\mathrm{d}t$) of the dissolved oxygen concentrations. Samples were then phase adjusted to reconstruct three cycles where θ_1 was closest to 0°. All time series were then analysed using statistical methods to obtain a phase angle, a signal-to-noise ratio and a significance of the signal-to-noise ratio. The metabolites utilise the nomenclature developed for a large-scale model of yeast metabolism. All gene names were retrieved from the *Saccharomyces* genome database.

3 Results and Discussion

3.1 Real-Time Measurements of Redox State

Continuous online measurements (Fig. 21.1) illustrate the phase relationships between seven parameters measured. Dissolved O_2 was chosen as a benchmark of oscillatory activity because it was most conveniently measured by a relatively cheap immersed electrode, and it represents the residual dissolved O_2 concentration

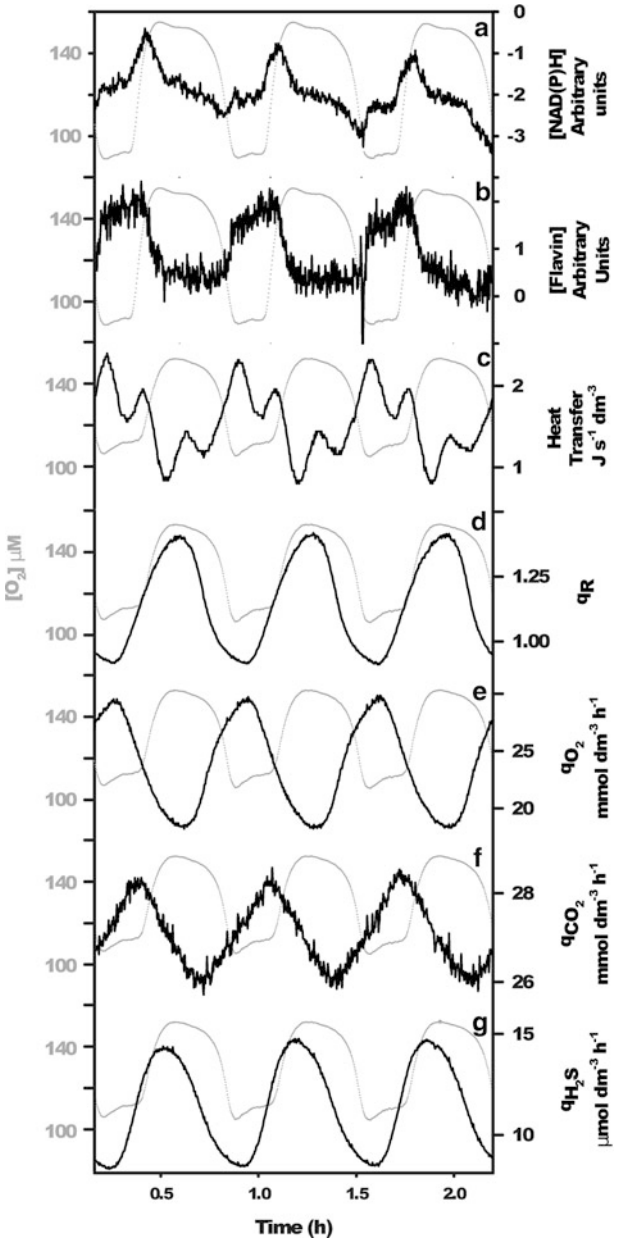


Fig. 21.1 The phase relationships between the continuously monitored parameters and the dissolved oxygen concentration (all figures; O_2) during the respiratory oscillation found during yeast continuous culture. NAD(P)H (**a**) and flavin (**b**) fluorescence were measured continuously on-line using a fluorimeter. Heat transfer (**c**) was calculated using Fourier's law from the water bath temperature and the reactor temperature. Respiratory quotient (q_R; d) was calculated by dividing the carbon dioxide production rates $\left(q_{CO_2};\ f\right)$; by the oxygen uptake rates $\left(q_{O_2};\ e\right)$; these were calculated from partial pressure measurements. Hydrogen sulphide production rates $\left(q_{H_2S};\ g\right)$ were calculated from the partial pressure of H_2S measured using an electrode

remaining after the organisms used what they require. The phase angles of all variables were referenced to the minimum first derivative of this oscillatory output. The troughs correspond to high respiratory activity or oxidative phase, where oxygen uptake rates are at a maximum, and the peaks correspond to low respiratory activity or reductive phase, where oxygen uptake is at a minimum. Its waveform has a period circa 40 min. Nicotinamide nucleotide (NAD(P)H) redox state was maximally reduced just before dissolved O_2 reached maximal values (Fig. 21.1a), at the onset of the reductive stage. Interestingly the minimum values for NAD(P)H coincided with the minimum first derivative of the dissolved oxygen concentration. Flavin fluorescence (Fig. 21.1b) peaked prior to NAD(P)H and was in-phase with oxygen uptake rates (Fig. 21.1e) and anti-phase to respiratory activity (Fig. 21.1d). The flavin signal (Fig. 21.1b) represents the sum of the fluorescent signal from free and bound flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) both excite at 450 nm and emit at 535 nm in the oxidised form [50]. FAD is the co-enzyme predominantly involved in transferring electrons from the TCA cycle reactions to the electron transport chain. FMN in both free and protein bound forms show auto-fluorescence and are key cofactors in many oxidoreductive enzymes including those involved in cellular iron regulation and mitochondrial assembly [27]. The data presented here clearly show that flavins are predominantly oxidised during the oxidative phase. Flavin fluorescence emission peaks at the end of the oxidative phase then rapidly declines at the transition to the reductive phase. Both NAD(P)H (Fig. 21.1a) and flavin signals (Fig. 21.1b) show more complex waveforms than that of dissolved oxygen. Additionally, heat transfer from the reactor showed a complex waveform with three local maxima occurring for each respiratory cycle (Fig. 21.1c), with major maximum and minimum rates corresponding to those for oxygen uptake rates (Fig. 21.1e). The higher frequency observed correlated with the three peaks of transcriptional activity previously observed [21, 24]. Maximum CO_2 production occurred at the end of the oxidative phase (Fig. 21.1f). Maximum H_2S production occurs in-phase with maximum dissolved oxygen (Fig. 21.1g). Therefore, online measurements provide a unique insight into the phase relationships between the products of respiration and the redox state of the cell. This is perhaps best illustrated by the increased flavin oxidation state and oxygen uptake rates thereby indicating that mitochondrial function is globally being influenced during each cycle.

3.2 The Relationship Between Redox State, Mitochondrial Function and the Cell-Division Cycle

Previously, we visualised the state of mitochondria during the oxidative and reductive phase by using ultra-thin sections of fixed organisms and electron microscopy [41]. These studies indicated that mitochondria structure drastically altered during each cycle (Fig. 21.2). Energised mitochondria in the oxidative stage were characterised by cristae that are not easily visible due to an expanded intermembrane space. In the reductive phase, mitochondria became de-energised, more

electron dense and the cristae were folded into sheets with most of the intra-mitochondrial space occupied by the matrix. However, the relationship between the cell-division cycle and the mitochondrial cycle as well as the details of the mitochondrial cycle remains to be elucidated. These changes are also mirrored in the cytoplasmic structure where the osmium/Pb fixation procedure yields organisms that are more electron dense during respiratory activity.

The cell-division cycle is gated by the respiratory oscillation where S-phase predominantly occurs during the reductive phase of the 40 min period cycle [21]. This occurs in a para-synchronous way so that ∼10% of the total cell number progress through S-phase (the cell-division time of the culture was set by the dilution rate of the culture to 8.1 h). To address the relationship between the oscillation, mitochondrial mass and the cell-division cycle we conducted two-dye flow cytometry using the vital dyes Hoechst 33342 for DNA and Mitotracker Green FM for mitochondrial mass (Fig. 21.2). Staining was conducted on live cells, making high-resolution time sampling impossible, therefore, we selected two time-points (the cytometry is a representative sample of triplicate plots). Small cells (forward scatter results not shown) with a low DNA content were assigned to be G_0 cells, these small cells also had a very low mitochondrial mass which was comparable in the oxidative (O) and reductive (R) phases. As cells progressed through the cell-division cycle the mitochondrial mass increased linearly for both time-points sampled. However, this increase was twice as much in cells sampled during the oxidative phase. So the average mitochondrial mass of G_1 cells in the oxidative phase was 225 rfu ± 52 rfu and 175 rfu ± 24 rfu in the reductive; in G_2 cells, the mitochondrial mass during the oxidative phase was 325 rfu ± 64 rfu and 248 rfu ± 38 rfu during the reductive phase. A model based clustering approach was used to define the cell cycle phases and the deviation of the cluster centres [32]. A distinct S-phase cluster was not detected during the oxidative phase. Therefore, the only cells that did show synchronous behaviour with respect to their mitochondrial mass were G_0 cells. S-Phase gating of 10% of the cells involves a reduction in mitochondrial mass of the entire population and G_1 appears to be the most affected phase of the cell-division cycle as mitochondrial mass showed significant differences in distributions compared to G_2 between the oxidative and reductive phases. We, therefore, provide an insight into how a short-time scale, i.e., that of mitochondrial energisation interact with the long-time scales of the cell-division cycle.

3.3 The Multi-oscillatory States of the Yeast "Redox Core"

NAD(P)H redox state is one of the most important indicators of intracellular function and also of oxidative stress [27]. In yeast, nitrosation of thiols and Fe cluster by NO^+ has a striking perturbative effect on respiratory oscillations [23]. Data on oxidative stress-mediated depletion of glutathione and cysteine levels confirm the close association of these core reactions with sulphate assimilation and H_2S

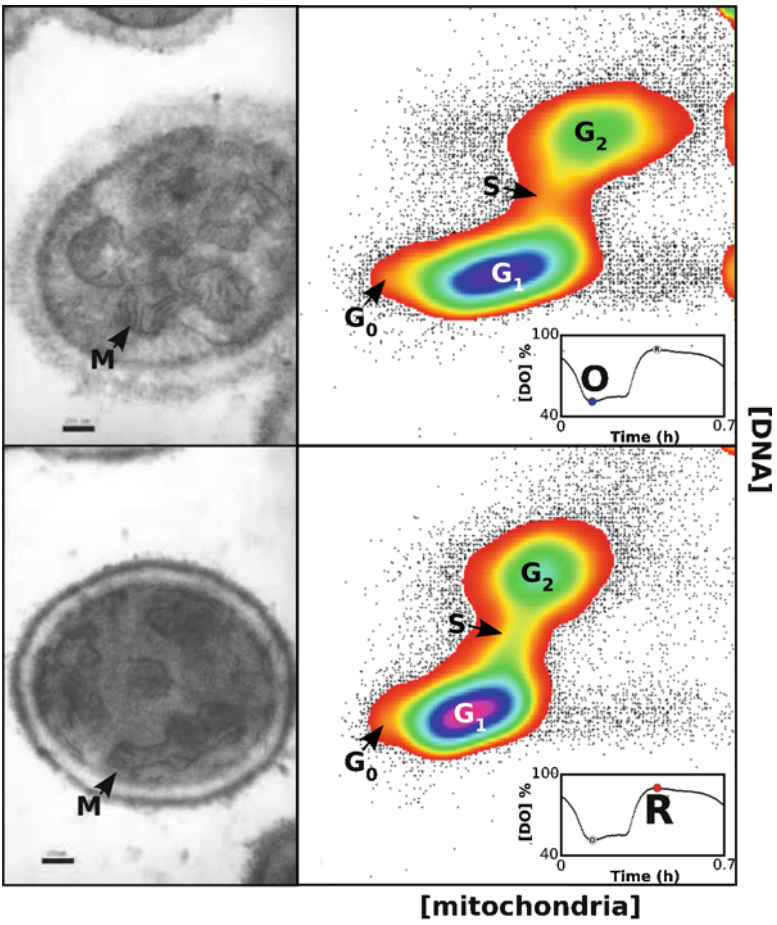


Fig. 21.2 The temporal organisation of mitochondria and the cell-division cycle, revealed by electron micrographs of thin sections of organisms sampled in the oxidative (O) and reductive (R) phases of the respiratory cycle (*left-hand* pictures; M – mitochondria), and flow-cytometry of DNA concentration, measured by Hoechst 33448 and mitochondrial mass [mitochondria], measured by Mitotracker Green™ (*right-hand* heatmaps). The pictures are electron micrographs taken from thin sections of the fixed samples. The *black bar* represents 200 μm. The heatmaps were produced from 50,000 cells analysed by flow cytometry. The maximum peak was 528 cells (*blue–purple* hue or *black*). Phases of the cell-division cycle are represented by G_0, G_1, S and G_2. The panel *insets* show a cycle of the respiratory oscillation found during continuous yeast growth where each *black dot* represents a concentration of dissolved oxygen [DO]. The *open circles* represent the oscillation phase where the samples were obtained

production [44]. H_2S, an evanescent, toxic and highly diffusible messenger in cell physiology [51–54], is also periodically produced in yeast with a fast rise time and a slower decline, from sulphite *via* sulphite reductase [42, 43, 48]. Regulation of this process has been studied using yeast disrupted in glutathione synthesis

and glutathione reductase, and with thiol redox modifying agents (diethylmaleate, *N*-ethylmaleimide, DL-buthionine[*S*, *R*]-sulphoximine or 5-nitro-2-furaldehyde). These compounds all caused perturbation of the respiratory oscillation [31], as did certain genetic modifications of the wild-type strain [47, 55]. However, hydrogen sulphide only phase shifts the oscillation during the reductive phase indicating that this potent inhibitor of the cytochrome c oxidase is acting as an amplitude modulator of the oscillation. Acetaldehyde too is a volatile and highly-diffusible messenger molecule and is capable of phase-shifting the respiratory oscillations during all phases making it an important synchronisation molecule [27, 46, 48, 49]. Acetaldehyde directly modulates NAD(P)H redox balance, and is a direct product of ethanol oxidation. Therefore, balance of these reactions is critical for phase modulation as revealed by the phase-resetting response curves obtained by pulsed addition of acetaldehyde to the spontaneously synchronous culture [4]. Recently, we analysed the data from continuous cultures exhibiting chaotic oscillatory behaviour in CO_2 and O_2 (10 s resolution) and in NAD(P)H fluorescence (10 ms resolution) for individual cells [1]. Both experiments indicated that temporal organisation in cells followed the rules of a scale-free self-similar system or statistical fractal over at least three orders of magnitude in time. These events encompass the cell-division cycle time-scale down to the millisecond time-scale of metabolic events. The data acquisition system of the fluorimeter employed to measure NAD(P)H can sample at 10 Hz; therefore, we carried out a similar analysis using NAD(P)H data sampled from continuous cultures (420,000 data points; Fig. 21.3a). The oscillation was allowed to free-run for 5 cycles, then we inhibited respiratory activity using a pulse-injection of 10 μmol ammonium sulphide. This resulted in the instantaneous release of H_2S which reversibly inhibits the mitochondrial respiratory chain [42]. This perturbation led to an initial rapid rise in NAD(P)H due to inhibition of respiration and NADH accumulation from the TCA cycle. The resulting redox imbalance was followed by the oxidation of the NADH pool as the TCA cycle undergoes product inhibition, and the antioxidant defences exhaust the electron donor capability of NADH. Importantly, recovery to the stable oscillatory state occurs only after 10 h stressing the overall impact across the metabolic and signalling network. Initially, the signal was processed using discrete Fourier analysis (DFT) where the signal could easily be delineated from the noise (where amplitude is proportional to frequency; Fig. 21.3b). The signal component of the spectra followed the scale-free behaviour as in our previous analysis during the unperturbed experiment [2]. The perturbation produced the major amplitude in the spectra (marked as 1 and 2 of Fig. 21.3b); the 40-minutes oscillation was the next most significant amplitude (marked as 3 of Fig. 21.3b). Possible sub-harmonics of the 40-minutes oscillation (20 and 10 min) were also observed (marked as 4 and 5 of Fig. 21.3b). Just prior to the appearance of noise in the signal, there was also significant components observed in the 2–5 min domain (marked as 6 and 7 of Fig. 21.3b). In order to minimise computational resources, we re-sampled the data to 1 Hz (42,000 data points) for further analyses.

Whilst DFT is very efficient at indicating whether a frequency occurred and for discriminating noise from signal in a particular time series, it does not indicate the

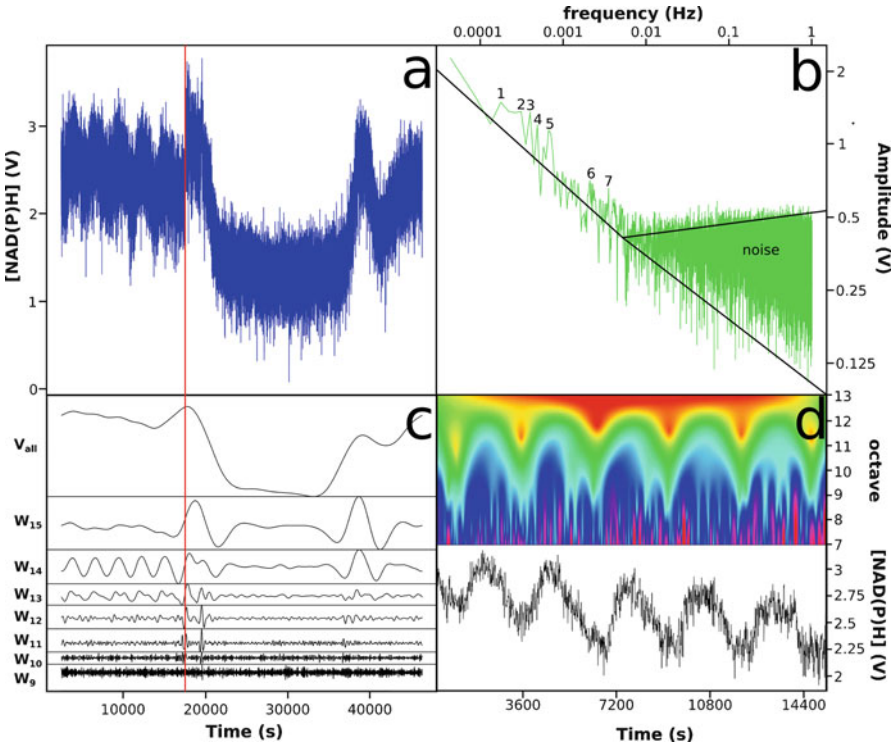


Fig. 21.3 Signal processing of the complex signal produced from continuous online measurement of NAD(P)H initially sampled at 10 Hz (**a**). Discrete Fourier transformation (DFT) spectra reveal that the relation between the amplitude was linear until 0.05 Hz indicating scale-free dynamics in this region (**b**); below this, we observed a region of coloured noise. Discrete wavelet transformation (DWT) using the Daubechies wavelet was then used to process the signal where windows (W) that had significant correlation were shown (**c**). The data were down-sampled to 1 Hz to reduce computation cost. Continuous wavelet transformation (CWT) using the derivative of Gaussian wavelet (DOG) of data down-sampled to 0.1 Hz reveals the finer grain temporal events of the signal (**d**). The heatmap intensity indicates the correlation of the signal to the wavelet. The *vertical line* in (**a**) and (**c**) represents the time of $(NH_3)_2S$ addition

temporal behaviour of this frequency [35]. For example, if a complex waveform has frequency or amplitude modulation, e.g., if a signal becomes damped when perturbed, then DFT will not differentiate this from a stable waveform. Wavelet analysis while computationally more expensive than DFT provides not only a correlation coefficient to the chosen wavelet of a particular component, but an indication of the behaviour of that frequency. Therefore, DFT cannot be used to determine if the perturbation causes changes to the signals we observe. Therefore, we analysed the re-sampled data using Daubechies discrete wavelet transform (DWT, Fig. 21.3c). Again, we find the major correlation in all the data-points was from the perturbation

(V_{all} and W_{14} on Fig. 21.3c) and the next major component was from the 40-minutes oscillation (W_{14} on Fig. 21.3c) where this signal was completely damped during the perturbation. Interestingly, the correlation at 20 min (W_{13} on Fig. 21.3c) was greater than half the correlation of the 40-minutes signal indicating that this correlation was not merely a sub-harmonic of the 40-minutes signal, i.e., there was an observable 20-minutes cycle in the reductive phase where two peaks are merged. During the initial stages of the perturbation, this signal separates then dampens. There is a residual 20-minutes oscillation during the perturbation but this may be artificial due to effects from the DWT analysis. The 10-minutes correlation (W_{12} on Fig. 21.3c) is also much larger than that expected from a sub-harmonic ($>1/3$ of the 40-minutes signal). Additionally, its behaviour is also different from the 20-minutes signal, in that the amplitude of the signal is larger during the oxidative phase indicating that the NAD(P)H oscillates faster in the oxidative phase when mitochondria are more active. On perturbation, the cycles also become more separated and increase in amplitude before becoming damped later on in the perturbation. The 5-minutes signal (W_{11} on Fig. 21.3c) follows a similar pattern to the 10-minutes signal. We also found significant correlation in the 1-minute signal (W_9 on Fig. 21.3c); however, this did not display any significant change during the perturbation. All other frequencies tested did not produce significant correlations.

Whilst DWT is a powerful signal processing tool, it is limited by only extracting sub-multiples of the original input dataset. We, therefore, utilised the continuous wavelet transformation (CWT) to enhance the fidelity of the picture of the time-structure [56]. The CWT produces wavelet coefficients that represent a collection of correlations to a chosen waveform in a range that represents the time-scales in the data (called octaves from the historical use of wavelets to analyse sound; set to 13) and octaves are then subdivided into a number of arbitrarily chosen correlations called voices to represent the sub-scale resolution required (set to 16). The wavelet of choice is also arbitrary and we selected the derivative of Gaussian (DOG) wavelet, as this produced the best temporal resolution for higher frequencies for our data. The resulting matrix of correlation coefficients was visualised in a heat-map (Fig. 21.3d). To minimise computational cost, we re-sampled the data to 0.1 Hz and focussed on the stable oscillatory region (5 cycles). DOG-CWT reveals that the 5 and 10-minutes component (octaves 8 and 9) can be successfully delineated and were of much higher amplitude in the transition between the reductive to oxidative phase and during the oxidative phase. It also showed that the 20-minutes component observed in the DWT was almost exclusively found in the reductive phase (octave 10), and its periodicity was closer to 12.5 min. The analysis indicates that NAD(P)H undergoes a switch in its multi-oscillatory behaviour from low-frequency NAD(P)H oscillations during the reductive phase to higher frequencies in the oxidative phase which we suggest is mediated by mitochondrial activity. Indeed the mitochondria-generated frequency oscillations have been directly imaged in time-series of images obtained by two-photon excitation of NAD(P)H auto-fluorescence (scan-rate 8 Hz) [2].

3.4 The Temporal Behaviour of the Cellular Network

The impact of the observed multi-oscillatory behaviour on previous work using the self-synchronised aerobic continuous culture of *Saccharomyces cerevisiae* is yet to be fully appreciated and mitochondrial activity (especially the transmembrane electrochemical potential of the inner membrane) and redox state of the nicotinamide nucleotide-coenzymes all are critical determinants of global metabolic state. On many of the measurements, including those of dissolved oxygen and transcript levels, we frequently observe shoulder peaks that correspond with the 12.5-minutes oscillation observed by CWT [3,21]. Interestingly, on many of these measurements, we also observed sample–sample variation, apparently confined to the oxidative phase. As these were sampled at every 4 min, this variation could in retrospect largely be due to the higher frequencies now observable during the oxidative phase. This leads to the striking conclusion that stable-multi-oscillatory dynamics gives a "noise-free" system in the metabolic and transcriptional time domains and indeed may be their major function. By sampling at successively higher time resolutions, we can repeatedly find another higher frequency responsible for apparently noisy behaviour.

To illustrate how these multiple time-scales can be generated, we examined the amino acid biosynthesis network, where carbon skeletons are derived from central carbon metabolism, chiefly TCA cycle intermediates and to a lesser extent glucogenesis/glycolytic intermediates. Thereby, we can generate a simple conceptual model (Fig. 21.4). The intracellular concentrations of glutamate, aspartate, threonine, valine, leucine, serine and cysteine all oscillate with ∼40-minutes period, but showed different oscillation amplitudes, apparent noise, dual peaks and maxima [3, 43, 45]. This implies that numerous regulatory circuits are utilised for amino acid biosynthesis during the respiratory oscillation. The cyclic functional changes of mitochondria result in the oscillation in the concentration of NAD(P)H which are directly linked with the rate of amino acid biosynthesis. When amino acids biosynthesis rates are high, increases in the intracellular amino acid pools lead to increased aminoacylation of tRNAs [57]. The aminoacylated tRNAs then inhibit the translation of Gcn4p, the socalled "master regulator" of amino acid biosynthesis [58, 59]. This in turn results in the suppression of amino acid biosynthesis genes, leading to a change in production of the enzymes which catalyse amino acid biosynthesis. This forms an auto-regulatory system where amino acids form a negative feedback loop on a transcriptional feed forward loop; such systems have long been known for their ability to generate sustained oscillations [13].

There are also numerous modifications to the basic system where ever increasing levels of complexity exist. The GATA amino acid regulation system of nitrogen catabolite repression provides a slower time-scale *via* regulatory feedback loops on the promoters of many of the genes encoding amino acid enzymes [60]. Additionally, the activities of many amino acid enzymes are rapidly modulated by allosteric inhibition. A particularly complex picture of regulation is emerging for the regulation of serine where increase in the concentration of serine causes suppression

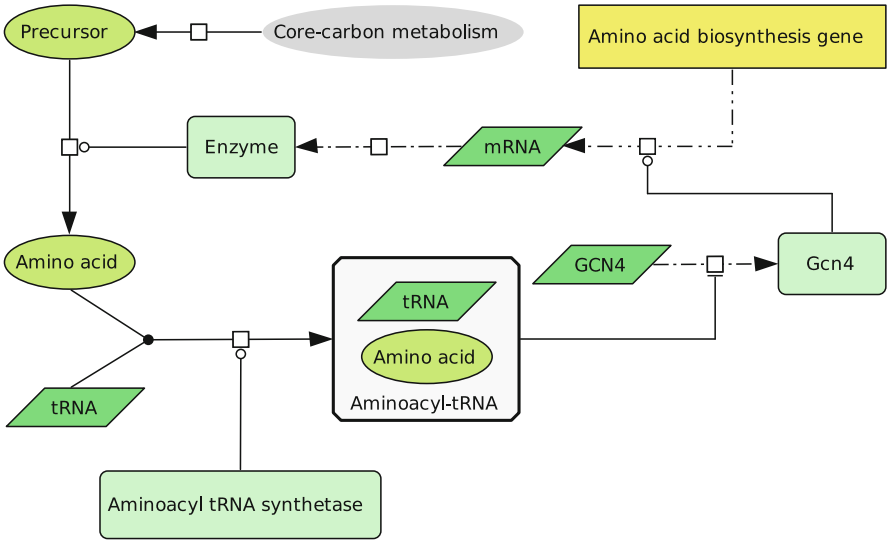


Fig. 21.4 Simplified CellDesigner [37] representation of *GCN4* regulation of amino acid biosynthesis during respiratory oscillation [3]. The transition from the precursor to the amino acid is tightly regulated by mitochondria and tRNA-mediated *GCN4* pathway. All symbols follow system biology graphical notation standards [38]

of the transcription of *SER3* (encoding phosphoglycerate dehydrogenase) [61] *via* the transcriptional regulator Cha4p (slow response) [62]. The Cha4p is primed on the promoter site and once serine concentrations increase, a rapid transcriptional response is elicited that is mediated by *SRG1* (non-coding RNA) and a rearrangement of nucleosome architecture at the promoter [63]. Serine also is an allosteric inhibitor for Ser3p (a fast response) [64]. The metabolic and transcriptional feedbacks and transcriptional feed-forward loops involved in *SER3* transcription and the activity of phosphoglycerate dehydrogenase have a high potential to oscillate or attenuate the oscillatory frequency and amplitude. The rapid feedback caused by allosteric inhibition, for example, may serve to dampen rapid changes in serine. However, when concentrations of the product are low this will have little effect on the fluctuation of the product.

The amino acid biosynthesis pathways and their regulation represent only a small part of the reactome; taken together, this indicates that there exist a depth of both intracellular and extracellular communication that at first seems unfathomably complicated [29]. However, we suggest that temporal coherence provided by a multi-oscillatory system simplifies and partitions the cellular processes. In order to summarise this, we constructed a polar plot of all the measured parameters involved in metabolic control (Fig. 21.5) and related this to some of the observed phenotypes during a cycle. Transcription is known to occur in at least three bursts; one in the oxidative phase and two in the reductive phase [21, 26, 65]. This correlates with heat production and the multi-oscillatory signal of NAD(P)H found in these studies.

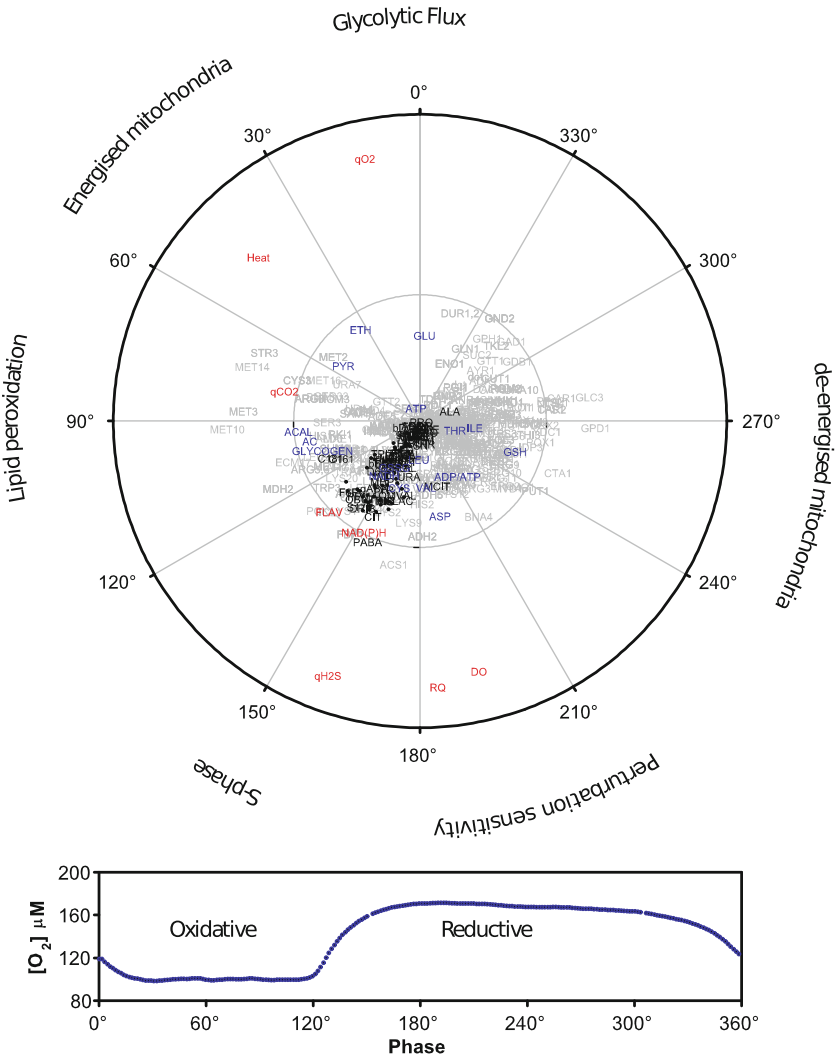


Fig. 21.5 The phase reconstruction of the signal-to-noise ratio (S/N) of measured parameters found during the respiratory oscillation in continuously growing yeast. The polar plot was constructed from the fast Fourier transform analyses on numerous datasets [3, 4, 21, 23, 31, 39–49]. The *red text* represents the online parameters where the signal-to-noise ratio had to be divided by 10 because sampling was much more frequent. The *blue text* represents metabolites measured in low-throughput enzymatic or HPLC methods. The *black text* shows data from high-throughput GC–MS measurements. The *grey text* represents transcripts involved in metabolism measured using Affymetrix microarrays. Physiological markers are indicated in text centred on the peak of the phenotype. For example, perturbation sensitivity refers to the sensitivity of this region to redox altering compounds such as ROS and glutathione, as well as other chemical agents. The lower plot represents a phase normalised cycle of dissolved oxygen and is meant to guide the reader. The transcript names are common names found in the yeast genome database and the metabolite names are from a model of yeast metabolism

During the oxidative phase, a biosynthetic program is initiated that ends with a peak in concentration of many of the measured cellular metabolites and NAD(P)H [3]. This coincides to the peak of DNA synthesis [21] and hydrogen sulphide production. From the polar plot, many transcripts involved in catalysis show peak production in the reductive phase, for example, Enolase 1 transcript (*ENO1*), oxidising the glycolytic intermediate 2-phosphoglycerate to phosphoenolpyruvate [66], reaches a maximum concentration in the late reductive phase; about 8–10 min prior to maximum glycolytic flux. It is important to note that Fig. 21.5 was produced exclusively from cells grown on glucose media. The oscillation still occurs when cells are grown on ethanol [31, 49] as the sole carbon source and, therefore, glycolytic activity is not required for the respiratory oscillation. Periodic accumulation of storage carbohydrates (glycogen and trehalose) does not occur during growth on ethanol media.

4 Outlook and Conclusions

Here we show a global overview of yeast respiratory organisation based on online measurements, i.e., by using high- and low-throughput data "snapshots". This is by no means complete and we are utilising and developing computational, proteomic, transcriptomic, metabolomic and single-cell analysis technologies to improve our understanding of cellular organisation in time. Indeed, using optical tweezers, we managed to monitor synchrony occurring between two individuals [2]. Recently, yeast cells were sorted based on glutathione content and cadmium resistance by fluorescence activated cell sorting (FACS), to reveal that the cell produces the appropriate rhythmic output by integrating the environmental signals *via* a "redox core" [55]. The "redox core" encompasses the cycling of intracellular redox biochemistry, i.e., thiol, flavin and nicotinamide adenine dinucleotide pools. Moreover, inhibiting mitochondrial respiration at the level of cytochrome c oxidase with H_2S to a synchronous continuous yeast culture abates all oscillatory frequencies including the 40-minutes period ultradian clock.

The systems described in this chapter are some of the most highly conserved during the long evolution of eukaryotes, and therefore it would seem likely that similar systems operate in all higher eukaryotes. Emerging evidence from circadian biology confirms that the majority of gene expression in mammals is oscillatory [67–69] and fundamental redox changes underpin these oscillatory behaviours [70, 71]; it would seem likely that temporal architecture in these systems is also multi-oscillatory.

Acknowledgements We thank Rainer Machné for helpful discussions. DL and DBM are grateful to the Royal Society and the Japan Society for the Promotion of Science for supporting this work. KS, DBM and MT are supported in part by funds from Yamagata Prefectural Government and Tsuruoka-city. MT and DBM are also supported by a Japan partnering award (Japan Science and Technology agency and the Biotechnology and Biological Sciences Research Council, UK).

References

1. Aon MA et al (2008) The scale free network organization of yeast and heart systems biology. PLoS One 3:e3624
2. Aon MA, Cortassa S, Lemar KM, Hayes AJ, Lloyd D (2007) Single and cell population respiratory oscillations in yeast: a 2-photon scanning laser microscopy study. FEBS Lett 581: 8–14
3. Murray DB, Beckmann M, Kitano H (2007) Regulation of yeast oscillatory dynamics. Proc Natl Acad Sci USA 104:2241–2246
4. Murray DB, Lloyd D (2007) A tuneable attractor underlies yeast respiratory dynamics. Bio Systems 90:287–294
5. Jacquet M, Renault G, Lallet S, De Mey J, Goldbeter A (2003) Oscillatory nucleocytoplasmic shuttling of the general stress response transcriptional activators Msn2 and Msn4 in *Saccharomyces cerevisiae*. J Cell Biol 161:497–505
6. Shedden K, Cooper S (2002) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. Nucleic Acids Res 30:2920–2929
7. Arreguin de Lorencez M, Käppeli O (1987) Regulation of gluconeogenic enzymes during the cell cycle of *Saccharomyces cerevisiae* growing in a chemostat. J Gen Microbiol 133: 2517–2522
8. Wiemken A, Matile P, Moor H (1970) Vacuolar dynamics in synchronously budding yeast. Arch Mikrobiol 70:89–103
9. Dawson PS, Westlake DW (1975) Changes in pattern of respiration and glucose utilisation in *Candida* utilis during the cell cycle: some variations with growth rate. Can J Microbiol 21:1013–1019
10. Creanor J (1978) Oxygen uptake during the cell cycle of the fission yeast *Schizosaccharomyces pombe*. J Cell Sci 33:399–411
11. Creanor J (1978) Carbon dioxide evolution during the cell cycle of the fission yeast *Schizosaccharomyces pombe*. J Cell Sci 33:385–397
12. Lloyd D, Poole RK, Edwards SW (1982) The cell division cycle: temporal organization control of cellular growth and reproduction. Academic, London
13. Goodwin BC (1965) Oscillatory behavior in enzymatic control processes. Adv Enzyme Regul 3:425–437
14. Lloyd D, Rossi EL (2008) Ultradian rhythms from molecules to mind: a new vision of life. Springer, New York
15. Lloyd D, Rossi EL (1992) Ultradian rhythms in life processes: an inquiry into fundamental principles of chronobiology and psychobiology. Springer-Verlag Berlin and Heidelberg GmbH & Co. K
16. von Meyenburg HK (1968) The budding cycle of *Saccharomyces cerevisiae*. Pathol Microbiol 31:117–127
17. von Meyenburg HK (1969) Energetics of the budding cycle of *Saccharomyces cerevisiae* during glucose limited aerobic growth. Arch Microbiol 66:289–303
18. Münch T, Sonnleitner B, Fiechter A (1992) [Repeat] The decisive role of the *Saccharomyces cerevisiae* cell cycle behaviour for dynamic growth characterization. J Biotechnol 22: 329–351
19. Sonnleitner B (1991) Dynamics of yeast metabolism and regulation. Bioprocess Eng 6: 187–193
20. Satroutdinov AD, Kuriyama H, Kobayashi H (1992) Oscillatory metabolism of *Saccharomyces cerevisiae* in continuous culture. FEMS Microbiol Lett 77:261–267
21. Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. Proc Natl Acad Sci USA 101:1200–1205
22. Lloyd D, Murray DB (2006) The temporal architecture of eukaryotic growth. FEBS Lett 580:2830–2835

23. Murray DB, Engelen FA, Keulers M, Kuriyama H, Lloyd D (1998) NO^+, but not NO^-, inhibits respiratory oscillations in ethanol-grown chemostat cultures of *Saccharomyces cerevisiae*. FEBS Lett 431:297–299
24. Li CM, Klevecz RR (2006) A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change. Proc Natl Acad Sci USA 103:16254–16259
25. Tu BP et al (2007) Cyclic changes in metabolic state during the life of a yeast cell. Proc Natl Acad Sci USA 104:16886–16891
26. Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science 310:1152–1158
27. Murray DB, Haynes K, Tomita M (2011) Redox regulation in respiring *Saccharomyces cerevisiae*. Biochim Biophys Acta http://www.ncbi.nlm.nih.gov/pubmed/21549177
28. Locher G, Sonnleitner B, Fiechter A (1992) On-line measurement in biotechnology: exploitation, objectives and benefits. J Biotechnol 25:55–73
29. Herrgård MJ et al (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26:1155–1160
30. Lloyd D, Murray DB (2005) Ultradian metronome: timekeeper for orchestration of cellular coherence. Trends Biochem Sci 30:373–377
31. Murray DB, Engelen F, Lloyd D, Kuriyama H (1999) Involvement of glutathione in the regulation of respiratory oscillation during a continuous culture of *Saccharomyces cerevisiae*. Microbiol 145(Pt 10):2739–2745
32. Lo K, Hahne F, Brinkman RR, Gottardo R (2009) flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinform 10:145
33. Chance B, Cohen P, Jobsis F, Schoener B (1962) Intracellular oxidation-reduction states in vivo. Science 137:499–508
34. Chance B, Thorell B (1959) Fluorescence measurements of mitochondrial pyridine nucleotide in aerobiosis and anaerobiosis. Nature 184:931–934
35. Carmona R, Hwang W-L, Torresani B (1998) Practical time-frequency analysis, volume 9: gabor and wavelet transforms, with an implementation in s (wavelet analysis and its applications). Academic Press, San Diego
36. Du P, Kibbe WA, Lin SM (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics 22:2059–2065
37. Funahashi A, Tanimura N, Morohashi M, Kitano H (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BioSilico 1:159–162
38. Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23:961–966
39. Murray DB, Roller S, Kuriyama H, Lloyd D (2001) Clock control of ultradian respiratory oscillation found during yeast continuous culture. J Bacteriol 183:7253–7259
40. Sohn H, Kuriyama H (2001) Ultradian metabolic oscillation of *Saccharomyces cerevisiae* during aerobic continuous culture: hydrogen sulphide, a population synchronizer, is produced by sulphite reductase. Yeast 18:125–135
41. Lloyd D, Salgado LE, Turner MP, Suller MT, Murray DB (2002) Cycles of mitochondrial energization driven by the ultradian clock in a continuous culture of *Saccharomyces cerevisiae*. Microbiology 148:3715–3724
42. Sohn HY, Murray DB, Kuriyama H (2000) Ultradian oscillation of *Saccharomyces cerevisiae* during aerobic continuous culture: hydrogen sulphide mediates population synchrony. Yeast 16:1185–1190
43. Sohn H, Kuriyama H (2001) The role of amino acids in the regulation of hydrogen sulfide production during ultradian respiratory oscillation of *Saccharomyces cerevisiae*. Arch Microbiol 176:69–78
44. Kwak WJ, Kwon GS, Jin I, Kuriyama H, Sohn HY (2003) Involvement of oxidative stress in the regulation of H_2S production during ultradian metabolic oscillation of *Saccharomyces cerevisiae*. FEMS Microbiol Lett 219:99–104

45. Sohn H-Y, Kum E-J, Kwon G-S, Jin I, Kuriyama H (2005) Regulation of branched-chain, and sulfur-containing amino acid metabolism by glutathione during ultradian metabolic oscillation of *Saccharomyces cerevisiae*. J Microbiol 43:375–380
46. Keulers M, Satroutdinov AD, Suzuki T, Kuriyama H (1996) Synchronization affector of autonomous short-period-sustained oscillation of *Saccharomyces cerevisiae*. Yeast 12: 673–682
47. Sohn H-Y et al. (2005) *GLR1* plays an essential role in the homeodynamics of glutathione and the regulation of H_2S production during respiratory oscillation of *Saccharomyces cerevisiae*. Biosci Biotechnol Biochem 69:2450–2454
48. Murray DB, Klevecz RR, Lloyd D (2003) Generation and maintenance of synchrony in *Saccharomyces cerevisiae* continuous culture. Exp Cell Res 287:10–15
49. Keulers M, Suzuki T, Satroutdinov AD, Kuriyama H (1996) Autonomous metabolic oscillation in continuous culture of *Saccharomyces cerevisiae* grown on ethanol. FEMS Microbiol Lett 142:253–258
50. Aubin JE (1979) Autofluorescence of viable cultured mammalian cells. J Histochem Cytochem 27:36–43
51. Elsey DJ, Fowkes RC, Baxter GF (2010) Regulation of cardiovascular cell function by hydrogen sulfide (H(2)S). Cell Biochem Funct 28:95–106
52. Gadalla MM, Snyder SH (2010) Hydrogen sulfide as a gasotransmitter. J Neurochem 113: 14–26
53. Wang M-J, Cai W-J, Zhu Y-C (2010) Mechanisms of angiogenesis: role of hydrogen sulphide. Clin Exp Pharmacol Physiol 37:764–771
54. Lloyd D (2006) Hydrogen sulfide: clandestine microbial messenger? Trends Microbiol 14:456–462
55. Smith MCA, Sumner ER, Avery SV (2007) Glutathione and Gts1p drive beneficial variability in the cadmium resistances of individual yeast cells. Mol Microbiol 66:699–712
56. Walker JS (2008) Beyond wavelets. In: Walker, JS (ed.) A primer on wavelets and their scientific applications. Chapman and Hall, London, pp 223–254
57. Wek RC, Jackson BM, Hinnebusch AG (1989) Juxtaposition of domains homologous to protein kinases and histidyl-tRNA synthetases in GCN2 protein suggests a mechanism for coupling GCN4 expression to amino acid availability. Proc Natl Acad Sci USA 86:4579–4583
58. Dever TE et al (1992) Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. Cell 68:585–596
59. Natarajan K et al (2001) Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol Cell Biol 21:4347–4368
60. Boczko EM et al (2005) Structure theorems and the dynamics of nitrogen catabolite repression in yeast. Proc Natl Acad Sci USA 102:5647–5652
61. Albers E, Laizé V, Blomberg A, Hohmann S, Gustafsson L (2003) Ser3p (Yer081wp) and Ser33p (Yil074cp) are phosphoglycerate dehydrogenases in *Saccharomyces cerevisiae*. J Biol Chem 278:10264–10272
62. Martens JA, Wu P-YJ, Winston F (2005) Regulation of an intergenic transcript controls adjacent gene transcription in *Saccharomyces cerevisiae*. Genes Dev 19:2695–26704
63. Thebault P et al (2011) Transcription regulation by the noncoding RNA *SRG1* requires Spt2-dependent chromatin deposition in the wake of RNA polymerase II. Mol Cell Biol 31: 1288–1300
64. Dubrow R, Pizer LI (1977) Transient kinetic studies on the allosteric transition of phosphoglycerate dehydrogenase. J Biol Chem 252:1527–1538
65. Murray DB (2006) The respiratory oscillation in yeast phase definitions and periodicity. Nat Rev Mol Cell Biol 7:696–701
66. Cohen R, Holland JP, Yokoi T, Holland MJ (1986) Identification of a regulatory region that mediates glucose-dependent induction of the *Saccharomyces cerevisiae* enolase gene *ENO2*. Mol Cell Biol 6:2287–2297
67. Etchegaray JP, Lee C, Wade PA, Reppert SM (2003) Rhythmic histone acetylation underlies transcription in the mammalian circadian clock. Nature 421:177–182

68. Ptitsyn AA et al. (2006) Circadian clocks are resounding in peripheral tissues. PLoS Comput Biol 2:e16
69. Yamada R, Ueda HR (2007) Microarrays: statistical methods for circadian rhythms. Meth Mol Biol 362:245–264
70. Lloyd D, Murray DB (2000) Redox cycling of intracellular thiols: state variables for ultradian, cell division cycle and circadian cycles? In: Driessche TV, Guisset JL, Vries GMP-de (eds.) The redox state and circadian rhythms. Kluwer, Dordrecht, pp 85–94
71. O'Neill JS et al (2011) Circadian rhythms persist without transcription in a eukaryote. Nature 469:554–558

Chapter 22
Coarse Graining *Escherichia coli* Chemotaxis: From Multi-flagella Propulsion to Logarithmic Sensing

Tine Curk, Franziska Matthäus, Yifat Brill-Karniely, and Jure Dobnikar

Abstract Various sensing mechanisms in nature can be described by the Weber–Fechner law stating that the response to varying stimuli is proportional to their relative rather than absolute changes. The chemotaxis of bacteria *Escherichia coli* is an example where such logarithmic sensing enables sensitivity over large range of concentrations. It has recently been experimentally demonstrated that under certain conditions *E. coli* indeed respond to relative gradients of ligands. We use numerical simulations of bacteria in food gradients to investigate the limits of validity of the logarithmic behavior. We model the chemotactic signaling pathway reactions, couple them to a multi-flagella model for propelling and take the effects of rotational diffusion into account to accurately reproduce the experimental observations of single cell swimming. Using this simulation scheme we analyze the type of response of bacteria subject to exponential ligand profiles and identify the regimes of absolute gradient sensing, relative gradient sensing, and a rotational diffusion dominated regime. We explore dependance of the swimming speed,

T. Curk
Department of Chemistry, University of Cambridge, Cambridge, UK

Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
e-mail: tcurk5@gmail.com

F. Matthäus
Center for Modeling and Simulation in the Biosciences (BIOMS),
University of Heidelberg, Heidelberg, Germany
e-mail: franziska.matthaeus@iwr.uni-heidelberg.de

Y. Brill-Karniely
Department of Chemistry, University of Cambridge, Cambridge, UK
e-mail: yb233@cam.ac.uk

J. Dobnikar (✉)
Department of Chemistry, University of Cambridge, Cambridge, UK

Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia
e-mail: jd489@cam.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_22, © Springer Science+Business Media, LLC 2012

average run time and the clockwise (CW) bias on ligand variation and derive a small set of relations that define a coarse grained model for bacterial chemotaxis. Simulations based on this coarse grained model compare well with microfluidic experiments on *E. coli* diffusion in linear and exponential gradients of aspartate.

1 Introduction

The bacterium *Escherichia coli* propels itself by rotating a set of long filaments, the flagella. Depending on whether the rotation is clockwise (CW) or counter-clockwise (CCW), the flagella, which have a chiral helical structure, take on different conformations. As a result the bacterium either swims along a (more or less) straight trajectory, or it engages in a tumble in which it randomly changes the direction. The motion behavior, swim and tumble, is controlled by a network of enzymatic reactions called the signaling pathway [1, 2]. Ligands (chemoattractants or chemorepellents) bind to a receptor complex in the bacterial membrane and, through the signaling cascade, influence the bias of flagellar rotation. The signaling pathway provides a certain "memory" of previously encountered ligand concentrations, and allows a comparison to present concentrations. In this way *E. coli* is able to bias its motion toward or against chemical gradients. This chemotactic behavior of the bacteria exhibits complex properties. One example is a very high sensitivity that enables *E. coli* to respond to very small as well as to very high ligand concentrations. The sensitivity is the result of a strong signal amplification by the signaling pathway [3] and ultrasensitivity of the flagellar motor on the response control CheYp [4]. Also receptor clustering has been found to enhance sensitivity [5]. Another example of complex behavior is a very precise adaptation to constant ligand concentrations [6, 7]. In constant ligand concentrations, the flagellar bias returns to the bias of unstimulated bacteria with very high accuracy. In this way, *E. coli* bacteria are able to respond to changes in the ligand concentration, irrespective of the absolute concentration level. Through the adaptation process, the bacteria sense relative gradients: they react to very small changes if the absolute concentration is small, but large changes are necessary to influence behavior in large absolute concentrations. The response to a relative gradient can be measured via the average run time or run length $\langle l \rangle$. If the bacteria sense relative changes the average run length relates to changes in the gradient as $\langle l \rangle \sim \mathrm{d}L/L = \mathrm{d}(\log(L))$, where L denotes the ligand concentration. This so-called *logarithmic sensing* is also a feature of visual or auditory perception, and has been described in the Weber–Fechner law. It has been shown only recently that also the chemotactic sensing of *E. coli* bacteria follows the Weber–Fechner law [8]. Although the logarithmic sensing had been anticipated, only the development of microfluidic chambers that generate stable chemical gradients, made an experimental verification possible.

Here we develop a coarse grained mathematical model for *E. coli* chemotaxis. We extend a previous model for propelling [9] by accounting for the combined action of four-flagella, which are regulated through the same signaling pathway. This extension of the model will eliminate a prevalent inconsistency in the modeling of

E. coli chemotaxis, namely, that the tumbling frequency of free swimming bacteria ($p \approx 0.1$) does not correspond to the CW bias of a single flagella ($\tau \approx 0.5$) [10]. We then use numerical simulations to determine several macroscopic relations and dependencies that constitute the coarse grained model for *E. coli* chemotaxis. Performing coarse grained computer simulations we are able to reproduce the bacterial density profiles in microfluidic experiments with various chemoattractand landscapes.

The following sections are organized as follows: In Sect. 2 we introduce our multi-flagella model to describe swimming of a single bacterium and derive the macroscopic relations that constitute the coarse grained model. In Sect. 3 we use the four-flagella model to simulate bacterial motion in exponential gradients. With this approach we identify different regions of chemotactic sensing: gradient sensing, relative gradient (logarithmic) sensing, and rotational diffusion dominated response. Finally, in Sect. 4 we simulate bacterial behavior in linear and exponential gradients based on the coarse grained model and we reproduce the experimental density distributions of bacteria in defined ligand gradients.

2 Modeling the Interaction of Four-Flagella

In our previous work we have adapted a model of the chemotaxis signaling pathway of *E. coli* to study the influence of noise on bacterial motion in various chemical landscapes [9]. This model included a detailed description of the regulatory processes: the signal transduction through steps of phosphorylation, from the receptor to the motor control CheY, and receptor methylation regulating adaptation to absolute ligand concentrations. The model was given as a set of differential equations adapted from [2], complemented by an algebraic relation to couple the concentration level of CheYp to the tumbling frequency. Another approach to study *E. coli* chemotaxis is to use stochastic simulation of the interaction of regulatory enzymes [11]. Bray et al. have developed a model that describes not only the signaling pathway, but also the assembly of the receptor complex (see for instance [12]). All these approaches model *E. coli* chemotaxis as a process, where the internal signaling cascade directly controls the tumbling frequency through the motor switching curve defining the CW bias. In other words, the propelling is described as if it was governed by the action of a single effective flagellum.

In reality, the process is much more complicated, as the tumbling frequency, as well as the swimming speed, are both affected by the interplay of several (typically four) flagella. The molecular motors that drive the flagellar rotation switch independently from each other, and not all of them have to turn CCW to impose swimming. On the other hand, a tumble can occur upon reversal of one or several flagella. Experimental studies by Turner et al. [13] have provided an insight into flagellar forms during different stages of propelling: depending on the CW or CCW spinning of the motor a flagellum spins in a so-called right handed or left handed form, respectively. Actually, there exist different right handed forms, but

Table 22.1 Shape (k) and scale (θ) parameters of the Γ-distribution (density function $f(x;k,\theta) = x^{(k-1)}\frac{\exp(-x/\theta)}{\theta^k \cdot \Gamma(k)}$) for different values of the CW bias following Korobkova et al. [14]. Intermediate values are approximated using a stepwise linear function

	CCW		CW	
τ	k_{CCW}	$\theta_{\mathrm{CCW}}^{-1}$	k_{CW}	$\theta_{\mathrm{CW}}^{-1}$
0.1	2	0.8	5	16.5
0.2	3	2.3	5	15
0.3	3	3.5	5	13.3
0.4	4	5.9	5	11
0.5	5	9.3	5	9.3
0.6	5	12	4	6.5
0.7	5	14	3	3.5
0.8	5	17	3	2.5
0.9	5	21	2	1

our modeling approach will not include this level of detail. When all the motors spin CCW, the flagella form a bundle and *E. coli* performs a run. Upon reversal of one motor (or more), the bundle is disturbed. The respective flagellum escapes the bundle and changes its waveform to right handed. This causes a change in the direction of swimming and a reduction of the speed. The more filaments break out of the bundle, the greater is the effect. After reversal back to CCW rotation, the flagellum rejoins the bundle and the initial speed is attained. It has been found that *E. coli* with four-flagella in total will change the direction by approximately 30° when one motor switches from CCW to CW. The right handed waveform can also contribute to propulsion, but is less efficient than left handed form. Thus, in the unlikely case that the bundle is formed out of CW spinning flagella, the speed of bacteria would be significantly lower than in the case of a left handed bundle.

To model the interplay of four-flagella and the resulting motion behavior we use experimental data of Korobkova et al. [14] derived for a single flagellar motor. They found that for an individual motor the duration of CW and CCW events follows a Γ-distribution for a given CW bias. In homogeneous environments, bacteria have a CW bias around 0.2, and the CW and CCW intervals have an average duration of 0.3 s and 1.3 s, respectively. In our model the dependence of the CW bias on the level of CheYp is described by a Hill function as given in [4]:

$$\tau = \frac{{Y_{\mathrm{p}}}^{H_{\mathrm{c}}}}{{Y_{\mathrm{p}}}^{H_{\mathrm{c}}} + K_{\mathrm{c}}^{H_{\mathrm{c}}}} \tag{22.1}$$

with Hill coefficient $H_{\mathrm{c}} = 10.3$ and dissociation constant $K_{\mathrm{c}} = 3.1\ \mu$M. Depending on the CW bias, the duration of CW and CCW intervals for every individual motor is determined by a Γ-distribution, following Korobkova et al. [14] (see Table 22.1). Whenever at least one of the motors switches from CCW to CW rotation the direction of bacteria is changed according to a Γ-distribution with shape 5 and scale 0.14, which provides a good fit to the experimental data of Turner et al. [13].

The speed of the bacterium v is determined based on the forms of all flagella at any given moment. Each form is designated by a form number which is a measure of

how much the flagellum in the particular form contributes to the motion. We assign a value $f = 1$ to all left handed forms, which are the most efficient. The right handed forms can have two different form values. Relying on the experimental observations of Turner et al. [13] we choose $f = 0$ for the first 80 ms after the switch, where the bundle is disturbed and the direction is changing, and $f = 0.5$ for CW spinning after 80 ms. This relates to the finding that the flagellum can contribute to the speed even if it is not part of the bundle. Summarizing,

$$f = \begin{cases} 1 \text{ left handed form,} \\ 0 \text{ right handed form less than 80 ms after the switch,} \\ \frac{1}{2} \text{ right handed form more than 80 ms after the switch.} \end{cases} \tag{22.2}$$

The combined action of four-flagella can be described by a single parameter $\Phi_4 = \sum_{i=1}^{4} f_i \,/\, 4$, where f_i denotes the form of the ith flagellum. Now we define the speed of the bacterium, subject to a respective combination of different flagellar forms as:

$$v = v_{\max} \cdot \Phi_4^{2.5}, \tag{22.3}$$

where $v_{\max} = 30\,\mu$m/s is the maximum speed when all flagella contribute to the bundle, i.e., $\Phi_4 = 1$. The exponent of 2.5 is chosen such that the average speed in the simulations is compatible with the experimentally determined value of 20 μm/s [15, 16]. It has to be stressed that the choice of the particular functional form in (22.3) and the choice of the numerical values of f in different states in (22.2) are somewhat arbitrary. The constraints on the function $v/v_{\max}(\Phi_4)$ are that it must be a monotonically increasing function on the interval $[0, 1]$ with a value of 1 at $\Phi_4 = 1$ and a positive value smaller than 1 at $\Phi_4 = 0$, and that the average speed in a simulation using this relation is equal to 20 μm/s. We have constructed several analytical expressions meeting the above constraints; however, the final result is always very similar and the particular choice of the function does not seem to have any noticeable effect on the results of the modeling. It should be noted that the arbitrary parameters in the model are only involved with calculating the speed. The run–tumble transitions do not depend on the speed, therefore there are no ambiguities connected with the distribution of the run and tumble events. Furthermore, the speed variation with time in a homogeneous environment predicted by our model (Fig. 22.1c) qualitatively agrees with the dynamics observed in the experiments by Berg et al. [17].

In our model we also account for the effects of the rotational diffusion, which is a consequence of the thermal interaction of the bacteria with the liquid medium. Rotational diffusion causes the bacteria to deviate from a straight trajectory when the cells try to swim straight for a long period of time. We model the rotational diffusion as a change in the angle during a run – as a random walk with step $0.52\,\text{rad}\cdot\sqrt{\Delta t}$ [17].

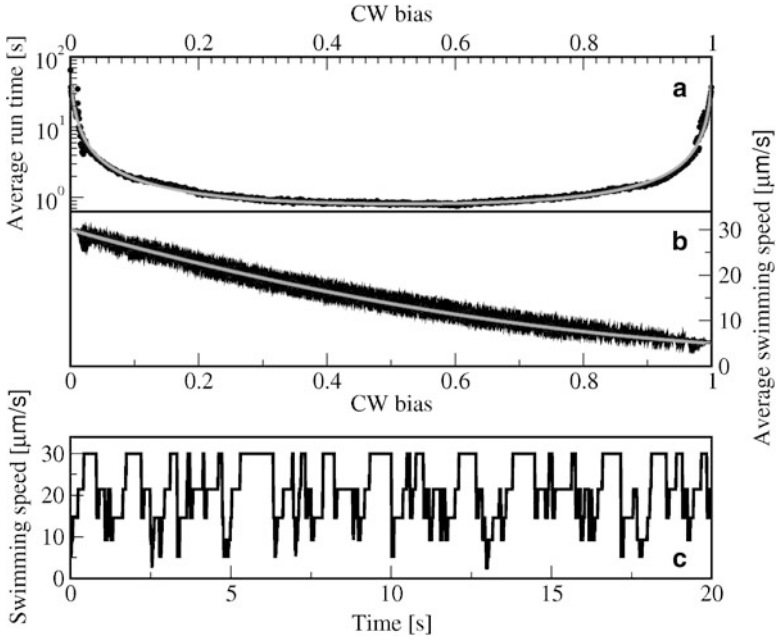


Fig. 22.1 (**a**) and (**b**): Average run length and speed as a function of CW bias as a result of the four-flagella model. The black curves are numerically determined from our model and the gray curves are the fitting functions (22.4) and (22.5). (**c**) Time variation of the speed for a single bacterium in homogenous environment (CW bias $\tau \approx 0.2$). The figure looks similar to the experimental data of Brown and Berg [17], and provides additional justification for the model

2.1 Model Verification

In order to compare simulations based on our model to the original tracking experiments of Berg et al. [17], we need to "measure" tumbling and swimming in the way defined in their original work. The beginning of a run is identified, when the bacterium changes its direction by less then 35° for three consecutive steps of 80 ms duration. On the other hand, if during a period of 80 ms the direction changes by more then 35°, the run ends and the cell is considered to tumble (for exact procedure description see Berg et al. [17]).

It needs to be noted that in this model a tumbling event is defined based on the rate of change of the direction of swimming and is not directly connected to the change of the flagellar states. Of course, the conformational switch of flagella indirectly contributes to the probability for tumbling. However, even in the case when the flagella do not switch at all, rotational diffusion alone can, with very low probability, cause a change in the angle that would be identified as a tumble.

To verify our model, we performed numerical simulations considering $N = 5000$ bacteria with fixed CW bias. At each time step we determined the form for

each flagellum on every cell according to the Γ-distribution mentioned above. For every cell we determined the speed according to (22.3) and the change of the direction (either through a switch in the flagellar conformation or due to the rotational diffusion). Our simulation results agree very well with the tracking experiments of Berg [17]. The run and tumble intervals are exponentially distributed with mean 1 s and 0.13 s, respectively. Also, the average change in direction during a tumble is about 60°. In Fig. 22.1c we display the dynamics of the current speed $v(t)$ of a single bacterium as a function of time. By visual inspection our simulation qualitatively agrees with experimentally observed dynamics ([17]).

We repeat the simulations for various values of the CW bias τ and measure the average run time $\langle l \rangle$ defined by the time elapsed between two consecutive tumbling events, and the average speed $\langle v \rangle$. The resulting dependencies are displayed in Fig. 22.1a and b, together with the appropriate fitting expressions given by:

$$\frac{\langle l \rangle(\tau)}{s} = 0.1 + \frac{0.3 - 0.03\tau - 0.5(\tau - 0.5)^2}{\left[(\tau + 3 \cdot 10^{-4})\,(1 + 3 \cdot 10^{-3} - \tau)\right]^{\frac{2}{3}(1+2(\tau-0.5)^2)}}\,, \tag{22.4}$$

$$\frac{v(\tau)}{\mu\mathrm{m/s}} = 30 - 40\tau + 15\tau^2\,. \tag{22.5}$$

The fit, (22.4), describing the average run time dependence on the CW bias was, obtained in an iterative process, where we started with a simpler fitting function $\langle l \rangle = A + \frac{B+C\tau}{(\tau(1-\tau))^D}$, with constants A, B, C, and D. We then improved the fit step by step by adding higher order terms. Even though (22.4) looks rather complicated we believe that it is necessary to have an accurate fitting expression, and we could not find any simpler way to express it in terms of analytic functions. For the dependence of the average speed on the CW bias (22.5), the rather simple second order polynomial provided a very good fit to the simulation data.

3 Logarithmic Sensing

Since our model seems to reproduce experimentally observed motion behavior well, we will now use it to study the response of large bacterial populations to time-varying chemical stimuli. *E. coli* bacteria use chemotaxis to move toward or against *spatial* gradients. Their way of detecting a spatial gradient is, however, to compare concentration values while moving. Effectively they thus respond to concentration changes in time. Studies on how time-varying stimuli affect the motion of *E. coli* date back to the seminal experiments of Brown and Berg [18]. Using their tracking microscope and an enzymatic reaction to either build up or degrade L-glutamate, they showed for the first time that the average run length increases when the L-glutamate concentration increases with time. More recently, Tu, Shimizu and Berg [19–21], developed a simple coarse grained model that describes the relationship

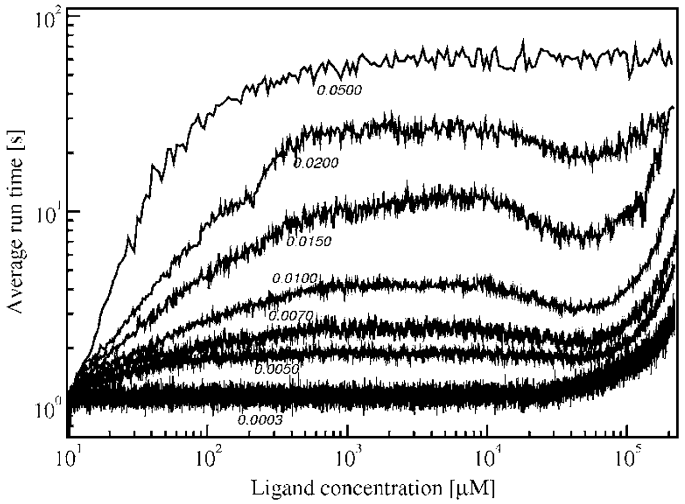


Fig. 22.2 Average run time dependence on the ligand concentration (α-methyl DL-aspartate) for different relative rates of temporal change $g = \dot{L}/L$

between the ligand, the average methylation level (adaptation control) and the average kinase activity (motor control). The model is used to study the chemotactic response for steps, linear and exponential increases in the ligand concentration, and oscillatory concentration dynamics. In combination with an experimental approach involving a microfluidic device producing stable spatial gradients, it is also used to estimate the concentration range of logarithmic sensing [8], and to show that it is the adaptation kinetics that underlies the logarithmic sensing.

In this section, we will use the four-flagella model to study the behavior of bacterial populations when the ligand concentration changes exponentially in time. For an exponential change $L = L_0 e^{gt}$, the relative change in the concentration is a constant: $\dot{L}/L = g$. Vladimirov et al. [22] modeled chemotaxis signaling pathway mathematically, based on a Monod–Wyman–Changeux model for mixed chemoreceptor clusters and a simplified description of methylation and kinase activity. They also accounted for multiple flagella through a simple voting model. Based on this model, they derived a ligand gradient on which the chemotactic activity and drift velocity are constant. In the intermediate region, this gradient can be well described by an exponential, and thus, also defines a range where *E. coli* sense logarithmically.

We simulated populations of bacteria in environments with different values of g and measured their average run times as a function of the ligand concentration L. In Fig. 22.2 we can easily identify the regime of logarithmic sensing as the range of values of L where, for a given g, the average run time does not depend on L (plateaus in the curves). In this range the bacterial response depends solely on the relative ligand change g.

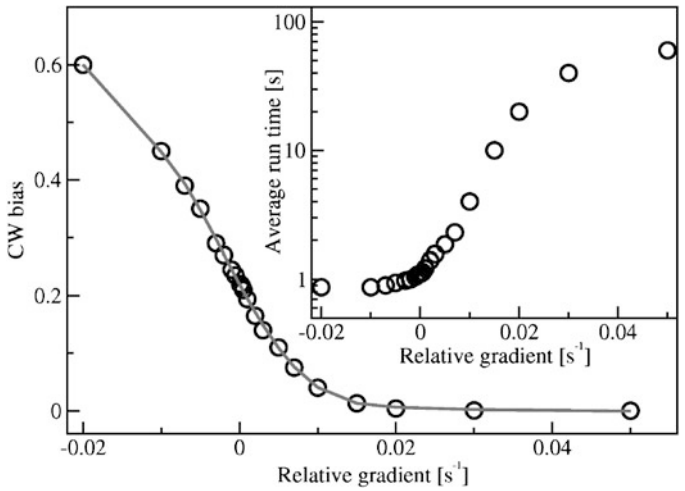


Fig. 22.3 CW bias as a function of the relative temporal change in ligand g in the regime where the sensing is logarithmic. Open symbols correspond to the simulation and the gray solid curve is the fitting function (22.6). The inset shows the dependence of the average run length on g

If we limit ourselves to the regime of logarithmic sensing, we can analyze the variation of the plateau heights in Fig. 22.2 with g and hence plot the dependence of the CW bias (Fig. 22.3) and of the average run time (Fig. 22.3 (inset)) on the relative ligand change g.

The relation of the CW bias on the relative concentration change g can be described by a fit function

$$\tau = \begin{cases} 0.2215 - 26.57\, g - 381\, g^2 & g < 0.002 \text{ (also for negative gradients)} \\ 0.24\, e^{-183.54\, g} & g \geq 0.002 \end{cases}, \tag{22.6}$$

displayed as the solid gray line in Fig. 22.3.

In Fig. 22.4 we schematically illustrate the different regimes of bacterial sensing observed in simulations. The figure displays the average run time as a function of L for a relative temporal concentration change $g = 0.02$. Here one can see that the response depends linearly on the temporal ligand change $\dot{L}$ at small values of L, followed by the logarithmic regime where the response is independent of L. For large ligand concentrations the response decreases as a function of L until saturation occurs. At saturation the receptors are fully methylated and the signal pathway is unable to adapt to the very high levels of L. For comparison we show a prediction (blue dashed curves) by a simple model of Brown and Berg [18] where the response is considered proportional to

$$\langle l \rangle \propto e^{K\dot{L}/(K+L)^2}\,. \tag{22.7}$$

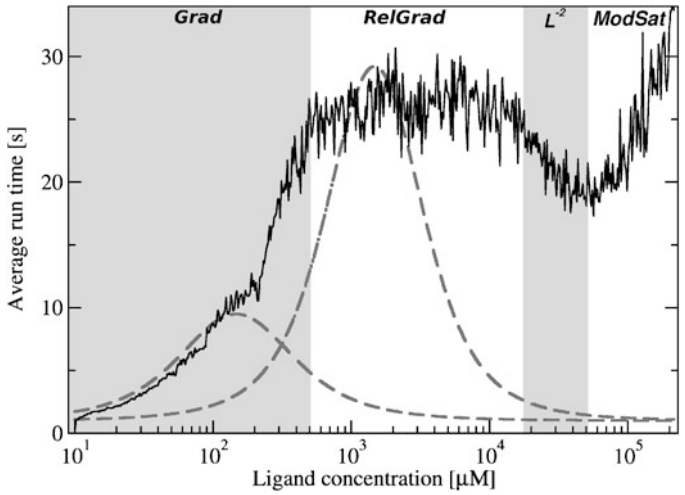


Fig. 22.4 Different sensing regimes in constant relative gradient g. For small ligand concentrations L the response is roughly proportional to $\dot{L}$ (*Grad*), in the intermediate region it is constant meaning that it depends only on $\dot{L}/L$ (*RelGrad*). For high concentrations it depends on $\dot{L}/L^2$ (L^{-2}). This region is very narrow. For yet higher values of L the model predicts receptor saturation, where the simulated bacteria cannot adapt to the concentrations any longer, but swim constantly regardless of their direction (*ModSat*). The dashed blue curves show the response predicted by a simple dissociation model (22.7) for two values of the dissociation constant K taken from [2]: $K_2 = 150\mu$M and $K_3 = 1500\mu$M. The model accounts for the *Grad* and L^{-2} regimes; however, the logarithmic sensing (*RelGrad*) regime is highly suppressed. The broadening of the *RelGrad* regime is a consequence of the multiple methylation levels of chemotactic receptors

Hereby, K is a dissociation constant describing the response in an effective way. Equation (22.7) describes the average run length being proportional to the gradient for small values of L, and also features the $e^{\dot{L}/L^2}$ dependence for large values where the response decreases with L. It fails, however, to properly describe the regime of logarithmic sensing. We have plotted two curves for two different dissociation constants associated with double and triple methylated receptors [2]. None features a broad flat logarithmic regime. The broader logarithmic regime present in bacteria and observed in our simulations is a consequence of the receptor multiple methylation levels and the slow adaptation kinetics. The range in which bacteria sense logarithmically is also affected by the rate of temporal change $\dot{L}$. For large $\dot{L}$ the regime of logarithmic sensing is much wider, and spans several magnitudes (see Fig. 22.5), which fit very well to the results of [8].

In Fig. 22.5 we illustrate the various response regimes of *E. coli* to aspartate in the form of a "phase diagram". Depending on L and $\dot{L}$ the type of response changes from linear gradient sensing (*Grad*) to logarithmic sensing (*RelGrad*). There is a small (probably insignificant) region where the response decreases (L^{-2}), and a region at high L or $\dot{L}$ where the behavior is dominated by the rotational diffusion

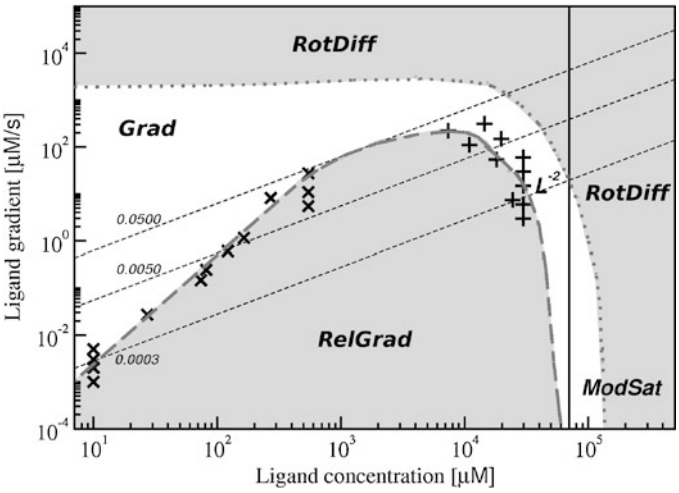


Fig. 22.5 Phase plot showing different regimes of bacterial sensing. The intermediate region between the dashed and the dotted line relates to logarithmic sensing. For small L bacteria sense the gradient $\dot{L}$. In very high gradients $\dot{L}$ or very high concentrations L the response is dominated by the effects of rotational diffusion. For very high values of L, the receptors are fully saturated and our model predicts constant swimming regardless of the orientation. Here, the run lengths are limited only by the effects of rotational diffusion. The dotted lines on the diagram represent the paths with constant relative gradient g

(*RotDiff*). In this regime the run times are very long, but Brownian rotational diffusion causes deviations from a straight trajectory, and in turn reduces the drift velocity.

4 Coarse Grained Model

The relations (22.4) and (22.5), together with (22.6) provide a coarse grained description of the bacterial response to chemical gradients, which we will use in this section to simulate chemotactic motion in different spatial gradients. There have been several attempts to model the dependence of the chemotactic response on the ligand profile. One example is an exponential dependence on the derivative of the fraction of bound receptors (see equation (8) in Brown and Berg [18]). Other models describe the bacterial memory using a kernel function (e.g., the chemotactic response function [23, 24]), or a positive delta function [25]). A different approach was proposed by Kalinin et al. [8], according to which the chemotactic response depends on the relative (spatial) gradient of the ligand. We adapt the latter, and take it one step forward by assuming that the tumble probability depends on the relative change in the ligand felt by an individual bacterium over time.

We use (22.4), (22.5), and (22.6) to carry out simulations of bacterial chemotaxis in a three-dimensional environment in the following way: During each simulation time step, a bacterium can either run or tumble, with tumble probability p. If a bacterium tumbles, the angle of its new direction relative to the old one is distributed as a Γ-distribution [11], and the azimuthal angle between the old and new direction is chosen randomly. The tumbling probability p is given for each bacterium in every time step as follows:

$$p = \frac{\Delta t}{\langle l \rangle}, \tag{22.8}$$

where $\Delta t = 0.02\,\mathrm{s}$ is the simulation time step. The average run length, $\langle l \rangle$, is calculated as a function of the CW bias according to (22.4). The CW bias is calculated given the relative gradient felt by the bacterium from (22.6). Since the velocity of the bacteria is not constant, but depends on the form and interaction of the four-flagella, the distance of the bacterial movement during Δt is determined by the CW bias-dependent velocity, (22.5). In a coarse grained way therefore our model includes the chemotactic response of the signaling pathway, the effects of the rotational diffusion and the multi-flagellar dynamics. Formulated in this way, its natural constraint is that it is valid only in the regime of chemoattractand concentrations and gradients where the logarithmic sensing applies (*RelGrad* in Fig. 22.5). It could easily be extended to include the other sensing regimes as well; however, this was not our intention here as we wanted to end up with a simple tractable coarse grained model.

With this setting we simulate the motion of bacterial populations in linear and exponential spatial gradients. Hereby, the ligand concentration is constant in time and changes only in x-direction. We choose periodic boundary conditions in the y- and z-direction, and reflecting boundary conditions in x. We traced 1,500 bacteria in the linear and 15,000 in the exponential case. Our results are compared to experimental data obtained under constant ligand profiles that are provided by microfluidic devices. Kalinin et al. [8] measured the chemotactic response of *E. coli* in linear gradients. Ahmed et al. [26] created microfluidics where the geometry of the channel allows arbitrary nonlinear gradients.

4.1 Linear Ligand Profile

Figure 22.6 shows the bacterial density at steady state (inset) and the profile decay constant. As expected, the bacterial density at steady state shows an exponential distribution (compare Fig. 2c in [8]). The decay constant of the exponential probability density is given as $\delta = v_\mathrm{d}/\mu$, where v_d is the chemotaxis drift velocity and μ the motility (or diffusion) constant [8]. Assuming that μ is constant (125–150 $\mu\mathrm{m}^2/\mathrm{s}$, [8]), δ is proportional to the drift velocity. The monotonic relationship between the decay constant δ and the relative gradient fits qualitatively very well to the results of Kalinin et al. [8].

The drift velocity reaches a maximum for high relative gradients, both in the experimental study of Kalinin et al. [8] and in our simulations. This, we believe, is

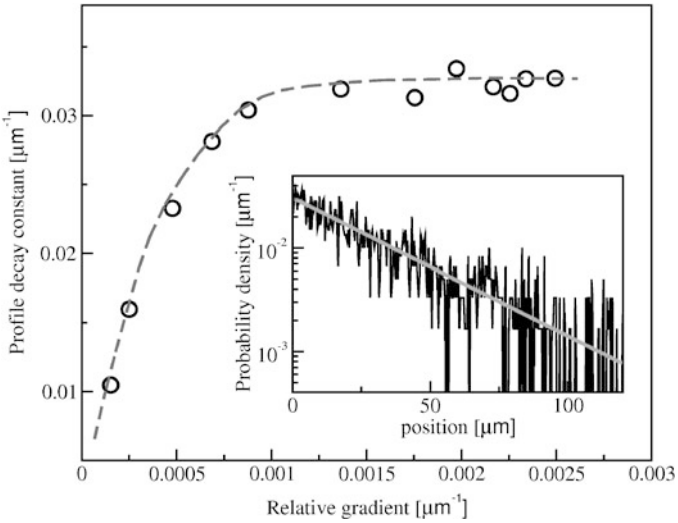


Fig. 22.6 Decay constant $\delta = v_d/\mu$ as a function of the relative gradient. The dependence of the decay constant on the relative gradient is monotonic and agrees very well with the findings of [8]. The density distribution of the bacteria at steady state, shown in the inset, is exponential as expected

due to the rotational diffusion, which becomes dominant for high relative gradients where the run lengths are large. Our results predict the regime at which the drift velocity stops growing very well (around a relative gradient of $0.001\,\mu\text{m}^{-1}$). Furthermore, the maximum drift velocity in our simulation ($\delta \cdot \mu$) is $\sim 4.8\,\mu$m/s, which is in good agreement with available data ($7\,\mu$m/s according to Berg and Turner [27], $3\,\mu$m/s according to [8]).

4.2 Exponential Ligand Profile

We next simulated the chemotactic behavior in exponential ligand gradients as shown in Fig. 22.7a. We determine numerically the bacterial density distribution in the gradient at steady state (see Fig. 22.7b), and compare it to experimental data on *E. coli* profiles in microfluidic devices that produce similar gradients [26]. We observe that our steady state density profile (black line in Fig. 22.7b) deviates from the experimental data. We argue that the time given to the bacteria in the experiments to settle to a density profile (1 h) was far too short to guarantee a steady state profile. With a diffusion constant of about $300\,\mu\text{m}^2/\text{s}$ the cells diffuse during this time only about 1 mm and the observed profiles in the microfluidic chamber of 15 mm length cannot be equilibrated. In our simulations a steady state was reached only after 25 h. This argument can be nicely supported by a very good agreement of early-time density profiles from our simulations with the experimental data, including the

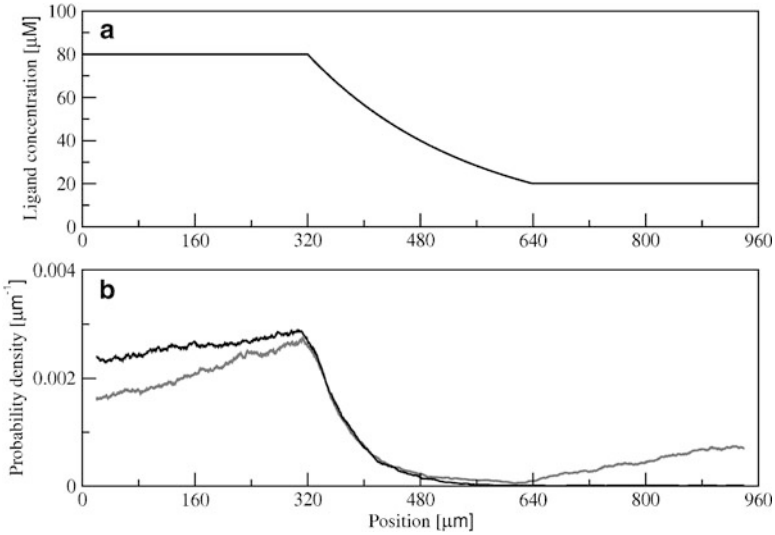


Fig. 22.7 (**a**) Ligand profile used in the coarse grained simulations. (**b**) Profiles of bacterial density at steady state (25 h after start, black line) and early-time profile (3.3 h after start, gray line). Especially the early-time profile agrees very well with the experimental data [26]

peak at the onset of exponential decay, and the increase in sensitivity at low ligand levels further away from the ramp. The gray curve shown in Fig. 22.7 corresponds to a simulation time of 3.3 h. The early-time profile (gray curve in Fig. 22.7b), and the experimentally obtained density profiles after 1 h from [26] show an increase in density next to the right boundary although the ligand concentration is constant. This is due to interaction with the boundary (in the simulations we have reflecting boundaries). In the long time course, all bacteria have the chance to escape from this region and the profile at the steady state becomes flat. Note that also the bacterial density in $x \leq 320\mu$m is not constant even though the ligand is constant there. This is because this area is affected by the exponential regime: bacteria that swim around $x = 320\mu$m have high drift velocity to the left.

Summarizing, our simulations show that the our coarse grained model, given by the three algebraic relations of (22.4), (22.5), and (22.6) provides a very simple but efficient description of *E. coli* chemotactic swimming behavior in spatial gradients falling in the region of logarithmic sensing defined as *RelGrad* in Fig. 22.5.

5 Conclusions

We have shown that the chemotactic response of *E. coli* to ligand gradients can be effectively coarse grained to describe the observed bacterial dynamics in various situations. Our present work is based on experiments of the *E. coli* response to

aspartate as a ligand; however, the bacteria will generally be exposed to various nutrients and poisons. Different ligands bind different receptors and certainly the response to them is different. Depending on the number of methylation levels involved for each receptor, there might be a smaller or a larger region of logarithmic sensing for each substance. If the experiments were available for a greater variety of ligands, the bacterial dynamics in a complex environment with many types of chemoattractands and chemorepellents could be analyzed along the same lines as we did here for aspartate. However, other phenomena, disregarded here, are also important when studying bacterial dynamics. One example is bacterial communication via chemoattractand secretion, cell death, reproduction, and hydrodynamics. The latter becomes very important in crowded environments and close to interfaces and can lead to complex pattern formation even without chemotaxis [28]. Our coarse grained approach is well suited to be combined with such effects in the future.

Acknowledgments We acknowledge the support of the following funding agencies: the Center for Modeling and Simulation in the Biosciences (BIOMS) of the University of Heidelberg (FM), the German Ministry of Education and Research (grant Nr. 03BOPAL1) (MSM), the Slovenian Research Agency (P1-0055), the European Research Council (COLSTRUCTIION 227758), and the 7th Framework Programme (ITN-COMPLOIDS 234810) (JD, JBK & TC). JD wants to acknowledge the hospitality of the Aspen Center for Physics during the summer workshop programme in August 2010.

References

1. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913–917
2. Kollmann M, Løvdok L, Bartholomé K, Timmer J, Sourjik V (2005) Design principles of a bacterial signalling network. Nature 438:504–507
3. Sourjik V, Berg HC (2002) Receptor sensitivity in bacterial chemotaxis. PNAS 99(1):123–127
4. Cluzel P, Surette M, Leibler S (2000) An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. Science 287:1652–1655
5. Bray D (1998) Receptor clustering as a mechanism to control sensitivity. Nature 393:85–88
6. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168–171
7. Hansen CH, Endres RG, Wingreen NS (2008) Chemotaxis in *Escherichia coli*: a molecular model for robust precise adaptation. PLOS Comp Biol 4(1):14–27
8. Kalinin YV, Jiang LL, Tu Y, Wu M (2009) Logarithmic sensing in *Escherichia coli* bacterial chemotaxis. Biophys J 96:2439
9. Matthäus F, Jagodić M, Dobnikar J (2009) *E. coli* superdiffusion and chemotaxis – search strategy, precision and motility. J Biophys 97(4):946–957
10. Strong SP, Freedman B, Bialek W, Koberle R (1998) Adaptation and optimal chemotactic strategy for *E. coli*. Phys Rev E 57(4):4604–4616
11. Emonet T, Macal CM, North MJ, Wickersham CE, Cluzel P (2005) Agent Cell: a digital single-cell assay for bacterial chemotaxis. Bioinformatics, 21(11):2714–2721
12. Bray D, Levin MD, Lipkow K (2007) The chemotactic behavior of computer-based surrogate bacteria. Curr Biol 17(4):R132–R134
13. Turner L, Ryu WS, Berg HC (2000) Real-time imaging of fluorescent flagellar filaments. J Bacteriol 182(10):2793–2801

14. Korobkova E, Emonet T, Park H, Cluzel P (2006) Hidden stochastic nature of a single bacterial motor. Phys Rev Lett 96(5):058105
15. Staropoli JF, Alon U (2000) Computerized analysis of chemotaxis at different stages of bacterial growth. Biophys J 78:513–519
16. Berg HC (2003) *E. coli* in motion. Biol Med Phys Biomed Eng. Springer, New York
17. Berg HC, Brown DA (1972) Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. Nature 239(5374):500–504
18. Berg HC, Brown DA (1974) Temporal stimulation of chemotaxis in *Escherichia coli*. Proc Natl Acad Soc USA 71(4):1388
19. Tu Y, Shimizu TS, Berg HC (2008) Modeling the chemotactic response of *Escherichia coli* to time-varying stimuli. PNAS, 105(39):14855–14860
20. Shimizu TS, Tu Y, Berg HC (2010) A modular gradient-sensing network for chemotaxis in *Escherichia coli* revealed by responses to time-varying stimuli. Mol Sys Biol 6:Art. no. 382
21. Jiang L, Ouyang Q, Tu Y (2010) Quantitative modeling of *Escherichia coli* chemotactic motion in environments varying in space and time. PLOS Comp Biol 6(4):e1000735
22. Vladimirov N, Løvdok L, Lebiedz D, Sourjik V (2008) Dependence of bacterial chemotaxis on gradient shape and adaptation rate. PLOS Comp Biol 4(12):e1000242
23. Block SM, Segall JE, Berg HC (1982) Impulse responses in bacterial chemotaxis. Cell 31:215
24. Kafri Y, daSilveira RA (2008) Steady-state chemotaxis in *Escherichia coli*. Phys Rev Lett 100:Art. no. 238101
25. de Gennes PG (2004) Chemotaxis; the role of internal delays. Eur Biophys J 33:691–693
26. Ahmed T, Shimizu TS, Stocker R (2010) Bacterial chemotaxis in linear and nonlinear steady microfluidic gradients. Nano Lett 10:3379
27. Berg HC, Turner L (1990) Chemotaxis of bacteria in glass capillary arrays. Biophys J hbox58:919–930
28. Cates ME, Marenduzzo D, Pagonabarraga I, Tailleur J (2010) Arrested phase separation in reproducing bacteria creates a generic route to pattern formation. Proc Natl Acad Sci USA 107:11715–11720

Chapter 23
Self-Feedback in Actin Polymerization

Anders E. Carlsson

Abstract Polymerization of actin, which is crucial for functions such as cell migration, membrane ruffling, cytokinesis, and endocytosis, must be tightly regulated in order to preserve an adequate supply of free actin monomers to respond to changing external conditions. The paper will describe mechanisms by which F-actin feeds back on its own assembly, thus regulating itself. I will present the experimental evidence for such feedback terms, discuss their use in current models of actin dynamics in cells, and present preliminary calculations for the role of feedback in transient endocytic actin patches. These calculations suggest a partial homeostasis of F-actin, in which the F-actin peak height depends only weakly on the actin filament nucleation rate.

1 Introduction

The precise regulation of actin polymerization is crucial for appropriate polymerization of actin in response to external stimuli, or for the internal dynamics of cells exploring their local environment. Upstream regulation of actin polymerization has both positive elements, which enhance actin polymerization, and negative elements, which inhibit it. Well-established positive elements include nucleation-promoting factors (NPFs), which act upstream of proteins/complexes nucleating new actin filaments [1]. These proteins include the Arp2/3 complex, which generates new actin filaments as branches on existing filaments, and formins, which generate new filaments in the absence of preexisting filaments. Well-known negative elements include severing and depolymerization induced by proteins such as cofilin. Actin has a bound nucleotide, ATP or ADP or an intermediate state denoted ADP-P_i.

A.E. Carlsson (✉)
Department of Physics, Washington University, St. Louis, MO 63130, USA
e-mail: aec@wustl.edu

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_23, © Springer Science+Business Media, LLC 2012

The hydrolysis of ATP–actin to ADP–actin favors depolymerization, and this process is accelerated by cofilin. In addition to such top–down control, feedback mechanisms can be useful in tailoring the dynamic response of cells to stimuli [2]. Positive feedback can lead to strong rapid responses, and negative feedback to stability or homeostasis. Positive F-actin feedback motifs could thus be useful in obtaining a rapid response to stimuli favoring polymerization, while negative feedback motifs could avoid excessive F-actin accumulation. The combination of positive feedback and delayed negative feedback could also provide an mechanism for rendering F-actin dynamic – in the sense of causing F-actin accumulations to be transient. This article outlines the experimental evidence for such feedback, explores possible feedback circuits embodying both positive and negative feedback, reviews the use of such circuits in recent theories of actin dynamics in cells, and presents preliminary results for the protein dynamics of endocytic actin patches based on a three-state model. Calculations of the F-actin peak height as a function of the filament nucleation rate gives a weak dependence, suggesting a partial homeostasis mechanism.

It is important to distinguish between "direct" and "indirect" feedback. "Direct" feedback occurs on a rapid enough time scale that the assembly or disassembly of F-actin at a given time is effectively proportional to the amount of F-actin present at that time; "indirect" feedback occurs after a delay, which may be generated by one or more intermediate nodes in a signaling pathway. This distinction is important because direct positive and direct negative feedback will cancel each other, simply leading to a reduced magnitude of feedback; on the other hand direct positive feedback combined with delayed negative feedback can lead to oscillatory behavior.

2 Experimental Evidence for Positive and Negative Self-Feedback of F-actin

Evidence for positive feedback of F-actin lies in the dynamics of actin polymerization in vitro induced by Arp2/3 complex. The Arp2/3 complex, when activated by upstream agents such as NPFs, forms new growing branches on the sides of existing filaments both in cells and in vitro. Because the rate of nucleation of new filaments increases with the amount of F-actin present, a small number of filaments can rapidly multiply and the F-actin concentration initially grows exponentially. This means that the time course of polymerization often appears to have a "lag phase", where polymerization is limited, followed by an explosive growth of polymerization. Figure 23.1 compares time courses of actin polymerization assuming positive feedback (solid line), in which the number of filaments increases proportionally to the F-actin concentration, and no feedback (dashed line), in which the number of filaments is constant. The positive feedback curve clearly demonstrates a lag phase, followed by a rapid jump to the final value which is the total actin concentration minus the critical concentration. The

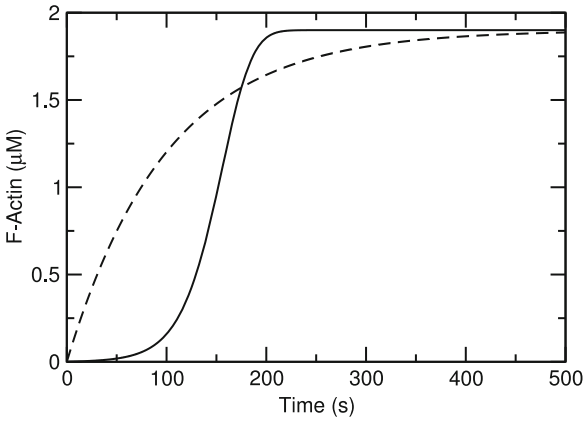


Fig. 23.1 Schematic of actin polymerization with (*solid line*) and without (*dashed line*) positive feedback. Curves obtained using rate equations described in [5] with arbitrary parameters

no feedback curve increases much less dramatically. There are many examples of Arp2/3-induced polymerization time courses in the literature which are similar to the positive feedback curve shown here [3–7]. The positive feedback involves two steps, the generation of new free barbed ends on existing filaments, and the growth of these filaments to generate new F-actin. This is indicated by the positive feedback lines in Fig. 23.2a. Thus it is in principle an indirect feedback mechanism. However, if the delay between the generation of new filaments and the subsequent F-actin accumulation is small, the feedback may be considered direct, for practical purposes. This delay is determined by the lifetime of a free barbed end, which is determined by the capping rate. In general, capping rates in cells are on the order of $1\ s^{-1}$, which is faster than most of the dynamic F-actin processes to be discussed below. Therefore positive feedback of F-actin may legitimately be viewed as direct, as indicated in Fig. 23.2b.

Evidence for indirect negative feedback of F-actin, as indicated by the negative feedback lines in Fig. 23.2, comes from studies in cells and in vitro. Weiner et al. [8] measured the dynamics of the Hem-1 component of the WAVE2 complex in neutrophils, an NPF. The WAVE2 complex acts upstream of actin polymerization by activating Arp2/3 complex, and is required for proper leading edge morphology; its homologs regulate cell shape and movement in a variety of organisms. Weiner et al. performed fluorescence recovery after photobleaching (FRAP) experiments using fluorescently labeled Hem-1. In these experiments, a 1–2 μm spot was bleached, and the recovery was followed as a function of time. The recovery involves Hem-1 leaving the membrane and being replaced. The experiments were performed both in the presence and absence of latrunculin (Lat), an agent which inhibits actin polymerization. It was found that Lat greatly slowed the recovery process. This implies that actin polymerization is essential for the dynamics of Hem-1 leaving and entering the membrane.

Kaksonen et al. ([9] evaluated the effects of F-actin on the lifetimes of proteins acting upstream of actin in transient protein patches occurring during endocytosis in budding yeast. Under control conditions (Ctrl), a number of proteins, including

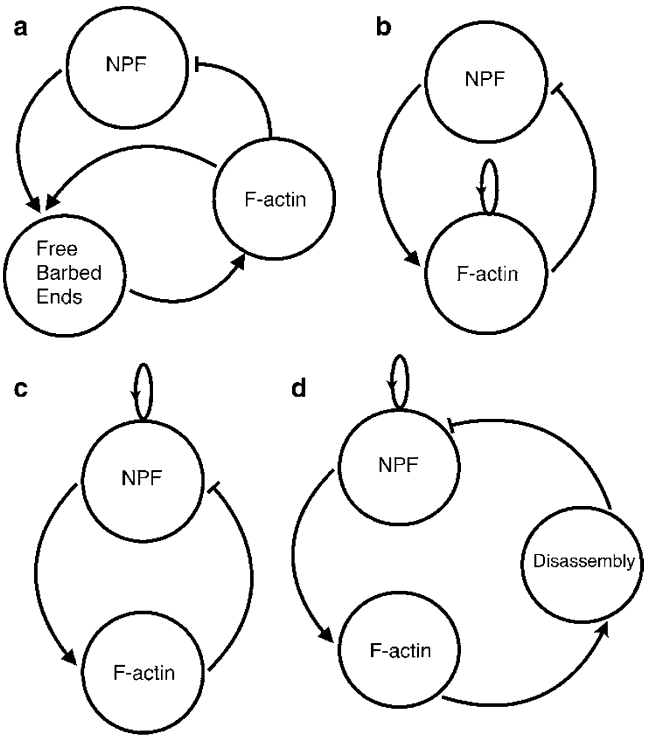


Fig. 23.2 Possible F-actin feedback circuits generating F-actin dynamics. *Arrows* indicate positive feedback and bars at ends of lines indicate negative feedback. *Loops* indicate positive feedback

Clc1, Bbc1, and End3, assemble first. These are followed by F-actin, and then by a series of proteins, including cofilin, which disassembles F-actin. It was found that suppression of actin polymerization by the addition of Lat greatly lengthened the lifetimes of Clc1, Bbc1, and End3, as indicated in Fig. 23.3. This shows that F-actin disassembles proteins which act upstream of actin polymerization, thus demonstrating an indirect negative feedback loop, as in Fig. 23.2.

In vitro studies [10] have shown that F-actin inhibits the Abl tyrosine kinase, which regulates actin polymerization [11]. Measurement of purified Abl-kinase activity as a function of increasing F-actin concentration revealed a steady drop. Furthermore, time-dependent growth of Abl-kinase activity was found to be inhibited by F-actin.

Ganguly et al. studied the effect of serotonin receptor activity on actin polymerization [12] in Chinese hamster ovary (CHO) cells, and conversely the effect of F-actin on receptor mobility and activity [13]. It was found that serotonin-mediated activation of the receptor led to increased F-actin content, presumably through inhibition of the production of cyclic AMP (cAMP), which causes actin depolymerization. F-actin, in turn, inhibited serotonin receptor mobility, which was

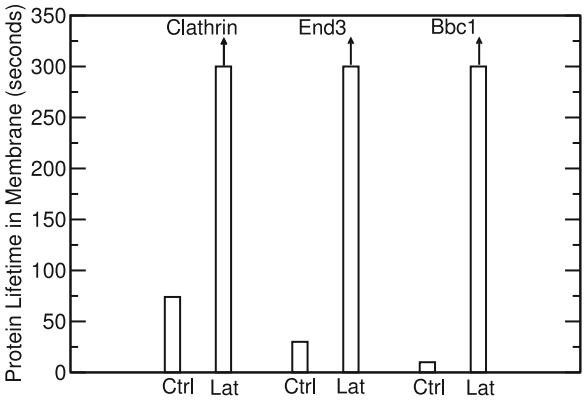


Fig. 23.3 Lifetimes of endocytic coat proteins, under control conditions (Ctrl) and with latrunculin (Lat). Data taken from Ref. [9]. *Arrows* mean that bars are lower bounds for lifetimes

found to reduce the receptor efficiency. These observations, taken together, suggest a multistep negative feedback loop of the type shown in Fig. 23.2d, where the disassembly module summarizes effects of F-actin on receptor mobility/efficiency, the effects of serotonin receptors on cAMP production, and the pathways allowing cAMP to depolymerize F-actin.

3 Calculations of the Effect of Self-Feedback on F-actin Dynamics in Cells

Mathematical modeling of F-actin dynamics in cells has shown that the types of feedback effects discussed here can lead to the formation of spontaneous dynamic phenomena such as F-actin waves and patches [14]. These models have used feedback architectures of the types illustrated in Fig. 23.2. The general mechanism by which such feedback loops lead to transient F-actin structures, using Fig. 23.2c for concreteness, is the following. Positive feedback of NPF causes a small fluctuation of NPF to grow. F-actin then builds up because NPF activates F-actin. Since F-actin inhibits NPF, the NPF will drop to zero after a while, and subsequently the F-actin will drop to zero. If diffusion is present, this can result in traveling waves or patches.

Two models in the literature [8, 15] used an architecture of the type shown in Fig. 23.2c. These models led to spontaneous waves of F-actin and the NPF Hem-1; [15] also found actin patches under some conditions. Other models have used positive F-actin feedback to treat the spontaneous formation of F-actin waves and patches in *Dictyostelium*. The model of reference [16] used a feedback circuit of the form shown in Fig. 23.2b, and assumed spontaneous polarization of filament orientations and diffusion-like spreading of F-actin. These assumptions were found to lead to the formation of patches which eventually coalesced into traveling waves. The treatment of [17] implemented the circuit in Fig. 23.2a,

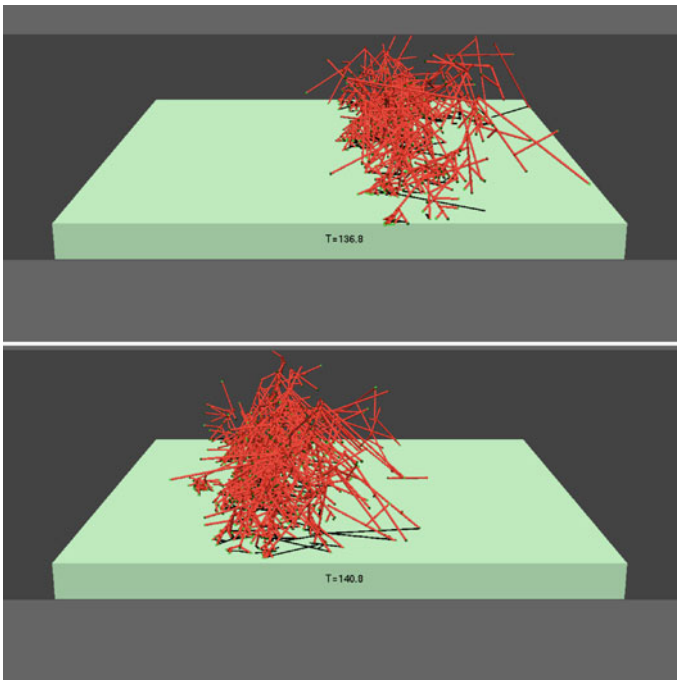


Fig. 23.4 Progression of actin wave in stochastic-simulation approach [17]. Red rods denote actin filaments; substrate-bound membrane of cell is green

using a dendritic-nucleation model of actin filament generation, in which new filaments are generated as branches on existing filaments by the action of NPFs in the plasma membrane. The calculations were performed using a stochastic-growth methodology in which explicit three-dimensional network structures were generated, and dendritic filament clusters moved by Brownian motion. This avoided the explicit parameterization of positive feedback and diffusion effects; instead, these effects emerged naturally from the known biochemistry. The calculations revealed, with increasing actin concentration, a series of phases beginning with patches, subsequently waves, and finally a phase which could not be described in terms of either traveling waves of patches, but still displayed fluctuations which appeared to be strongly out of equilibrium. Figure 23.4 shows an example of the wave phase. A forest of actin filaments is seen moving to the left. The "story line" of the motion is as follows: (1) there is a high concentration of membrane-bound (active) NPF ahead of the wave; (2) F-actin grows into this region; and (3) the F-actin removes and inactivates the NPF, causing subsequent depolymerization of F-actin at the back of the wave.

A larger set of feedback interactions has been treated in a recent study of endocytosis in budding yeast [18]. During this process, actin patches form at the cortex and, disappear after approximately 20 s. The model used has roughly the

structure of Fig. 23.2d. NPF was assumed to assemble in response to membrane-bound PIP2, which had a positive feedback rate law. F-actin, rather than directly inhibiting NPFs, was taken to cause membrane curvature. This curvature, in turn, caused PIP2 hydrolysis, which reduced NPF accumulation. In other words, the "disassembly" bubble in Fig. 23.2d would include curvature and PIP2 hydrolysis – a multistep negative feedback of F-actin on NPFs. This set of assumptions led to the appearance of transient patches, and made several predictions about the effects of key rates on the success of endocytosis.

4 The Systems Biology of a Transient Actin Patch

Motivated by the modeling of [18], we have attempted to abstract the key ingredients required to obtain transient actin patches by using a simpler model, based on Fig. 23.2d, with only one variable in the "disassembly" bubble. We do not specify the physical nature of this variable, but rather deduce its dynamics from its impact on the assembly/dissasembly of the NPF module and F-actin. Because of this simplicity, it is straightforward to evaluate the effects of key rates on measurable output properties. The model has three variables. The first, $[N]$, includes the coat proteins which arrive first in endocytosis, and the NPFs; the second, $[F]$, includes F-actin and associated actin-binding proteins such as Arp2/3 complex and capping protein; and the third $[D]$, includes the factors that are most important in disassembly of the coat proteins.

The initial assembly of the coat protein/NPFs is assumed to proceed by the generation of accumulation nuclei from membrane proteins in a restricted "corral" region. The nuclei grow when their size exceeds a critical size. The equations of motion are:

$$\frac{\mathrm{d}[N]}{\mathrm{d}t} = k_N^+(N_0 - [N]) - k_N^+ N_0 \exp\left[\varepsilon_\mathrm{s}\left(1/\sqrt{N} - 1/\sqrt{N_\mathrm{c}}\right)\right] - k_N^-[D]^i[N] \quad (23.1)$$

$$\frac{\mathrm{d}[F]}{\mathrm{d}t} = k_F^+[N]^j - k_F^-[F] \quad (23.2)$$

$$\frac{\mathrm{d}[D]}{\mathrm{d}t} = k_D^+[F] - k_D^-[D]. \quad (23.3)$$

Here N_0 is the initial number of coat/NPF proteins in the corral, N_c is a critical cluster size, and ε_s is a dimensionless surface energy parameter; i and j are powers characterizing the cooperativity of NPFs in assembling actin and the nonlinearity of coat/NPF disassembly. The k's are on- and off-rate parameters for the three species. The cooperativity of the NPFs in assembling actin (j) is based on experimental observations indicating a high degree of cooperativity in the activity of the NPF WASp [19], which has the analog Las17 in yeast. The cooperativity in disassembly

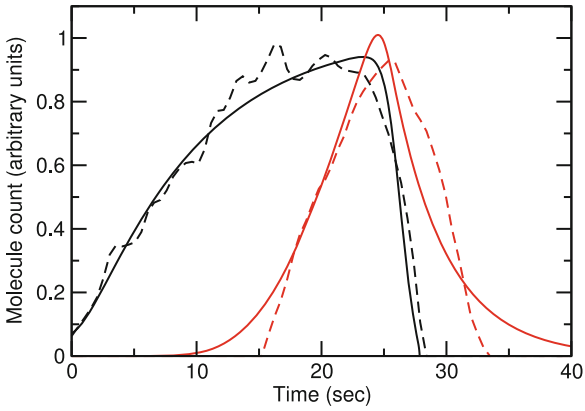


Fig. 23.5 *Solid lines*: time courses of coat/NPF (black) and F-actin (red) obtained from deterministic simulations. *Dashed lines* denote experimental data from [9]

(i) could come from several sources [18], one of which is an exponential dependence of coat protein off-rates on membrane curvature.

The form of (23.1) is motivated by classical nucleation theory [20], in which the ratio of the on-rates to the off-rates is determined by the free energy of cluster formation and its dependence on cluster size; the square-root terms in (23.1) reflect the contribution of the surface energy (which is proportional to $\sqrt{N}$ for a protein patch growing in two dimensions). Note that the factor in front of the exponential can safely be taken to be k_N^+, since adjusting this factor is equivalent to adjusting N_c.

The parameters in this model were adjusted to fit measured data for the dependence of a coat protein/NPF (Las17) and F-actin (as indicated by the actin-binding protein Abp1) in budding yeast. The parameters adjusted were: an overall factor scaling the model results to the experimental data, k_N^+, k_F^+, k_F^-, and k_D^+. The parameter k_N^- was assigned a fixed value since changes in k_N^- can be compensated for by changes in the normalization of $[D]$, and $[D]$ is not included in the fitting database. To reduce the number of fitting parameters, and because disassembly of D is expected to at least partly require disassembly of F, it was assumed that $k_D^- = 0.25k_F^-$. The exponents i and j were assigned the value 8; smaller values than this tended to give a worse fit to the data. As shown in Fig. 23.5, the modeling at this level gives a good fit to the averaged patch-count data. The life cycle of the patch is that first NPF builds up, causing F-actin buildup; the F-actin in turn causes disassembly factors to build up, which disassemble the NPF patch.

This model makes several predictions regarding the behavior of key observables on underlying rates and concentrations. For example, k_F^+ is expected to influence the height and lifetime of the actin peak, and thus the NPF lifetime as well. Figure 23.6 shows the dependence of the F-actin peak height and the patch lifetime on k_F^+. It is seen that the peak height increases with k_F^+, but at a rate slower than linear. The slowness of the increase is a manifestation of the negative feedback loop connecting

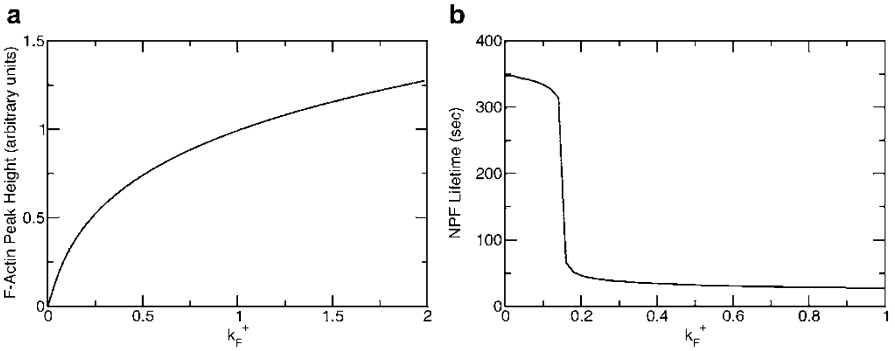


Fig. 23.6 F-actin peak height (**a**) and NPF patch lifetime (**b**) as functions of actin polymerization/nucleation parameter k_F^+, which is given in units of the best-fit value from the fits in Fig. 23.5

F-actin with itself: increased F-actin at a certain time leads to increased disassembly agent buildup, which in turn removes NPF and inhibits actin polymerization. On the other hand, decreasing k_F^+ relative to its fitted value has little effect until it is quite small; then the lifetime climbs very rapidly. (The "shelf" at small k_F^+ should be viewed as a lower bound, because the simulations were only run out to 400 s.) These predictions could be tested by currently available experimental techniques. Since k_F^+ is expected to increase with increasing actin concentration, agents such as Lat, which sequester free actin monomers, should reduce k_F^+. Similarly, mutations in those NPF domains which activate Arp2/3 complex directly or indirectly [21] should lead to reduced k_F^+. This theory predicts that such mutations, unless they affect all or nearly all NPF activators, would have little effect on the F-actin peak height. This may be viewed as a partial homeostasis mechanism in which the F-actin peak height is robust to weak perturbations.

5 Summary

Several lines of evidence indicate that F-actin, rather than simply following its upstream activators passively, feeds back on its own production in ways that can either enhance or inhibit polymerization, depending on time scale. Such feedback effects may be a key factor in regulating transient F-actin accumulations such as those in F-actin waves or endocytic patches, and in stabilizing quantities such as the F-actin peak height to perturbations in underlying rate parameters.

Acknowledgment This work was supported by the National Institutes of Health under Grant R01 GM086882.

References

1. Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. Cell 112:453–456
2. Alon, U (2007) An introduction to systems biology. Taylor and Francis, New York
3. Pantaloni D, Boujemaa R, Didry D, Gounon P, Carlier MF (2000) The Arp2/3 complex branches filament barbed ends: functional antagonism with capping proteins. Nat Cell Biol 2:385–391
4. Amann KJ, Pollard TD (2001) The Arp2/3 complex nucleates actin filament branches from the sides of existing filaments. Nat Cell Biol 3:306–310
5. Carlsson AE, Wear MA, Cooper JA (2004) End vs. side branching by Arp2/3 complex. Biophys J 86:1074–1081
6. Goley ED, Ohkawa T, Mancuso J, Woodruff JB, D'Alessio JA, Cande WZ, Volkman LE, Welch MD (2006) Dynamic nuclear actin assembly by Arp2/3 complex and a baculovirus wasp-like protein. Science 314:464–467
7. Tehrani S, Tomasevic N, Weed S, Sakowicz R, Cooper J (2007) Src phosphorylation of cortactin enhances actin assembly. Proc Natl Acad Sci 104:8827–88323
8. Weiner OD, Marganski WA, Wu LF, Altschuler SJ, Kirschner MW (2007) An actin-based wave generator organizes cell motility. PLoS Biol 5:2053–2063
9. Kaksonen M, Toret CP, Drubin DG (2005) A modular design for the clarhtin- and actin-mediated endocytosis machinery. Cell 123(2):305–320
10. Woodring PM, Hunter T, Wang JYJ (2001) Inhibition of c-abl tyrosine kinase activity by filamentous actin. J Biol Chem 276:27104–27110
11. Lanier LM, Gertler FB (2000) From Abl to actin: Abl tyrosine kinase and associated proteins in growth cone motility. Curr Opin Neurobiol 10:80–87
12. Ganguly A, Saxena R, Chattopadhyay A (2011) Reorganization of the actin cytoskeleton upon G-protein coupled receptor signaling. Biochim Biophys Acta-Biomembranes 1808:1921–1929
13. Ganguly S, Pucadyil A, Chattopadhyay A (2008) Actin cytoskeleton-dependent dynamics of human serotonin$_{1a}$ receptor correlates with receptor signaling. Biophys J 95:451–463
14. Carlsson AE (2010) Actin dynamics: from nanoscale to microscale. Ann Rev Biophys 39:91–110
15. Doubrovinski K, Kruse K (2008) Cytoskeletal waves in the absence of molecular motors. Europhys Lett 83:18003
16. Whitelam S, Bretschneider T, Burroughs NJ (2009) Transformation from spots to waves in a model of actin pattern formation. Phys Rev Lett 102:198103
17. Carlsson AE (2010) Dendritic actin filament nucleation causes traveling waves and patches. Phys Rev Lett 104:228102
18. Liu J, Sun Y, Drubin DG, Oster GF (2009) The mechanochemistry of endocytosis. PLoS Biol 7:e1000204
19. Padrick SB, Cheng HC, Ismail AM, Panchal SC, Doolittle LK, Kim S, Skehan BM, Umetani J, Brautigam CA, Leong JM, Rosen MK (2008) Hierarchical regulation of WASP/WAVE proteins. Mol Cell 32:426–438
20. Kelton KF, Greer AL (2007) Nucleation in condensed matter: applications in materials and biology. Elsevier, Boston
21. Galletta BJ, Chuang DY, Cooper JA Distinct roles for Arp2/3 regulators in actin assembly and endocytosis. PLoS Biol 6:72–85

Part IV
Computational Tools, Algorithms and Theoretical Methods for Systems Biology

Chapter 24
Global Optimization in Systems Biology: Stochastic Methods and Their Applications

Eva Balsa-Canto, J.R. Banga, J.A. Egea, A. Fernandez-Villaverde, and G.M. de Hijas-Liste

Abstract Mathematical optimization is at the core of many problems in systems biology: (1) as the underlying hypothesis for model development, (2) in model identification, or (3) in the computation of optimal stimulation procedures to synthetically achieve a desired biological behavior. These problems are usually formulated as nonlinear programing problems (NLPs) with dynamic and algebraic constraints. However the nonlinear and highly constrained nature of systems biology models, together with the usually large number of decision variables, can make their solution a daunting task, therefore calling for efficient and robust optimization techniques.

Here, we present novel global optimization methods and software tools such as cooperative enhanced scatter search (eSS), AMIGO, or DOTcvpSB, and illustrate their possibilities in the context of modeling including model identification and stimulation design in systems biology.

1 Introduction

The use of optimization has allowed biologists not only to describe patterns or mechanisms but also to predict, from first principles, how organisms should be designed [6, 41]. In particular, mathematical optimization (1) is the underlying hypothesis for model development in for example flux balance analysis [21] or the activation of metabolic pathways [22, 29, 47], (2) is at the core of model

E. Balsa-Canto (✉) • A. Fernandez-Villaverde • G.M. de Hijas-Liste • J.R. Banga
(Bio)Process Engineering Group, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain
e-mail: ebalsa@iim.csic.es; afvillaverde@iim.csic.es; gundian@iim.csic.es; julio@iim.csic.es

J.A. Egea
Department of Applied Mathematics and Statistics, Technical University of Cartagena (UPCT), Cartagena, Spain
e-mail: josea.egea@upct.es

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_24,

identification, including parameter estimation and optimal experimental design [7], or (3) enables the computation of optimal stimulation procedures to synthetically achieve a desired biological behavior [25,35].

Most of these problems are formulated as nonlinear programing problems (NLPs) where the objective is to find a set of decision variables (or functions) in order to minimize or maximize a given cost function (or functional) subject to a set of dynamic and algebraic constraints. The solution of such problems requires the use of advanced numerical optimization methods. In this regard, hundreds of different methods are at hand: from deterministic local methods to sophisticated metaheuristics. One aspect that should be taken into account at the time of selecting the most appropriate method is the nature of the problem under consideration.

Whereas convex problems present a unique solution which may be found with deterministic local methods, finding the global optimum for multimodal problems, i.e., those presenting multiple local optima, including noisy problems, typical in dynamic systems due to the numerical integration of complex partial and ordinary differential equations (ODEs), requires robust and efficient global optimization methods.

Some of these methods have been incorporated in software tools devoted to modeling, model analysis, simulation, and parameter estimation such as: COPASI [18], SBToolbox2 [37], or PottersWheel [26].

In this work we present novel global optimization methods and software tools developed at our group which are devoted to handle not only parameter estimation but also different optimization problems in the context of systems biology. In this regard we will introduce:

- DOTcvpSB [17], which is the first toolbox for dynamic optimization (DO) problem in Systems Biology, i.e., offering the possibility of handling dynamic FBA problems, optimal enzymatic activation problems, or the optimal design of stimulation profiles to achieve certain desired biological behaviors.
- AMIGO, which covers all the steps of the iterative identification procedure [2]: local and global sensitivity analysis, parameter estimation, identifiability analysis, and optimal experimental design. The robust identifiability analysis, parameter estimation, and optimal experimental design problems are formulated and solved as general (dynamic) optimization problems.
- A multi-thread cooperative scatter search approach based on enhanced scatter search (eSS) [14] is presented here as a means to handle large-scale multimodal problems. We remark that the cooperative eSS could be incorporated as an optimizer in both DOTcvpSB and AMIGO.

Four illustrative examples have been selected to show their applicability.

DOTcvpSB [17] is used to solve a problem related to the enzyme activation in a branched reaction network. The advantages of the cooperative scatter search approach are illustrated through the solution of a parameter estimation problem related to the modeling of the central carbon metabolism in *Escherichia coli*. AMIGO is used to solve an optimal experimental design problem related to a three-step metabolic pathway model. And the last example illustrates how hybrid optimization

methods incorporated in DOTcvpSB are able to solve a highly multimodal problem related to the computation of the optimal stimulation conditions to obtain a given multicellular structure in bacterial chemotaxis.

2 Optimization Problem Formulation

Consider a general dynamic and possibly distributed system described by the following state space equations:

$$\dot{\mathbf{y}} = \Xi(\mathbf{x}, \mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, t); \quad \mathbf{x}_t = \Psi(\mathbf{x}, \mathbf{x}_{\boldsymbol{\xi}}, \mathbf{x}_{\boldsymbol{\xi}\boldsymbol{\xi}}, \mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, t), \tag{24.1}$$

where $\boldsymbol{\xi} \in \Omega \subset \Re^3$ are the spatial variables, $\mathbf{x}(\boldsymbol{\xi}, t) \in X \subset \Re^\nu$ is the subset of state variables depending on both time and spatial location, $\mathbf{y}(t) \in Y \subset \Re^\mu$ is the subset of time-dependent variables, $\mathbf{x}_{\boldsymbol{\xi}} = \partial \mathbf{x}/\partial \boldsymbol{\xi}$, $\mathbf{x}_{\boldsymbol{\xi}\boldsymbol{\xi}} = \partial^2 \mathbf{x}/\partial \boldsymbol{\xi}^2$, $\mathbf{x}_t = \partial \mathbf{x}/\partial t$, $\dot{\mathbf{y}} = \mathrm{d}\mathbf{y}/\mathrm{d}t$, $\mathbf{u} \in U \subset \Re^\sigma$ are the control variables and $\boldsymbol{\theta} \in \Theta \subset \Re^\eta$ are time-independent parameters.

In addition, state variables are subject to initial and boundary conditions:

$$\mathbf{y}(t_0) = \Xi_0(\mathbf{x}(t_0), \mathbf{u}(t_0), \boldsymbol{\theta}, t_0) \tag{24.2}$$

$$\mathbf{x}(t_0) = \Psi_0(\mathbf{y}(t_0), \mathbf{u}(t_0), \boldsymbol{\theta}, t_0); \quad \mathscr{B}(\mathbf{x}, \mathbf{x}_{\boldsymbol{\xi}}, \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\xi}, t) = 0; \quad \boldsymbol{\xi} \in \Omega \tag{24.3}$$

Note that the formulation in (24.1)–(24.3) can be used to model many biological systems such as biochemical pathways, e.g., cell signaling or metabolic pathways; diffusion reaction systems, e.g., pattern formation or persistence and extinction of species, etc.

State and control variables may be also subject to algebraic constraints which force the satisfaction of particular biological conditions at particular time points or throughout the process:

$$\mathbf{r}_k^{\mathrm{eq}}(\mathbf{x}(\boldsymbol{\xi}, t_k), \mathbf{y}(t_k), \mathbf{u}(t_k), \boldsymbol{\theta}, t_k) = 0; \quad \mathbf{r}_k^{\mathrm{in}}(\mathbf{x}(\boldsymbol{\xi}, t_k), \mathbf{y}(t_k), \mathbf{u}(t_k), \boldsymbol{\theta}, t_k) \leq 0 \tag{24.4}$$

$$\mathbf{c}^{\mathrm{eq}}(\mathbf{x}(\boldsymbol{\xi}, t), \mathbf{y}(t), \mathbf{u}(t), \boldsymbol{\theta}, t) = 0; \quad \mathbf{c}^{\mathrm{in}}(\mathbf{x}(\boldsymbol{\xi}, t), \mathbf{y}(t), \mathbf{u}(t), \boldsymbol{\theta}, t) \leq 0. \tag{24.5}$$

Control variables and parameters may be also subject to bound constraints:

$$\mathbf{u}^L \leq \mathbf{u}(t) \leq \mathbf{u}^U; \quad \boldsymbol{\theta}^L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}^U. \tag{24.6}$$

The last element in the problem definition will be the objective functional that quantifies the quality of a solution:

$$J = \phi(\mathbf{x}(\boldsymbol{\xi}, t_{\mathrm{f}}), \mathbf{y}(t_{\mathrm{f}}), \boldsymbol{\theta}, t_{\mathrm{f}}) + \int_{t_0}^{t_{\mathrm{f}}} L(\mathbf{x}(\boldsymbol{\xi}, t), \mathbf{y}(t), \mathbf{u}(t), \boldsymbol{\theta}, \boldsymbol{\xi}, t)\mathrm{d}t \tag{24.7}$$

where the scalar functions ϕ (Mayer term) and L (Lagrangian term) are continuously differentiable with respect to all of their arguments, and the final time t_f can be either fixed or free.

This objective functional may be related to, for example, the quantity of metabolites produced in a metabolic pathway, to the distance among experimental data and model predictions in the case of parameter estimation or to the information provided by an experimental scheme in the case of optimal experimental design.

The general *dynamic optimization* problem considered here can be then formulated as: Find the controls $\mathbf{u}(t)$ and the time-invariant parameters $\boldsymbol{\theta}$ subject to the system dynamics in (24.1)–(24.3) and the algebraic constraints in (24.4)–(24.6) so as to minimize (or maximize) the objective functional in (24.7).

3 Numerical Methods

There are several alternatives for the solution of DO problems from which the indirect and the direct methods are the most widely used. The *indirect methods* make use of the Pontryagin's maximum principle so as to obtain the optimality necessary conditions. The method relies on the formulation of the Hamiltonian by summing the cost functional, the product of multiplier functions (co-states) with the dynamic equations in (24.1) and the product of the Lagrange multipliers with algebraic constraints and the subsequent derivation of the corresponding first and second order derivatives on the decision variables. The result will be a two or multi-point boundary value problem which must be solved for the state and co-state variables [11]. However, the complexity of the numerical solution of such boundary value problems has motivated the use of direct methods for most realistic applications.

Direct methods such as the complete parameterization (CP, [9]), multiple shooting (MS, [10]), or control vector parameterization (CVP, [44]) transform the DO problem into a NLP. These methods discretize and approximate either the control variables or both the control and state variables in such a way that the decision variables for the NLP are related to the given parameterization scheme. The three alternatives basically differ in the resulting number of decision variables, in the presence or absence of parameterization-related constraints and in the necessity of using a boundary value problem solver. While the CP or the MS approaches may become prohibitively expensive in computational terms, the CVP approach allows handling large-scale DO problems without solving very large NLPs and without dealing with extra junction constraints.

3.1 Control Vector Parameterization

The CVP method proceeds dividing the duration of the process into a number ρ of control intervals and the control function is approximated using a low order polynomial form over each interval. Each control variable approximation may be expressed using Lagrange polynomials as follows:

$$u_j(t) = \sum_{i=1}^{M_j} u_{ij} \Phi_i^{(M_j)}(\tau), \tag{24.8}$$

where, $j = 1, \ldots, \rho$, $t \in [t_0, t_f]$, and τ is normalized time given by,

$$\tau = \frac{t - t_0}{t_f - t_0} \tag{24.9}$$

and the Lagrange polynomials of order M, $\Phi_i^{(M_j)}$ are defined in the standard form,

if $M = 1$

$$\Phi_i^{(M)}(\tau) \equiv 1, \tag{24.10}$$

if $M \geq 2$

$$\Phi_i^{(M)}(\tau) \equiv \prod_{i'=1, i \neq 1}^{M} \frac{\tau - \tau_{i'}}{\tau_i - \tau_{i'}}. \tag{24.11}$$

The parameters of these polynomials, u_{ij}, will be used as decision variables in the optimization process together with time-independent parameters.

The generalization of the CVP approach for the case of optimal experimental design may be found in [1].

3.2 Boundary Value Problem Solution

The solution of the nonlinear dynamic, sometimes distributed, models describing biological systems (24.1) requires the use of suitable numerical techniques. For the most general case involving partial differential equations (PDEs) numerical methods use some type of space parameterization approach to transform the PDEs into an equivalent set of ODEs [36]. The numerical method of lines and the finite element method are the most widely used approaches for this transformation. The underlying idea is to discretize the domain of interest into many smaller subdomains and use local spatial functions to approximate the distributed variables in each subdomain. As a result a large-scale, usually stiff, set of ODEs is obtained which may be solved with a sparse implicit initial value problem solver.

3.3 Nonlinear Programing Methods

Nonlinear programing methods may be largely classified in two main groups: local and global. Local methods are designed to generate a sequence of solutions, using some type of pattern search or gradient and Hessian information, that will converge to a local optimum, usually the closest to the provided initial guess. However the NLPs with nonlinear dynamic constraints (such as in parameter estimation or the ones resulting from the application of the CVP approach) are frequently multimodal (i.e., presenting multiple local optima) [6, 7]. Therefore, local methods may converge to local solutions, especially if they are started far away from the global optimum. In order to surmount these difficulties, global methods must be used.

3.3.1 Global Optimization Methods

Global methods have emerged as the alternative to search the global optimum [30]. The successful methodologies combine effective mechanisms of exploration of the search space and exploitation of the previous knowledge obtained by the search. Depending on how the search is performed and the information is exploited the alternatives may be classified in three major groups: deterministic, stochastic, and hybrid.

Global deterministic methods [16, 31] in general take advantage of the problem's structure and guarantee global convergence for some particular problems that verify specific smoothness and differentiability conditions. Although they are very promising and powerful, there are still limitations to their application, particularly for nonlinear dynamic systems, since the computational cost increases rapidly with the size of the considered dynamic system and the number of decision variables.

Global stochastic methods do not require any assumptions about the problem's structure. They make use of pseudo-random sequences to determine search directions toward the global optimum. This leads to an increasing probability of finding the global optimum during the run time of the algorithm, although convergence may not be guaranteed. The main advantage of these methods is that, in practice, they rapidly arrive to the proximity of the solution.

The most successful approaches lie in one (or more) of the following groups: pure random search and adaptive sequential methods, clustering methods, or metaheuristics. Metaheuristics are a special class of stochastic methods which have proved to be very efficient in recent years. They include both population (e.g., genetic algorithms) or trajectory-based (e.g., simulated annealing) methods. They can be defined as guided heuristics and many of them try to imitate the behavior of natural or social processes that seek for any kind of optimality [42]. Some of these strategies have been successfully applied to, for example, parameter estimation [27, 28, 40] or optimal experimental design [1] in the context of systems biology.

Despite the fact that many stochastic methods can locate the vicinity of global solutions very rapidly, the computational cost associated to the refinement of the solution is usually very large. In order to surmount this difficulty, hybrid methods, and metaheuristics have been recently presented for the solution of DO problems [4, 13] or parameter estimation problems [3, 33, 34]. They speed up these methodologies while retaining their robustness and, provided a gradient-based local method is used, they guarantee convergence to a gradient zero solution.

In particular, the Scatter Search metaheuristic [15] is an evolutionary hybrid optimization method that has been successfully applied to the solution of not only parameter estimation problems [32, 34, 46] but also DO [13] and optimal experimental design [2] problems. The newest version, the eSS method (www.iim.csic.es/~gingproc/ssmGO.html, [14]), presents a simpler but more effective design which helps to overcome typical difficulties of nonlinear dynamic systems optimization such as noise, flat areas, nonsmoothness, and/or discontinuities.

4 Illustrative Examples

4.1 Optimal Enzyme Activation in Metabolic Networks

An example of the insight that optimization can provide concerns the enzyme activation in metabolic networks. Several authors have shown that the genetic regulation of metabolic networks may follow an optimality principle such as the minimization of the transition time or the maximization of the production of a given metabolite. For example, the optimal "just-in-time" activation pattern in enzyme expression for the case of unbranched pathways has been formulated and solved as a nonlinear optimization problem with dynamic constraints [22, 47]. More recently the problem has been considered as a general DO one and was solved through the use of the Pontryagin's maximum principle for linear pathways [8, 29]. However the difficulty (or impossibility) of analytically solving other more realistic cases, such as those considering nonlinear dynamics for the enzyme expression, or other arbitrarily complex networks, calls for the use of robust numerical DO approaches.

Here we consider one of such examples and approach its solution using the DOTcvpSB toolbox (http://www.iim.csic.es/~dotcvpsb, [17]) which combines the CVP approach with global stochastic and hybrid methods to solve DO problems.

The pathway considered is depicted in Fig. 24.1. It consists of four enzymatic reactions with one branch where the products are accumulated to be consumed later.

The hypothesis is that the pathway activation minimizes the time from the substrate to the product. The activation profile may then be found by computing $\mathbf{r}_i\ (t)$ over $t \in [t_0, t_f]$ to minimize $J = t_f$ subject to the system dynamics:

$$\frac{\mathrm{d}\mathbf{S}_i}{\mathrm{d}t} = N\upsilon \qquad \frac{\mathrm{d}\mathbf{E}_i}{\mathrm{d}t} = r_i - \lambda \mathbf{E}_i \tag{24.12}$$

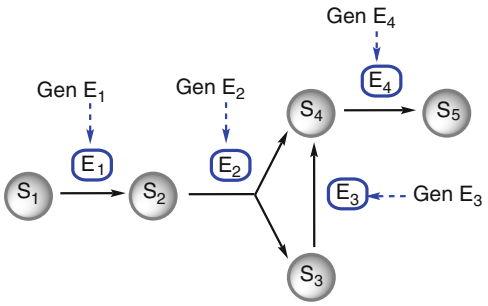


Fig. 24.1 Schematic representation of the branched pathway considered. The pathway consists of four enzymatic reactions catalyzed by a specific enzyme (E_i) where S_1 is the substrate, S_2–S_4 are intermediates and S_5 is the product. The enzyme dynamics are considered to be linear with a reaction rate r_i

where:

$$\upsilon = \frac{k_{\text{cat}}\mathbf{S}_i}{K_{\text{M}} + \mathbf{S}_i} \qquad N = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{24.13}$$

and the following end point and path constraints:

$$S_5\left(t_{\text{f}}\right) = P_{t_{\text{f}}} \qquad \sum_{i=1}^{4} \mathbf{E}_i \leq E_{\text{T}} \tag{24.14}$$

with $K_{\text{M}} = 1\text{mM}$, $k_{\text{cat}} = 1\, s^{-1}$, $P_{t_{\text{f}}} = 0.75\text{mM}$, and $\lambda = 0.5$.

The optimal activation profile corresponding to an optimal final time $t_{\text{f}} = 10.4\,\text{s}$ obtained with an evolutionary approach is presented in Fig. 24.2.

4.2 Parameter Estimation in Complex Systems Biology Models

The problem of parameter estimation in biochemical pathways, formulated as a NLP where the objective is to compute the model parameter values that maximize the fit to the experimental data, has received substantial attention [19, 28, 33]. Many difficulties found during parameter estimation are not only due to the highly nonlinear nature of the models and their size, but also due to the quality and quantity of experimental data. These result in poor practical identifiability, i.e., in the difficulty or impossibility to compute unique values for the parameters given a set of data, or the presence of suboptimal solutions.

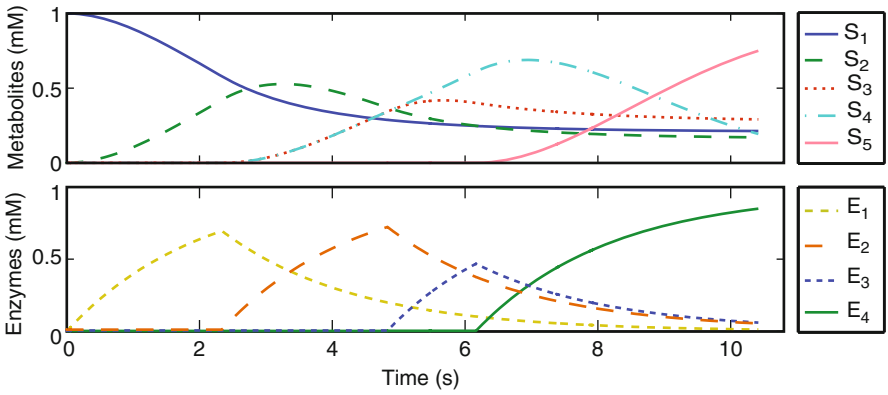


Fig. 24.2 Optimal enzyme activation profile and the corresponding metabolite dynamics. The optimal profiles for the expression rates follow a switching pattern that matches with the pathway topology leading to enzyme profiles that follow a sequential activation with protein degradation to synthesize another protein. The substrate (S_1) is converted into the product (S_5) through the intermediates (S_2, S_3, S_4), the intermediate (S_4) is accumulated and consumed in the last section of the pathway

The presence of suboptimal solutions may be tackled with global optimization methods. Recent works show how hybrid stochastic–deterministic methods handle small-to-medium size problems with reasonable computational efforts [3, 33, 34]. However, further developments are necessary to enhance, in so far is possible, the efficiency of the optimization while keeping robustness for large-scale complex biological models. Multi-thread approaches, i.e., those running several computations in parallel in different processors, seem to be the most suitable for this purpose.

Here we present a new *cooperative* strategy for the parallelization of the eSS algorithm [14]. The central idea is to run, in parallel, several *threads* of eSS, which may have different settings and/or random initializations, and exchange information among them as shown in Fig. 24.3. Taking into account the classification of cooperation schemes proposed in [43], the cooperative eSS can be described as follows:

1. There are η concurrent programs.
2. The best solution found and the eSS reference set, which contains valuable information about the diversity of solutions, are available for sharing.
3. All threads share the information.
4. The threads exchange information at a fixed time interval τ.

It should be noted that cooperation produces more than just speed-up since it can change the systemic properties of an algorithm and therefore its macroscopic behavior [43]. To illustrate this point an example related to the parameter estimation of a model describing the central carbon metabolism of *E. coli* that takes into account the enzymatic and transcriptional regulation layer [23] is considered.

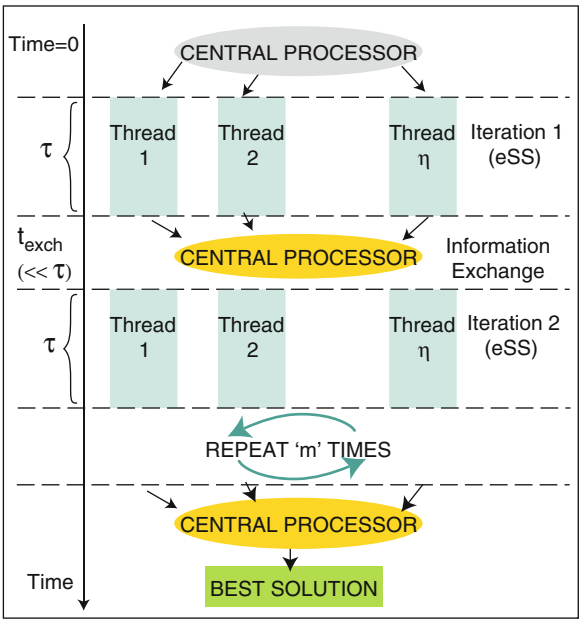


Fig. 24.3 Schematic representation of the cooperative eSS. Each of the η threads has a fixed degree of "aggressiveness". "Conservative" threads are used for increasing the probabilities of finding a feasible solution, even if the parameter space is "rugged" or weakly structured. "Aggressive" threads may speed up the calculations in "smoother" areas. Communication, which takes place at fixed time intervals, enables each thread to benefit from the knowledge gathered by the others. This knowledge includes not only information about the best solution found so far, but also about the sets of diverse parameter vectors that may be worth trying for improving the solution

The model consists of 47 nonlinear ODEs with 193 unknown parameters (affinity constants, specific activities, Hill coefficients, growth rates, expression rates, etc.). The objective is to compute those parameters so as to predict a given system behavior. Due to the stiff character of the equations and the time required for their solution, one evaluation of the least squares function takes a few seconds.

The cooperative and noncooperative multi-thread implementations of eSS are compared by launching ten threads in both cases. In the cooperative case, the ten threads exchange information as explained. In the noncooperative case, they simply run until the maximum computation time is reached. Figure 24.4 presents the corresponding convergence curves showing how the cooperative version outperforms the noncooperative one, being capable of finding a better value of the objective function while reducing computation time by 70%.

4.3 Optimal Experimental Design for Parameter Estimation

As mentioned above, poor practical identifiability has to do with the type of experimental scheme being used and the quality of the corresponding experimental

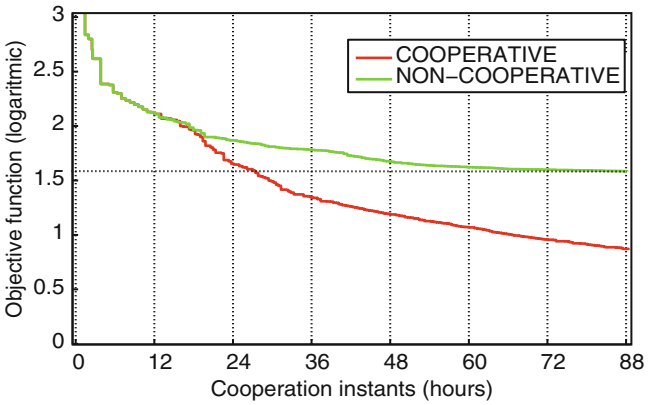


Fig. 24.4 Comparison of the performance of the parallel and cooperative eSS implementations in the solution of a large-scale parameter estimation problem. Each curve represents, at every time instant, the best value found by any of its 10 threads

data in terms of experimental noise. The purpose of optimal experimental design is to devise the necessary dynamic experiments in such a way that the parameters are estimated from the resulting experimental data with the best possible statistical quality, which is usually a measure of the accuracy and/or decorrelation of the estimated parameters. In this way, the model and a close-to-optimal solution for the parameters are being used to design new more informative experiments which in general will result in better practical identifiability properties. The information provided by the measurements is often quantified by means of the Fisher information matrix [1, 5].

AMIGO (http://www.iim.csic.es/~amigo, [2]) is a multi-platform toolbox which apart from covering model simulation, local and global sensitivity analysis, parameter estimation, and identifiability analysis, incorporates the optimal experimental design as a general DO problem.

Here we illustrate its possibilities in the context of optimal experimental design with an example related to a model describing a pathway consisting of three enzymatic steps including the enzymes and the mRNAs explicitly [28]. Previous works [28, 33, 34] considered a factorial plan consisting of 16 experiments under different amounts of substrate and product to estimate all 36 model parameters. We will consider here the case of estimating: na_2, na_3, k_1, k_2, k_3, k_4, k_6, V_1, V_2, V_3, V_5, K_5.

In a few seconds with eSS as implemented in AMIGO the global optimum is achieved corresponding to the following parameter values: $k_1 = 1.0 \pm 4.4$, $k_2 = 0.1 \pm 6.9$, $k_3 = 1.0 \pm 8.8$, $k_4 = 0.1 \pm 0.01$, $k_6 = 0.1 \pm 0.02$, $V_1 = 1.0 \pm 4.4$, $V_2 = 0.1 \pm 6.9$, $V_3 = 1.0 \pm 8.8$, $V_5 = 0.1 \pm 0.07$, $na_2 = 2.0 \pm 0.7$, $na_3 = 2.0 \pm 0.7$, $K_5 = 1.0 \pm 1.2$. Note that even though the global solution was found, the confidence regions for some of the parameters are rather large and in many cases (k_1, k_2, k_3, V_1, V_2, V_3, K_5) they are over the 100%.

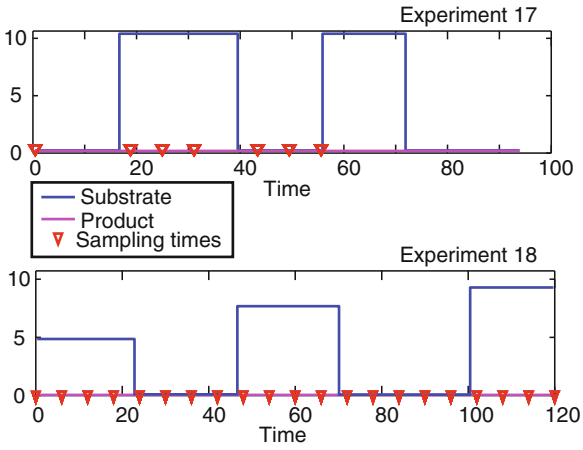


Fig. 24.5 Optimally designed experiments. For the experiment 17 the number and location of sampling times was optimally designed, resulting in a much reduced number of necessary sampling times. It is also remarkable that optimal step-wise experiment 18 results in a pulse-up type stimulation

In order to improve the practical identifiability we implemented in AMIGO the design of a parallel-sequential experimental scheme. In particular two experiments were designed under the following conditions:

- Experiment 17: pulsed stimulation of the substrate. The location and duration of the pulses as well as the number and location of sampling times and experiment duration were to be optimized.
- Experiment 18: a 5 step-wise stimulation of the substrate is allowed within the maximum and minimum values.

Even allowing for limited flexibility in the design of the experiments, results (Fig. 24.5) reveal a substantial reduction in the confidence regions for the parameters: $k_1 = 1.0 \pm 1.3(-70\%)$, $k_2 = 0.1 \pm 3.2(-54\%)$, $k_3 = 1.0 \pm 4.1(-53\%)$, $V_1 = 1.0 \pm 1.3(-70\%)$, $V_2 = 0.1 \pm 3.2(-54\%)$, $V_3 = 1.0 \pm 4.1(-53\%)$, $K_5 = 1.0 \pm 0.2(-83\%)$. Further improvements may be achieved by either allowing for further flexibility in the designs or by adding new experiments.

4.4 Stimulation Design

For some particular systems, once reliable models have been developed it is possible to design stimulation conditions so as to achieve a certain goal. In this context, it is possible, for example, to optimally design medical treatments or immune responses [12, 20], or to obtain particular behaviors such as in the case of pattern formation [24, 25, 35]. However the solution of such problems often results in the presence of suboptimal solutions [25]. Here, we illustrate with an example related to pattern formation in bacterial chemotaxis [24] how hybrid global optimization methods may successfully solve these problems.

Some types of cells are able to sense the presence of chemical signals (chemoattractants) and guide their movement in the direction of the concentration gradient of these signals. This process is called chemotaxis. The chemotaxis of the bacteria *E. coli* is one of the best understood chemotaxis processes. These bacteria, under given stress conditions, secrete chemoattractants. Other cells respond to these secreted signaling molecules by moving up their local concentration gradients and forming different types of multicellular structures.

The system may be described by a two-component diffusion reaction model:

$$\frac{\partial z}{\partial t} = D\frac{\partial^2 z}{\partial \xi} + \mu\frac{\partial}{\partial \xi}\left(\frac{z}{(1+c)^2}\frac{\partial c}{\partial \xi}\right) \qquad \frac{\partial c}{\partial t} = \frac{\partial^2 c}{\partial \xi} + \frac{z^2}{(1+z^2)} \tag{24.15}$$

with homogenous first order boundary conditions and initial conditions $z(\xi, 0) = 1$; $c(\xi, 0) = 0$, where $z(\xi, t)$ and $c(\xi, t)$ represent the cell density and the concentration of the chemoattractant, respectively.

Lebiedz and co-workers [24, 25] considered the problem of externally manipulating the process so as to achieve a particular Gaussian cell distribution. With

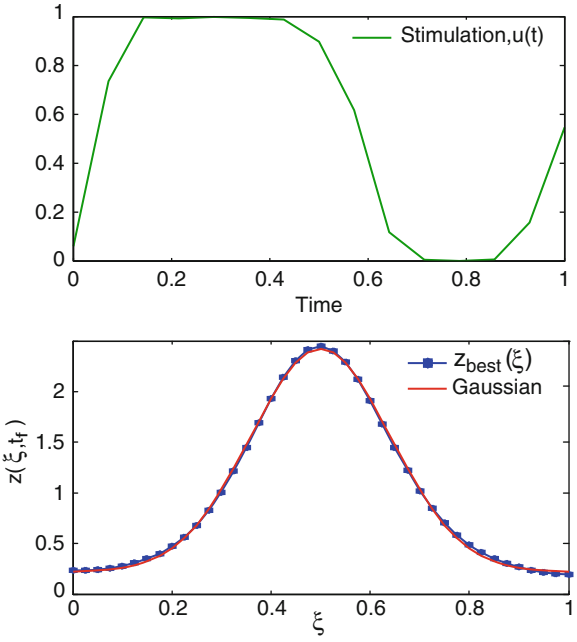


Fig. 24.6 Optimal stimulation profile and corresponding optimal behavior for the bacterial chemotaxis problem. The model in (24.15) with the corresponding boundary and initial conditions was numerically solved using the numerical method of lines with fourth order formulae and a mesh of 41 elements. In a first approximation to the DO problem the CVP approach with linear control interpolation was combined with a multi-start of a local method (i.e., the solution of the problem with a local method from 500 different initial guesses) which revealed the presence of several local optima. The global solution was found with a sequential hybrid-deterministic method [4]

this aim, a nonzero chemoattractant flux is introduced in the boundary $\frac{\partial c}{\partial \xi}(\xi = L, t) = -c(\xi = L, t) + u(t)$. The problem was formulated as DO problem where the objective is to find $u(t)$ so as to minimize the distance between the distribution of bacteria at final time ($z(\xi, t_f)$) and the desired Gaussian distribution, subject to the system dynamics (24.15) and bounds on the concentration of chemoattractant. These authors made use of the multiple shooting approach with a local optimization method to solve the problem reporting some difficulties due to the presence of local optima and the large computational cost associated. Here the problem is solved by means of the CVP approach in combination with a sequential hybrid-deterministic method [4]. The optimal solution (Fig. 24.6) was found in a few seconds.

5 Conclusions

In this work we have focused on typical optimization problems in systems biology and how their solution may be approached with novel global optimization methods and software tools developed in our group. In particular, a novel optimization approach, the multi-thread eSS, for the solution of large-scale optimization problems was presented, together with the AMIGO toolbox devoted to model identification and the DOTcvpSB toolbox devoted to DO.

As illustrative examples we considered the optimal enzyme activation in a branched metabolic pathway, the parameter estimation of a large-scale dynamic model, an optimal experimental design problem to improve identifiability, and the design of optimal stimulation conditions to achieve a given desired result in a reaction-diffusion system.

It should be noted that these software tools can be easily extended to handle multiobjetive optimization problems following the methods described in [38,39,45].

Acknowledgments This work was supported by the Spanish MICINN project "MultiSysBio" (ref. DPI2008-06880-C03-02), and by CSIC intramural project "BioREDES" (ref. PIE-201170E018).

References

1. Balsa-Canto E, Alonso AA, Banga JR (2008) Computational procedures for optimal experimental design in biological systems. IET Syst Biol 2(4):163–172
2. Balsa-Canto E, Alonso AA, Banga JR (2010) An iterative identification procedure for dynamic modeling of biochemical networks. BMC Syst Biol 4:11
3. Balsa-Canto E, Peifer M, Banga JR, Timmer J, Fleck C (2008) Hybrid optimization method with general switching strategy for parameter estimation. BMC Syst Biol 2:26
4. Balsa-Canto E, Vassiliadis VS, Banga JR (2005) Dynamic optimization of single- and multi-stage systems using a hybrid stochastic–deterministic method. Ind Eng Chem Res 44(5): 1514–1523

5. Bandara S, Sclöder J, Eils R, Bock H, Meyer T (2009) Optimal experimental design for parameter estimation of a cell signaling model. Plos Comput Biol 5(11):1–12
6. Banga JR (2008) Optimization in computational systems biology. BMC Syst Biol 2:47–53
7. Banga JR, Balsa-Canto E (2008) Parameter estimation and optimal experimental design. Essays Biochem 45:195–210
8. Bartl M, Li P, Schuster S (2010) Modelling the optimal timing in metabolic pathway activation-use of pontryagin's maximum principle and role of the golden section. Biosystems 101(1): 67–77
9. Biegler LT, Cervantes A, Wätcher A (2002) Advances in simulaneous strategies for dynamic process optimization. Chem Eng Sci 57(4):575–593
10. Bock H, Plitt K (1984) A multiple shooting algorithm for direct solution of optimal control problems. In: Proc 9th IFAC World Congress, Pergamon Press, New York, pp 242–247
11. Bryson AE, Ho YC (1975) Applied optimal control. Hemisphere Pub. Corp, New York
12. Castiglione F, Piccoli B (2007) Cancer immunotherapy, mathematical modeling, and optimal control. J Theor Biol 247(4):723–732
13. Egea JA, Balsa-Canto E, Garcia MSG, Banga JR (2009) Dynamic optimization of nonlinear processes with an enhanced scatter search method. Ind Eng Chem Res 48(9):4388–4401
14. Egea JA, Martí R, Banga JR (2010) An evolutionary method for complex-process optimization. Comp Oper Res 37(2):315–324
15. Egea JA, Rodriguez-Fernandez M, Banga JR, Marti R (2007) Scatter search for chemical and bio-process optimization. J Global Optim 37(3):481–503
16. Floudas C (2000) Deterministic global optimization: theory, methods and applications. Kluwer Academics, The Netherlands
17. Hirmajer T, Balsa-Canto E, Banga JR (2009) DOTcvpSB, a software toolbox for dynamic optimization in systems biology. BMC Bioinformatics 10:199–213
18. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) COPASI – A COmplex PAthway SImulator. Bioinformatics 22(24):3067–3074
19. Jaqaman K, Danuser G (2006) Linking data to models: data regression. Nat Rev Mol Cell Bio 7(11):813–819
20. Joly M, Pinto J (2006) Role of mathematical modeling on the optimal control of hiv-1 pathogenesis. AiChe J 52(3):856–884
21. Kauffman K, Prakash P, Edwards J (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14(5):491–496
22. Klipp E, Heinrich R, Holzhtte H (2002) Prediction of temporal gene expression. metabolic opimization by re-distribution of enzyme activities. Eur J Biochem 269:5406–5413
23. Kotte O, Zaugg J, Heinemann M (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Sys Biol 6:355
24. Lebiedz D (2005) Exploiting optimal control for target-oriented manipulation of (bio)chemical systems: A model-based approach to specific modification of self-organized dynamics. Int J Mod Phys B 19 3763–3798
25. Lebiedz D, Maurer H (2004) External optimal control of self-organisation dynamics in a chemotaxis reaction diffusion system. IEE Syst Biol 2:222–229
26. Maiwald T, Timmer J (2008) Dynamical modeling and multi-experiment fitting with PottersWheel. Bioinformatics 24(18):2037–2043
27. Mendes P, Kell D (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14(10):869–883
28. Moles CG, Mendes P, Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res 13:2467–2474
29. Oyarzun DA, Ingalls B, Middleton R, Kalamatianos D (2009) Sequential activation of metabolic pathways: a dynamic optimization approach. Bull Math Biol 71:1851–1872
30. Pardalos P, Romeijn HE, Tuyb H (2000) Recent developments and trends in global optimization. J Comp App Math 124:209–228
31. Pinter J (1996) Global optimization in action. Continuous and Lipschitz optimization: algorithms, implementations and applications. Kluwer Academics, Netherlands

32. Rateitschak K, Karger A, Fitzner B, Lange F, Wolkenhauer O, Jaster R (2010) Mathematical modelling of interferon-gamma signalling in pancreatic stellate cells reflects and predicts the dynamics of stat1 pathway activity. Cell Signal 22:97–105
33. Rodriguez-Fernandez M, Egea JA, Banga JR (2006) Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC Bioinformatics 7:483
34. Rodriguez-Fernandez M, Mendes P, Banga JR (2006) A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems 83(2–3):24
35. Salby O, Sager S, Shaik O, Kummer U, Lebiedz D (2007) Optimal control of self-organized dynamics in cellular signal transduction. Math Comp Model Dyn 13:487–502
36. Schiesser WE (1994) Computational mathematics in engineering and applied science: ODEs, DAEs, and PDEs. CRC Press, Inc., Florida, USA
37. Schmidt H, Jirstrand M (2006) Systems biology toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics 22(4):514–515
38. Sendin JOH, Exler O, Banga JR (2010) Multi-objective mixed integer strategy for the optimisation of biological networks. IET Syst Biol 4(3):236–248
39. Sendin JOH, Vera J, Torres N, Banga JR (2006) Model based optimization of biochemical systems using multiple objectives: A comparison of several solution strategies. Math Comp Mod Dyn Syst 12(5):469–487
40. Sugimoto M, Kikuchi S, Tomita M (2005) Reverse engineering of biochemical equations from time-course data by means of genetic programming. BioSystems 80:155–164
41. Sutherland W (2005) The best solution. Nature 435:569
42. Talbi EG (2009) Metaheuristics: from design to implementation. Wiley Publishing, New Jersey, USA
43. Toulouse M, Crainic T, Sansó B (2004) Systemic behavior of cooperative search algorithms. Parallel Comput 30:57–79
44. Vassiliadis VS, Pantelides CC, Sargent RWH (1994) Solution of a class of multistage dynamic optimization problems. 1. problems without path constraints. Ind Eng Chem Res 33(9): 2111–2122
45. Vera J, de Atauri P, Torres N, Banga JR (2003) Multicriteria optimization of biochemical systems by linear programming: application to production of ethanol by *Saccharomyces cerevisiae*. Biotechnol Bioeng 83(3):335–343
46. Vera J, Balsa-Canto E, Wellstead P, Banga JR, Wolkenhauer O (2007) Power-law models of signal transduction pathways. Cell Signal 19:1531–1541
47. Zaslaver A, Mayo A, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette M, Alon U (2004) Just-in-time transcription program in metabolic pathways. Nat Genet 36:486–491

Chapter 25
Mathematical Modeling of the Human Energy Metabolism Based on the Selfish Brain Theory

Matthias Chung and Britta Göbel

Abstract Deregulations in the human energy metabolism may cause diseases such as obesity and type 2 diabetes mellitus. The origins of these pathologies are fairly unknown. The key role of the brain is the regulation of the complex whole body energy metabolism. The Selfish Brain Theory identifies the priority of brain energy supply in the competition for available energy resources within the organism. Here, we review mathematical models of the human energy metabolism supporting central aspects of the Selfish Brain Theory. First, we present a dynamical system modeling the whole body energy metabolism. This model takes into account the two central control mechanisms of the brain, i.e., allocation and appetite. Moreover, we present mathematical models of regulatory subsystems. We examine a neuronal model which specifies potential elements of the brain to sense and regulate cerebral energy content. We investigate a model of the HPA system regulating the allocation of energy within the organism. Finally, we present a robust modeling approach of appetite regulation. All models account for a systemic understanding of the human energy metabolism and thus do shed light onto defects causing metabolic diseases.

1 Selfish Brain Theory

How does the human organism control its energy metabolism? The answer to this question is essential to understand disorders such as obesity and type 2 diabetes mellitus. However, the answer to the challenging question remains open.

M. Chung (✉)
Department of Mathematics, Texas State University, 601 University Drive, San Marcos, TX 78666, USA
e-mail: mc85@txstate.edu

B. Göbel (✉)
Graduate School for Computing in Medicine and Life Sciences, Institute of Mathematics and Image Computing, University of Lübeck, Maria-Goeppert-Strasse 1a, 23562 Lübeck, Germany
e-mail: goebel@mic.uni-luebeck.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_25,

The lipostatic and glucostatic theory gave a first basic theoretical understanding of control mechanisms of the human energy metabolism [20, 29]. However, studies show that these theories fail to explain many phenomena like certain metabolic diseases [18, 36]. The decisive role of the brain in the global regulation of the complex whole body energy metabolism is subject to recent research activities [28, 34, 37, 40]. To our best knowledge, the "Selfish Brain Theory" for the first time regards the brain as the central organ in the energy metabolism and as its hierarchical highest controller (a detailed review of the Selfish Brain Theory can be found in Peters et al. [37]). The goal of the Selfish Brain Theory is to understand the dynamics of the human energy metabolism in order to identify the origins of disorders such as obesity, diabetes mellitus type 2, and anorexia nervosa as well as to develop efficient therapies and treatments therefore, [32, 39, 42].

Why do we call the brain "selfish"? In contrast to other organs, the brain has unique characteristics. The blood brain barrier restricts cerebral substrate uptake. The brain almost exclusively metabolizes glucose. Compared to its size the energy consumption of the brain is very high due to neuronal activity accounting for about 25% of total body glucose utilization [8]. Despite its high energy consumption, the brain has little ability to store energy. Another key feature is the brain's plasticity. The brain is able to adapt to its environment, to learn, and to change neuronal and hormonal responses to external stimuli. Furthermore, the brain is connected to all other organs. Via afferent nerve fibers the brain receives status information from the organs and via efferent nerve fibers the brain is capable of sending control signals to the organs. The brain as subordinate controller needs to maintain its own functionality. Without energy the brain will shut down and the organism will collapse. Hence, highest priority for the brain is providing itself with sufficient energy – the brain acts "selfish".

How does the brain secure its energy supply? An overview of involved mechanisms is given in Fig. 25.1. In general, the brain has two mechanisms. First, the brain may direct available energy resources across the blood brain barrier into the brain [4, 26, 45] or reduce energy uptake of peripheral organs [10, 12, 19] (*allocation*). Secondly, ingestion of energy increases the total amount of energy in the body (*appetite*) [30, 41]. We conclude, regulating appetite and changing the allocation of energy are the main control mechanisms of the brain.

What are key control mechanisms to maintain cerebral energy supply? The Selfish Brain Theory identifies the "principle of biological homeostasis" as central control mechanism [9]. This mechanism can be found in various homeostatic systems throughout the body [5]. The principle of biological homeostasis bases on interacting negative and positive feedback signals naturally stabilizing the controlled target substrate:

1. The ligand A binds with high affinity to a receptor B but with low affinity to a receptor C.
2. The two ligand-bound receptors B and C act in opposing manners.
3. Ligand-bound receptors B increase the concentration of A (positive feedback) while ligand-bound receptors C decrease the concentration of A (negative feedback).

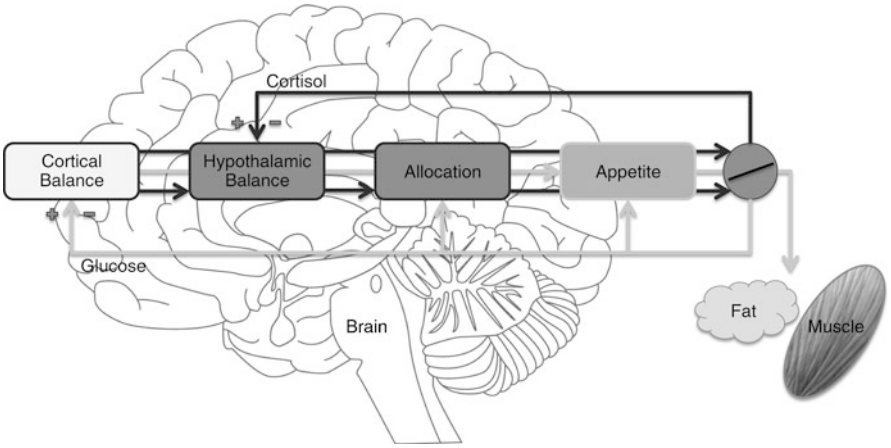


Fig. 25.1 The energy metabolism according to the Selfish Brain Theory. The brain has two principal mechanisms to secure its energy supply: allocation and appetite. The neocortex as well as the HPA system govern these two mechanisms. Feedback signals from cerebral energy content (glucose) and activity of the HPA system (cortisol) affect the hierarchical elements of the system. The model reflects the priority of cerebral energy supply while brain and body periphery compete for the available energy resources

These interactions lead to a homeostatic state of molecule A [9, 34]. They result in *cortical* and *hypothalamic balance* which govern allocation and appetite.

Mathematical models play a key role in theory validation. There exist many models describing the human energy metabolism, e.g., [1, 2, 6, 16, 25, 27]. These models base on the classical glucostatic and lipostatic theory, respectively. Recently, several mathematical models of the human energy metabolism have been developed and investigated to analyze the numerous findings and complex mechanisms of the Selfish Brain Theory [7,9,15,22,23,35]. These models take into account the decisive role of the brain with respect to the energy metabolism. That is, it is considered as superior regulatory instance and as energy consumer. The analytic investigations and numerical simulations do shed light onto the global regulation of the complex whole body energy metabolism. Moreover, there exist detailed modeling approaches of the brain energy metabolism [11, 46].

In this work, we review and discuss mathematical models of the human energy metabolism supporting the Selfish Brain Theory. First, we present a whole body model focusing on the main regulatory control mechanisms of the brain, i.e., allocation and ingestion (Sect. 2.1) [14, 15]. This compact dynamical system aims at understanding the basic mechanisms. The models in the following subsections can be seen as submodels focusing on details of the energy metabolism presented in Sect. 2.1. In Sect. 2.2, we present a neuronal model which identifies a potential cerebral mechanism to sense the brain's energy content and to send signals to the body [7]. This section is followed by a mathematical model of the HPA system which can be seen as allocation mechanism of the brain (Sect. 2.3) [9]. The last mathematical model in our consideration in Sect. 2.4 focuses on appetite regulation [13]. We close with a short discussion and conclusion in Sect. 3.

2 Mathematical Modeling

2.1 Brain Centered Energy Metabolism Model

A mathematical model describing the human energy metabolism on a whole body scale has been developed and investigated in Göbel et al. [15]. This model includes principal elements of the Selfish Brain Theory. The model takes into account the two central roles of the brain in the energy metabolism. The brain is considered as subordinate regulatory authority as well as the consumer with highest priority.

The model consists of separated compartments containing energy metabolites. It regards the time-dependent brain energy content $A = A(t) \in \mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$, which might be identified as cerebral adenosine triphosphate (ATP) concentration. The energy level in the blood is $G = G(t) \in \mathbb{R}_+$ consisting mostly of glucose. The energy resources in the body periphery $R = R(t) \in \mathbb{R}_+$ comprise all energy reserves like muscle, fat tissue, liver, and gastrointestinal tract.

The mathematical model integrates energy fluxes between these compartments and control signals directing the energy fluxes within the organism, see Fig. 25.2. Both mechanisms supplying the brain with adequate energy amounts are included in the model: allocation and appetite. First, the allocation mechanism of the brain is represented in the production of the control signal insulin $I = I(t) \in \mathbb{R}_+$. The brain may supply itself with energy by dropping the insulin level. In this way, the

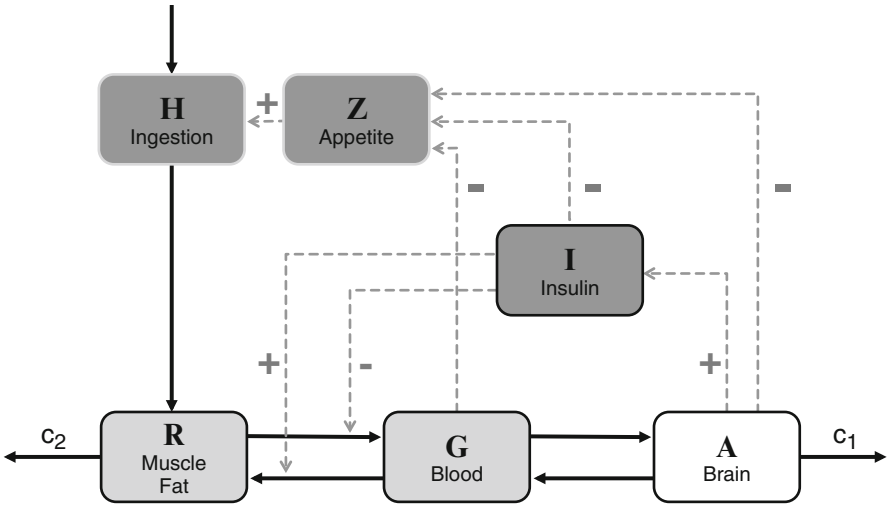


Fig. 25.2 Energy fluxes between compartments (*solid*) and control signals directing the energy fluxes in the organism (*dashed*). The ingested energy H passes the resources R and the blood G and is transported into the brain A. Energy is consumed by the brain (c_1) and the periphery (c_2). While the appetite Z affected by A, G, and I controls the ingestion H, insulin I controls the allocation of energy to the brain via control of the blood glucose flux

insulin-dependent glucose flux into the body periphery is suppressed. The available blood glucose is assimilated into the brain because the glucose flux across the blood brain barrier is insulin-independent. Regarding the hormone insulin not only as local signal but as central feedback signal of the brain is a central new aspect in the model. Secondly, the regulation of appetite $Z = Z(t) \in \mathbb{R}_+$ and the ingestion of energy $H = H(t) \in \mathbb{R}_+$ are included. The brain as well as the body periphery consume energy represented by $c_1 = c_1(t) \in \mathbb{R}_+$ and $c_2 = c_2(t) \in \mathbb{R}_+$, respectively.

The model of the whole body energy metabolism is given by the system of five ordinary differential equations:

$$\begin{aligned}
\dot{A} &= p_1 \frac{G}{A} - c_1, \\
\dot{G} &= -p_1 \frac{G}{A} - p_2 G I + p_3 \frac{R}{G}, \\
\dot{I} &= p_4 A - p_5 I, \\
\dot{R} &= p_2 G I - p_3 \frac{R}{G} + p_6 H - c_2, \\
\dot{H} &= p_7 \left(f\,(Z) - H \right) \\
\text{with } Z &= \frac{p_8}{AGI}
\end{aligned} \tag{25.1}$$

and with positive parameters $p_1, \ldots, p_8$. All parameters have a physiological interpretation. For instance, the sigmoid function f represents ingestion activation depending on appetite.

It is well known that ingestion is mildly regulated on a short-time scale [38, 43]. In contrast, on a long-time scale extending over months or years we observe an extremely strict food intake regulation [47]. In the presented dynamical system, the parameter p_7 reflects the sensitivity of the organism to food intake consistent with its need for energy. A low value of p_7 indicates a slow adaption to the body's energy needs. However, a rather high value of p_7 reflects that ingestion is strongly regulated and the energy uptake immediately satisfies the needs of the organism. The transition $p_7 \to \infty$ in system (25.1) leads to a lower-dimensional dynamical system describing the mean regulation of food intake. It is the related long-term model of the energy metabolism giving insight into the long-term behavior of the model.

The presented model realistically describes the qualitative and quantitative behavior of the human whole body energy metabolism even for a large class of physiological interventions. Short-time observations demonstrate the physiological periodic ingestive behavior generating the circadian blood glucose and insulin oscillations, see solid lines in Fig. 25.3. However, if ingestion activation is not sensitively dependent on appetite and thereby energy deprivation, we observe permanent feeding and fixed energy levels, see dashed lines in Fig. 25.3. A stable

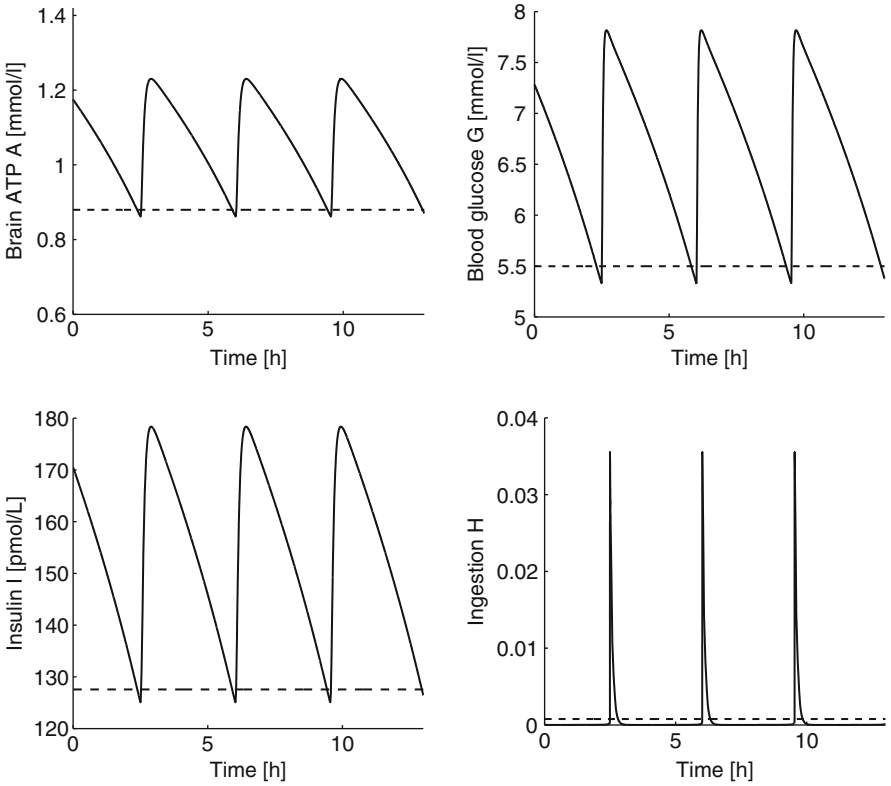


Fig. 25.3 Simulation results of model (25.1) for sensitive activation of food intake (*solid*) and moderate sensitivity of ingestion activation (*dashed*)

long-term behavior in accordance with the homeostatic regulation of the energy metabolism on a long-time scale can be observed. The model properties have been analyzed in detail in [14, 15]. It can be shown that the long-term model has an asymptotically stable stationary point, whereas the short-term model features stable limit cycles. Therefore, the analytic results are in line with the numerical calculations.

The presented mathematical model of the human whole body energy metabolism reproduces central aspects of the Selfish Brain Theory. Key elements like the preeminence of the brain's energy supply as well as the competition for available energy resources between brain and body periphery are reflected in the obtained results. In the following sections, we introduce mathematical models of regulatory subsystems of the energy homeostasis, namely cortical balance, HPA system (allocation), and appetite regulation, compare Fig. 25.1.

2.2 Neuronal Model

The cortical balance of brain energy supply is essential to maintain the brain's functionality. Hence, a key question arises in understanding the regulation of the energy metabolism: How does the brain sense its energy level and how is the regulatory feedback generated?

The principle of biological homeostasis generates the cortical balance of the energy metabolite ATP via interaction of ATP sensitive potassium (KATP) channels. On the one hand, ATP binds at low concentrations to high affine KATP channels predominantly distributed on excitatory dopaminergic neurons. On the other hand, ATP binds at high concentrations to low affine KATP channels mainly located on inhibitory GABAergic neurons. Activated KATP channels close and potassium ions are prevented to cross the cell membrane. Thereby, the neuron fires and releases its neurotransmitter. Action potentials of GABAergic neurons inhibit dopaminergic neurons via $GABA_A$ receptors which are located at the postsynaptic membrane of dopaminergic neurons. The neurotransmitter dopamine is prevented to be released. Maximum dopamine outflow can be seen as feedback signal at the homeostatic ATP level while low dopamine signals correspond to unbalanced brain energy content [37].

Mathematical modeling of these interactions must incorporate coupled GABAergic and dopaminergic neurons as well as the relevant ATP sensitive ion channels, see Fig. 25.4. Mathematical models of brain activity need to consider neuronal models with their signaling, interactions, and patterns. The Hodgkin–Huxley model introduced in 1952 [17] is the standard type model to simulate neuronal activity of a single neuron. This model is given by a system of four ordinary differential equations for each neuron. In [7], a coupled Hodgkin–Huxley model of GABAergic and dopaminergic neurons is used to investigate the neurological plausibility of the described mechanisms. The differential equations:

$$\begin{aligned} \dot{V}^{\text{dopa}} &= \frac{1}{C}\left(-I_{\text{Na}}^{\text{dopa}} - I_{\text{K}}^{\text{dopa}} - I_{\text{L}}^{\text{dopa}} - I^{\text{GABA}} + I_{\text{P}}\right), \\ \dot{V}^{\text{GABA}} &= \frac{1}{C}\left(-I_{\text{Na}}^{\text{GABA}} - I_{\text{K}}^{\text{GABA}} - I_{\text{L}}^{\text{GABA}} + I_{\text{P}}\right) \end{aligned} \tag{25.2}$$

describe the membrane potential V of dopaminergic and GABAergic neurons, respectively. The constant C is the membrane capacitance, and I_{P} is an externally applied current. The currents of sodium and potassium ion channels are represented by I_{Na}, I_{K}, while I_{L} is a leakage current. These currents are described by membrane potential-dependent gating variables in the Hodgkin–Huxley model.

The basic Hodgkin–Huxley model is expanded by the synaptic coupling of GABAergic and dopaminergic neurons as well as by the interaction of high and low affine KATP channels in [7]. The inhibitory effect of GABAergic neurons on

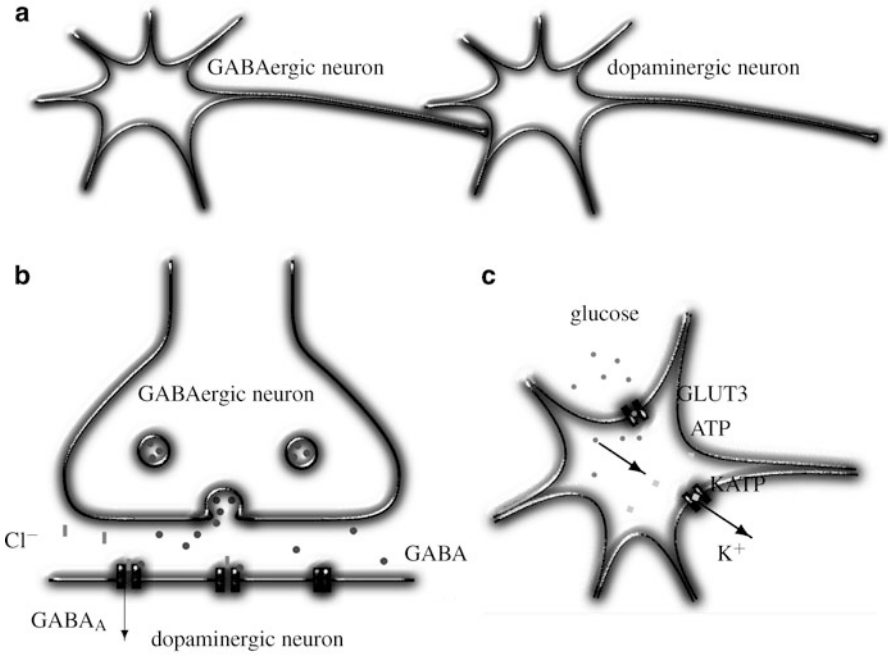


Fig. 25.4 (**a**) Schematic illustration of the interaction between GABAergic and dopaminergic neurons. (**b**) GABA (black dots) is released into the synaptic cleft between the presynaptic GABAergic and the postsynaptic dopaminergic neuron. Activated $GABA_A$ receptors open and Cl^- ions (gray bars) pass into the dopaminergic neuron leading to a hyperpolarization. An action potential cannot be generated. (**c**) Insulin-independent glucose receptors GLUT3 are permeable for glucose (black dots) predominantly accounting for neuronal glucose uptake. Through glycolysis and respiratory chain, glucose is decomposed into ATP (gray squares). In turn, ATP binds to KATP channels closing these channels for K^+ ions and action potentials become more likely

dopaminergic neurons due to $GABA_A$ receptors is represented by the current I^{GABA} in equation (25.2), see Fig. 25.4b. The potassium current I_K is modified since the closing probability of KATP channels is ATP dependent, see Fig. 25.4c.

The developed mathematical model realistically reproduces the results of an in vitro experiment in which a biphasic dopamine release under varying extracellular glucose concentration is shown [44]. The simulations show a biphasic dopamine release with low activity at low and high glucose levels and an elevated activity at moderate glucose levels. The Selfish Brain Theory specifies KATP channels to be involved in cerebral energy supply sensing [37]. According to this theory, the biphasic dopamine release can be explained by the dynamics of high and low affine KATP channels, and the biphasic release might be interpreted as an energy sensing mechanism of the brain. The obtained results support the plausibility of interacting high and low affine KATP channels controlling dopamine and GABA outflow [7]. The opposing effects of the channels may generate a homeostatic regulation of

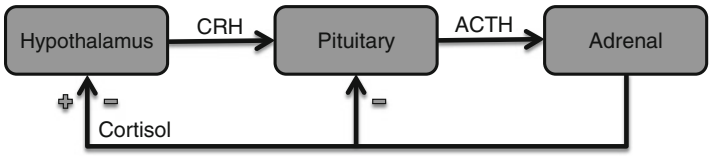


Fig. 25.5 CRH in the brain stimulates ACTH in the hypothalamus. In turn, ACTH is released in the blood and stimulates the adrenals to produce cortisol. Cortisol crosses the blood brain barrier and alters CRH as well as ACTH release. The HPA system is a closed loop feedback system

cerebral ATP concentration consistent with the proposed principle of biological homeostasis. Note, the homeostatic feedback on the cerebral ATP is conjectured by [37] and supported by references therein.

The link between the neuronal model and the whole body model (25.1) lies in the description of the regulation of the energy supply of the organism f. Our neuronal model represents the biphasic release of the neurotransmitter dopamine depending on energy content. Integration of the biphasic dopamine release corresponds to the sigmoid dynamics of the function $f(p_8/(AGI))$ in our whole body model (25.1) regulating the energy supply of the organism according to its needs.

2.3 Model of the HPA System

Hypothalamic–pituitary–adrenal (HPA) dynamics are closely related to the energy metabolism. Deregulation of cortisol may cause depression, diabetes, and visceral obesity [34]. The Selfish Brain Theory identifies the HPA system as central control mechanism of energy allocation [34].

Model (25.1) includes the allocation mechanism in terms of the control signal insulin I, mainly regulated by the brain energy content A. The HPA system shows similar central signaling behavior as insulin [37]. To investigate the allocation mechanism in more detail we turn to a mathematical modeling of the HPA system.

Activation of hypothalamic neurons causes a release of corticotropin releasing hormone (CRH). In turn, CRH is secreted into the pituitary, where it subsequently stimulates the release of adrenocorticotropic hormone (ACTH) into the blood circulation. ACTH stimulates the release of cortisol in the adrenal cortexes into the blood stream (Fig. 25.5). Serum cortisol concentration has to be sufficiently regulated within a physiological range. The HPA system is a closed loop control system since cortisol reaches all areas of the brain and stimulates/inhibits via high affine mineralocorticoid (MR) and low affine glucocorticoid receptors (GR), respectively. The stimulation/inhibition of CRH via glutamatergic pathways closes the loop.

Inhibition of cortisol secretion is an essential component of the regulation within this system. However, maintaining cortisol concentrations above a critical threshold is vital since low cortisol may result in pathological conditions.

The homeostatic regulation of cortisol follows the principle of biological homeostasis. Cortisol binds at low concentrations to MRs and only at high concentrations to GRs. Activated MR and GR operate in an opposing manner. Cortisol raises its own serum concentration via activated MRs, while the hormone reduces its concentration via activated GRs.

The dynamics of the HPA system can now be described by two coupled differential equations incorporating the principle of biological homeostasis:

$$\begin{aligned} \dot{y} &= -b_1 y + \frac{e_1 z}{z + K_1} - \frac{e_2 z}{z + K_2}, \\ \dot{z} &= -b_2 z + b_3 y. \end{aligned} \tag{25.3}$$

Here, CRH and ACTH are pooled from different brain regions in one molecular cue of the brain/pituitary compartment with a physiological interpretation as plasma ACTH concentration $y = y(t) \in \mathbb{R}_+$. The concentration of cortisol is given by $z = z(t) \in \mathbb{R}_+$. The change in cortisol concentration at time t is determined by a natural decay rate b_2 and a stimulus from ACTH with a rate b_3. The feedback on the ACTH compartment is formed by a decay rate b_1 combined with the feedback from cortisol via activated MR and GR. This feedback is modeled by nonlinear biochemical receptor kinetics. The coefficients e_1 and e_2 represent the integrated maximal efficacies, and K_1 and K_2 are the binding affinities of all MRs and GRs in various brain regions, respectively.

It can be shown that the system of ordinary differential equations (25.3) has one positive asymptotically stable stationary point if the condition

$$\frac{e_1}{K_1} > \frac{b_1 b_2}{b_3} + \frac{e_2}{K_2}$$

holds [9]. This is a weak assumption. It only states that the stimulating feedback via MR with e_1, K_1 dominates the inhibitory feedback via GR with e_2, K_2 plus the ratio of inhibitory and excitatory first order terms b_1, b_2, and b_3.

Simulations show that this dynamical system is tightly regulated and tends quickly to its stationary point (see Fig. 25.6). Compared to other HPA models this model does not depend on peripheral signals to generate a stationary point. It is basically generated by feedback signals via MR and GR from the brain and can be seen as a central allocation feedback signal. Comparisons with clinical trials show the validity of the mathematical model [9]. Note, the dynamics of this model are mainly generated by the principle of biological homeostasis. The balance is primarily induced by the interaction of receptors located in various brain regions. The association of the HPA system with the energy allocation mechanism results in a novel view of deregulations in the energy metabolism. Hence, deregulations may originate in the brain and medical conditions such as obesity and type 2 diabetes mellitus can be seen as “brain disease”.

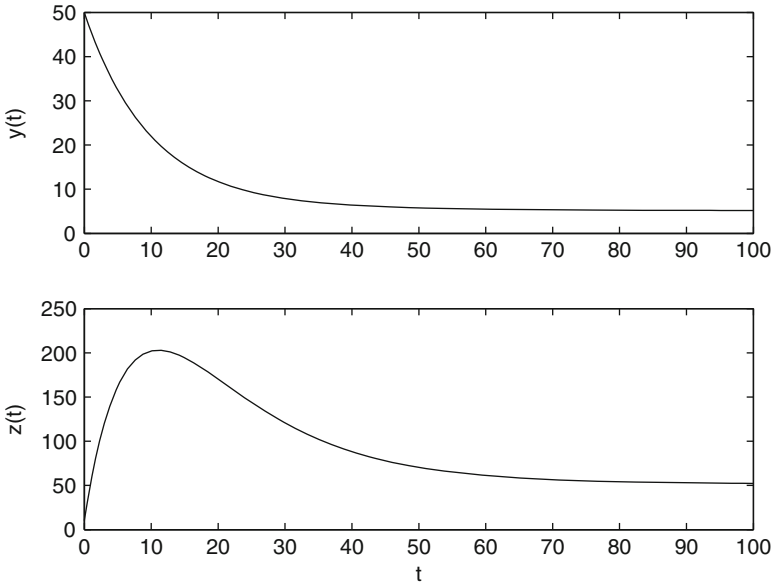


Fig. 25.6 Simulation of the HPA model. After an initial deflection the system quickly tends to the stationary point

2.4 Appetite-Regulation Model

A central part of the Selfish Brain Theory is the appetite regulation in the lateral hypothalamus [3,30,31,41], see Fig. 25.1. As discussed above, food intake is one of two mechanisms to supply the brain with adequate energy amounts. In this section, we present mathematical models that investigate the control of appetite. These models analyze appetite regulation in detail which is represented by $p_8/(AGI)$ in our whole body model (25.1).

Especially in medical applications, quantitative specifications of the numerous metabolites and complex regulatory pathways are unknown, incomplete, or incompletely understood. Our modeling approach features characteristics in order to handle the named difficulties. In [13], a robust modeling approach is chosen to describe appetite regulation in the lateral hypothalamus. Robust modeling approaches merely use reliable information and formulate results which are valid for a large class of models including all feasible realizations of unknown mechanisms [24]. Abstract compartments with their energy contents and general influence functions describing energy fluxes and regulatory pathways are used. The general influence functions are only constrained by two weak, physiologically reasonable assumptions. First, they are monotonous. In the majority of cases, it is known whether a mechanism is governed by the supplier or by the receiver. Secondly, the functions are saturated assuming that saturation is the regular case in physiological

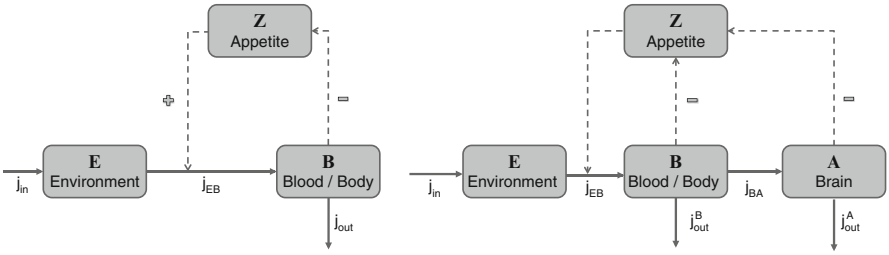


Fig. 25.7 Small three-compartment (*left*) and expanded four-compartment appetite-regulation model including the brain (*right*). Food intake is regulated by appetite in the lateral hypothalamus which is activated by lacking energy in body periphery and brain

systems. The influence functions are not further specified so that the obtained results remain valid for a variety of possible interactions. The obtained class of models is analytically investigated [21,33] with respect to circadian periodicity of human food intake.

First, a three-compartment model is regarded consisting of an environmental, a body and an appetite compartment, see Fig. 25.7(left). Time-dependent total energy contents $E = E(t) \in \mathbb{R}_+$ in the environment and $B = B(t) \in \mathbb{R}_+$ in the blood or body are considered. The compartment B could be regarded as consolidation of the compartments G and R in our whole body model (25.1). The appetite activation is described by $Z = Z(t) \in \mathbb{R}_+$ which is regulated by the energy level in the body. It can be shown that all possible interactions lead to a dynamical system with an unique, asymptotically stable stationary point. Hence, this model describes permanent feeding and does not reproduce the circadian periodicity of food intake. Therefore, it is not an adequate appetite-regulation model.

The three-compartment model is expanded by a brain compartment with the energy level $A = A(t) \in \mathbb{R}_+$, see Fig. 25.7(right). The brain is regarded as regulatory instance of appetite activation and as energy consumer.

The influx into the near environment is described by the saturated, monotonously decreasing function $j_{\text{in}} = j_{\text{in}}(E)$. Energy is consumed by the body $j^{\text{B}}_{\text{out}} = j^{\text{B}}_{\text{out}}(t)$ as well as by the brain $j^{\text{A}}_{\text{out}} = j^{\text{A}}_{\text{out}}(t)$. Ingestion is represented by the flux $j_{\text{EB}} = j_{\text{EB}}(E, Z)$ from the environment E into the body B. Note, this flux corresponds to the function H in model (25.1). The flux j_{EB} is represented by a monotonously increasing function since appetite Z as well as increased offer in the environment E enhance the flux. The energy flux $j_{\text{BY}} = j_{\text{BY}}(B, A)$ from the body to the brain compartment increases with B and decreases with A. The activation of the appetite is described by the saturated, monotonously decreasing function $u = u(B, A)$. Physiologically, appetite activation is suppressed at high energy levels in the body and in the brain. Appetite is assumed to be a subject of self-inhibition. Energy conservation gives the dynamical system:

$$\begin{aligned}\dot{E} &= j_{\text{in}}(E) - j_{\text{EB}}(E, Z),\\ \dot{B} &= j_{\text{EB}}(E, Z) - j_{\text{BY}}(B, A) - j_{\text{out}}^{\text{B}},\\ \dot{A} &= j_{\text{BY}}(B, A) - j_{\text{out}}^{\text{A}},\\ \dot{Z} &= \beta\,(u(B, A) - Z)\end{aligned} \tag{25.4}$$

with $\beta > 0$ reflecting the appetite sensitivity.

The four-compartment model (25.4) including the brain describes circadian periodic ingestive behavior in case of a sensitive cerebral appetite activation. Therefore, the indispensable role of the brain in the regulation of appetite and food intake could be verified supporting central ideas of the Selfish Brain Theory. This supports our simulation results in Sect. 2.1 where the need for a sensitive ingestion activation was shown in order to model the physiological periodicity of food intake, compare Fig. 25.3.

This robust modeling approach generalizes former mathematical models of appetite regulation, e.g., [23]. This work presents a simple mathematical model using almost exclusively linear functions to describe appetite regulation and energy transport from the body into the brain. The model also demonstrates the important role of the human brain in appetite regulation and periodic food intake as well as the dependence of appetite activation on cerebral energy supply.

3 Discussion and Conclusion

Mathematical modeling provides a key tool to investigate and validate theoretical formulations such as the Selfish Brain Theory. In this review paper, we discussed mathematical models of the human energy metabolism considering the decisive role of the brain. In Sect. 2.1, we introduced a compact model describing the whole body energy metabolism [15]. Sections 2.2–2.4 focused on the mathematical descriptions of regulatory subsystems within the Selfish Brain Theory, namely, the cortical balance [7], the HPA system controlling the allocation mechanism [9], and the appetite regulation in the lateral hypothalamus [13]. Therefore, the given models form a theoretical representation of the Selfish Brain Theory. The systemic analysis of the mathematical models allows to understand the qualitative and even the quantitative behavior of the energy metabolism. The presented models support the theoretical formulations given in Sect. 1.

The Selfish Brain Theory states that diseases such as obesity and type 2 diabetes mellitus might originate in deregulations in the brain [36]. The presented mathematical models give new evidence that the origins of these pathologies are indeed brain centered. The analysis of the presented new models provides the opportunity to design new strategies to target diseases caused by deregulations in the energy metabolism and to develop effective therapies.

The brain centered energy metabolism model in Sect. 2.1 represents a compact whole body model. In future work, mathematical models of the regulatory subsystems in Sects. 2.2–2.4 need to be integrated to form a comprehensive whole body model of the energy metabolism.

Acknowledgements The authors thank Kerstin M. Oltmanns for her invaluable physiological expertise in the development of the presented models. We also thank the reviewer for his/her thorough review and highly appreciate the comments and suggestions, which significantly contributed to improving the quality of the publication. This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by the German Research Foundation [DFG GSC 235/1].

References

1. Ackerman E (1964) A mathematical model of the glucose-tolerance test. Phys Med Biol 9(2):203–213
2. Bergman R, Phillips L, Cobelli C (1981) Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. J Clin Invest 68(6):1456–1467
3. Berthoud HR (2004) Mind versus metabolism in the control of food intake and energy balance. Physiol Behav 81(5):781–793
4. Brown A, Ransom B (2007) Astrocyte glycogen and brain energy metabolism. Glia 55: 1263–1271
5. Calabrese E, Baldwin L (2003) Toxicology rethinks its central belief. Nature 421(6924): 691–692
6. Chow C, Hall K (2008) The dynamics of human body weight change. PLoS Comput Biol 4(3):1–11
7. Chung M, Göbel B, Peters A, Oltmanns K, Moser A (2011) Mathematical modeling of the biphasic dopaminergic response to glucose. J Biomed Sci Eng 4:36–145
8. Clark D, Sokoloff L (1999) Basic neurochemistry: molecular, cellular and medical aspects. Lippincott Williams & Wilkins, Philadelphia
9. Conrad M, Hubold C, Fischer B, Peters A (2009) Modeling the hypothalamus–pituitary–adrenal system: homeostasis by interacting positive and negative feedback. J Biol Phys 35:149–162
10. Dunning B, Ahrén B, Veith R, Taborsky G (1988) Nonadrenergic sympathetic neural influences on basal pancreatic hormone secretion. Am J Physiol 255:E785–E792
11. Gaohua L, Kimura H (2009) A mathematical model of brain glucose homeostasis. Theor Biol Med Model 6(26):1–24
12. Gerendai I, Halász B (2000) Central nervous system structures connected with the endocrine glands. findings obtained with the viral transneuronal tracing technique. Exp Clin Endocrinol Diabetes 108(6):389–395
13. Göbel B, Chung M, Oltmanns K, Peters A, Langemann D (2011) Robust modeling of appetite regulation. J Theor Biol 291:65–75
14. Göbel B, Langemann D (2011) Systemic investigation of a brain-centered model of the human energy metabolism. Theory Biosci 130(1):5–18
15. Göbel B, Langemann D, Oltmanns K, Chung M (2010) Compact energy metabolism model: Brain controlled energy supply. J Theor Biol 264:1214–1224
16. Hall K (2006) Computational model of in vivo human energy metabolism during semistarvation and refeeding. Am J Physiol Endocrinol Metab 291(1):E23–E37

17. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol (Lond) 117(4):500–544
18. van Itallie T (1990) The glucostatic theory 1953–1988: roots and branches. Int J Obesity 14: 1–10
19. Jansen A, Hoffman J, Loewy A (1997) CNS sites involved in sympathetic and parasympathetic control of the pancreas: a viral tracing study. Brain Res 766:29–38
20. Kennedy G (1953) The role of depot fat in the hypothalamic control of food intake in the rat. Proc Royal Soc Lond B Biol Sci 140(901):578–592
21. Khalil H (2002) Nonlinear systems, 3 edn. Pearson Higher Education, Prentice Hall, Upper Saddle River
22. Langemann D (2007) Selfish brain theory: mathematical challenges in the top–down analysis of metabolic supply chains. Grundy, J. (Ed.) Proc. Tutorials, posters, panels and industrial contributions at the 26th Int Conf on conceptual modeling – ER 2007 Auckland, New Zealand, CRPIT 83:39–49
23. Langemann D, Peters A (2008) Deductive functional assignment of elements in appetite regulation. J Biol Phys 34:413–424
24. Liu D, Michel A (1994) Dynamical systems with saturation nonlinearities: analysis and design. Springer, New York
25. Liu W, Tang F (2008) Modeling a simplified regulatory system of blood glucose at molecular levels. J Theor Biol 252(4):608–620
26. Magistretti PJ, Pellerin L, Rothman DL, Shulman RG (1999) Energy on demand. Science 283(5401):496–497
27. Man CD, Rizza R, Cobelli C (2007) Meal simulation model of the glucose–insulin system. IEEE Trans Biomed Eng 54(10):1740–1749
28. Marty N, Dallaporta M, Thorens B (2007) Brain glucose sensing, counterregulation, and energy homeostasis. Physiology 22:241–251
29. Mayer J (1953) Glucostatic mechanism of regulation of food intake. N Engl J Med 249(1): 13–16
30. Morton GJ, Cummings DE, Baskin DG, Barsh GS, Schwartz MW (2006) Central nervous system control of food intake and body weight. Nature 443(7109):289–295
31. Neary N, Goldstone A, Bloom S (2004) Appetite regulation: from the gut to the hypothalamus. Clin Endocrino 60:153–160
32. Oltmanns K, Melchert U, Scholand-Engler H, Howitz M, Schultes B, Schweiger U, Hohagen F, Born J, Peters A, Pellerin L (2008) Differential energetic response of brain vs. skeletal muscle upon glycemic variations in healthy humans. Am J Physiol Regul Integr Comp Physiol 294(1):R12–R16
33. Perko L (2001) Differential equations and dynamical systems, 3 edn. Springer, New York
34. Peters A, Conrad M, Hubold C, Schweiger U, Fischer B, Fehm H.L (2007) The principle of homeostasis in the hypothalamus–pituitary–adrenal system: new insight from positive feedback. Am J Physiol Regul Integr Comp Physiol 293(1):R83–R98
35. Peters A, Langemann D (2009) Build-ups in the supply chain of the brain: on the neuroenergetic cause of obesity and type 2 diabetes mellitus. Front Neuroenerg 1, Art. 2:1–15
36. Peters A, Pellerin L, Dallman MF, Oltmanns KM, Schweiger U, Born J, Fehm HL (2007) Causes of obesity: looking beyond the hypothalamus. Prog Neurobiol 81(2):61–88
37. Peters A, Schweiger U, Pellerin L, Hubold C, Oltmanns KM, Conrad M, Schultes B, Born J, Fehm HL (2004) The selfish brain: competition for energy resources. Neurosci Biobehav Rev 28(2):143–180
38. Rumpler W, Kramer M, Rhodes D, Paul D (2006) The impact of the covert manipulation of macronutrient intake on energy intake and the variability in daily food intake in nonobese men. Int J Obes 30:774–781
39. Schmoller A, Hass T, Strugovshchikova O, Melchert U, Scholand-Engler H, Peters A, Schweiger U, Hohagen F, Oltmanns K (2010) Evidence for a relationship between body mass and energy metabolism in the human brain. J Cereb Blood Flow Metab 30(7):1403–1410
40. Schwartz MW, Porte D Jr (2005) Diabetes, obesity, and the brain. Science 307:375–379

41. Schwartz MW, Woods S, Porte D, Seeley R, Baskin D (2000) Central nervous system control of food intake. Nature 404:661–671
42. Schweiger U, Greggersen W, Rudolf S, Pusch M, Menzel T, Winn S, Hassfurth J, Fassbinder E, Kahl K, Oltmanns K, Hohagen F, Peters A (2008) Disturbed glucose disposal in patients with major depression; application of the glucose clamp technique. Psychosom Med 70(2):170–176
43. Stanley S, Wynne K, McGowan B, Bloom S (2005) Hormonal regulation of food intake. Physiol Rev 85:1131–1158
44. Steinkamp M, Li T, Fuellgraf H, Moser A (2007) K(ATP)-dependent neurotransmitter release in the neuronal network of the rat caudate nucleus. Neurochem Int 50(1):159–163
45. Vannucci SJ, Maher F, Simpson IA (1997) Glucose transporter proteins in brain: delivery of glucose to neurons and glia. Glia 21(1):2–21
46. Vatov L, Kizner Z, Ruppin E, Meilin S, Manor T, Mayevsky A (2006) Modeling brain energy metabolism and function: a multiparametric monitoring approach. Bull Math Biol 68(2): 275–291
47. Westerterp K, Donkers J, Fredrix E, Boekhoudt P (1995) Energy intake, physical activity and body weight: a simulation model. Br J Nutr 73:337–347

Chapter 26
Identification of Sensitive Enzymes in the Photosynthetic Carbon Metabolism

Renato Umeton, Giovanni Stracquadanio, Alessio Papini, Jole Costanza, Pietro Liò, and Giuseppe Nicosia

Abstract Understanding and optimizing the CO_2 fixation process would allow human beings to address better current energy and biotechnology issues. We focused on modeling the C_3 photosynthetic Carbon metabolism pathway with the aim of identifying the minimal set of enzymes whose biotechnological alteration could allow a functional re-engineering of the pathway. To achieve this result we merged in a single powerful pipe-line Sensitivity Analysis (SA), Single- (SO) and Multi-Objective Optimization (MO), and Robustness Analysis (RA). By using our recently developed multipurpose optimization algorithms (PAO and PMO2) here we extend our work exploring a large combinatorial solution space and most importantly, here we present an important reduction of the problem search space. From the initial number of 23 enzymes we have identified 11 enzymes whose targeting

R. Umeton (✉)
Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: umeton@mit.edu

G. Stracquadanio
The Johns Hopkins University, 217 Clark Hall, Baltimore, MD 21218, USA
e-mail: stracquadanio@jhu.edu

A. Papini
University of Florence, Via La Pira, 4, Firenze, FI I-50121, Italy
e-mail: alpapini@unifi.it

J. Costanza • G. Nicosia
University of Catania, Viale A. Doria 6, Catania, CT 95125, Italy
e-mail: costanza@dmi.unict.it; nicosia@dmi.unict.it

P. Liò
University of Cambridge, William Gates Bldg, 15 J J Thomson Avenue, Cambridge CB3 0FD, UK
e-mail: pl219@cam.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_26,

in the C_3 photosynthetic Carbon metabolism would provide about 90% of the overall functional optimization. Both in terms of maximal CO_2 Uptake and minimal Nitrogen consumption, these 11 sensitive enzymes are confirmed to play a key role. Finally we present a RA to confirm our findings.

1 Introduction

The Calvin Cycle (C_3 cycle) is a biochemical pathway of plants capable of fixing atmospheric inorganic CO_2 into an organic compound. This biochemical pathway is hence the basis of primary (plants) productivity. The modeling of this pathway (and allied pathway in Carbon metabolism plants) aims at optimization with respect to some specific functional targets of interest for possible biotechnological applications. Photosynthesis models showed that modifying enzyme concentration would allow the increase of the C_3-cycle efficiency, while maintaining the total amount of Nitrogen constant in the plant [1–3]. Where the biochemical pathway of the Calvin cycle is concerned, a seminal work by Zhu et al. [1], based on the Farquhar model [4], showed, with the help of an evolutionary algorithm, how enzyme concentration rearrangements could be capable of increasing the total amount of CO_2 Uptake by a factor of 76% with respect to the results obtained with the initial concentrations characteristic of the natural leaf. CO_2 Uptake rate at the natural enzyme concentration was calculated in the latter work as 15.5 μ mol m^{-2} s^{-1} in normal "air". This value is within the range of typical CO_2 Uptake rates calculated in the field for C_3 leaves [5] and can then be considered a good approximation. More recently, new efficient models showed that strategies modifying enzyme concentrations may lead to an increase in CO_2 amount of 135% with respect to the initial natural value [2, 3]. In particular, Stracquadanio et al. [2] used also the concepts of robustness and sensitivity for assessing the evaluation of confidence limits in the results obtained by perturbing the new identified solutions; this simulates typical "in-vitro" implementation variables (refer to Sensitivity and Robustness in Sect. 2). An important question regarding this re-optimization is: *why did the evolution process and not optimize enzyme concentration in order to maximize* CO_2 *fixation*? One hypothesis to answer this non-trivial question is that the Calvin cycle pathway evolved during a time in which CO_2 atmospheric concentration was much lower compared to current values. Additionally, some of the enzymes (such those belonging to the photorespiration pathway) whose reduction in concentration would result in a theoretical increase in photosynthesis efficiency, might be strongly linked to other biological functions (e.g., photosystem protection and Nitrogen assumption [6, 7]).

In this paper, we improve and advance the work started in [2] by defining and exploiting an investigation pipe-line that composes Sensitivity Analysis (SA), Single- and Multi-objective optimization, and Robustness Analysis (RA) to move toward the study of the artificial photosynthesis. More in detail, these techniques are composed into the following pipe-line: beginning with a (1) system of ODEs

(refer to coming paragraph) to have a computational model of C_3 photosynthetic pathway, (2) we exploit SA to identify which are the sensitive tuning gears of the system, then (3) we exploit SO and MO optimization to re-optimize the pathway in a functional fashion; (4) we adopt RA to have a quantitative prediction about the stability of the lately found optimizations. Each step is iterated a number of times, until a reasonable solution stability and then experimental feasibility is achieved. Then (5) we compare the newly optimized solutions with natural values to assess solution key changes and to possibly read new insights in the pathway mechanisms. The paper is structured as follows. Section 2 details all of the methods adopted in our design workflow. Section 3 presents the results obtained exploiting these methods. In Sect. 4 conclusions and future directions are presented.

2 Methods

As mentioned above, the computational simulation of the C_3 Carbon metabolism requires the definition of a set of ODEs; in our research work, it is considered the model proposed by [1]. The model takes into account rate equations for each discrete step in photosynthetic metabolism, check-point equations for conserved quantities (e.g., total leaf Nitrogen) and a set of ODEs to describe each pathway mechanism: from initial concentration of nutrients of a cell of a leaf, toward enzyme-mediated reactions, and having a consequent CO_2 uptake. The model assumes that the total protein–nitrogen in the enzymes is $1\,\mathrm{g\,m^{-2}}$; the mass nitrogen in each enzyme, in a $1\,\mathrm{m^2}$ leaf area, is computed on the basis of the number of active sites, catalytic rate per active site, molecular mass of each enzyme, and the ratios between V_m of different enzymes. Mole of each protein is then calculated based on the molecular mass and the mass of each protein, i.e., the total concentration of the adenylate nucleotides ([CA]) in the chloroplast stroma (i.e., the sum of [ATP] and [ADP]) is assumed to remain constant. The V_m for each enzyme is then calculated based on the amount of each enzyme and the volume of the compartment that it occupies in $1\,\mathrm{m^2}$ leaf area. The total concentration of the adenylate nucleotides ([CA]) in the chloroplast stroma, the sum of [ATP] and [ADP], are assumed to remain constant. Similarly, the sum of [NADPH] and [NADP] in the chloroplast stroma ([CN]) are assumed constant. The export of PGA, GAP, or DHAP from the chloroplast to the cytosol is associated with a counterimport of the phosphate, mediated by a phosphate translocator. Consequently, the total concentration of phosphate in the stroma ([CP]) is assumed constant. Finally, a set of ODEs encodes the rates of changes in concentration for each metabolite; the latter is represented by the difference between the rates of those reactions generating the metabolite and the rate of the reactions consuming it. It is clear that the volume of the chloroplast stroma can be different from the cytosol one in a typical higher plant cell; in this scenario, it has been assumed a 1:1 ratio in the computation of concentrations within these two compartments.

2.1 Pathway Sensitivity Analysis

Morris method [8] has been adopted for the evaluation of sensitive components in the set of ODEs mentioned. The main idea behind SA is the identification of crucial pathway gears, whose tuning results in a major system response. More in detail, the SA here adopted belongs to the class of the *one-factor-a-time* (OAT) methods [9], aiming at the evaluation of pathway sensitivity by means of a series of stimuli in a way such that only one input is perturbed while all of the others are kept at their nominal value. Considering for the moment our pathway as a black-box with certain inputs and certain outputs, each SA step-variation, computed for each input is calculated as: $u_i = (P(x_1, x_2, \ldots, x_i + \Delta x_i, \ldots, x_k) - P(x_1, x_2, \ldots, x_i, \ldots, x_k))/\Delta x_i$, where P is the result computed from the pathway model with input x; in particular $x_1, x_2, \ldots, x_i + \Delta x_i, \ldots, x_k$ is the perturbed input vector and $x_1, x_2, \ldots, x_i, \ldots, x_k$ is the nominal input vector. For each factor, a group of outcomes u_i is collected and, as metrics for sensitivity, the mean ${\mu_i}^*$ and the standard deviation σ_i are computed. Highly linear behaviors should be expected from those inputs with a high value of ${\mu_i}^*$. A completely different behavior, such as highly non-linear or counterintuitive responses should be expected from those inputs with high σ_i values. For each enzyme (i.e., input) we use the five concentrations under consideration (refer to Table 26.1 in Sect. 3) as nominal values, computing 20 different factor levels, altered for 10 times each one. As bounds of this SA we adopt $\pm 100\%$ of the nominal value, for each input enzyme concentration. The result of this analysis, highlighted how there are *eleven* enzymes that have to be considered extremely sensitive when compared to all of the others [2]. These enzymes are: *Rubisco, PGA kinase, GAP dehydrogenase, FBP aldolase, FBPase, SBP aldolase, SBPase, Phosphoribulose kinase, ADPGPP, Phosphoglycolate phosphatase, and GDC*. These enzymes showed indeed high values of σ_i (i.e., $1 < \sigma_i < 15$), when compared to all of the others (i.e., $10^{-4} < \sigma_i < 1$).

The definition of a set of linked ODEs gives a mathematical description of the chemical process and, successively, the Morris analysis gives useful insights on linear and non-linear contribution of enzymes to the Carbon metabolism. However, it is important to validate these results by taking into account the interaction map defined by the pathway; it is plausible to assume that sensitive enzymes should be hubs of the photosynthesis pathway. This information has been obtained using Rosvall community detection method [10]; the interaction map we gained confirms these assumptions. Figure 26.1 shows how Rubisco and GAP dehydrogenase are the most strongly regulated enzymes of the Calvin Cycle. Both enzymes are light regulated. Transketolase is another key enzyme, since it uses as substrates Fructose-6-P (otherwise destined to exit from the cycle toward the starch biosynthetic pathway) and 3-P-Glyceraldehyde, that is produced by the enzyme GAP dehydrogenase itself. These enzymes correspond to the main nodes of the Calvin Cycle leading to the other biosynthetic pathways. Phosphoglycolate phosphatase is the first enzyme of the photorespiration pattern linked to the Oxygenase activity of Rubisco and Glycine

Table 26.1 Concentrations of the enzymes, individual robustness, CO_2 uptake rate (at $c_i = 270\,\mu$ mol mol^{-1}, reflecting current CO_2 atmospheric concentration), global and local robustness values. The first enzyme value column reports touchstone concentrations used in our simulations: the initial/natural leaf. The second value column reports the result of the optimization in which only the eleven sensitive enzymes are altered, while all of the others are kept at their nominal values. Third column reports the best-known leaf design, in terms of CO_2 uptake and robustness. The fourth column reports the result of a simulation, where the enzymes Cytosolic FBP aldolase, Cytosolic FBPase, UDP-Glc pyrophosphorylase have been maintained to their initial values. Last column reports the most efficient known point in terms of CO_2, but corresponds to a highly instable solution

Enzyme name	Initial concentration mg N m^{-1} (the natural leaf)	Optimal concentration of 11 sensitive enzymes mg N m^{-1}	Optimal and robust concentration mg N m^{-1}	Optimal and robust concentration mg N m^{-1} (3 fixed enzymes)	Optimal but *not* robust concentration mg N m^{-1}
Rubisco	517.00 (100)	784.27 (**84**.5)	860.226 (100)	856.44 (100)	861.93 (**39**)
PGA kinase	12.20 (100)	4.66 (100)	3.989 (100)	3.63 (100)	3.98 (**0**)
GAP DH	68.80 (100)	69.03 (**81**.5)	64.483 (100)	65.08 (100)	63.55 (**17**)
FBP aldolase	6.42 (100)	10.40 (100)	9.050 (100)	10.86 (100)	9.29 (**30.5**)
FBPase	25.50 (100)	29.44 (100)	26.889 (100)	32.24 (100)	27.03 (**0**)
Transketolase	34.90 (100)	*34*.90 (100)	8.247 (100)	16.93 (100)	16.98 (*100*)
SBP aldolase	6.21 (100)	5.55 (100)	6.661 (100)	5.75 (100)	5.94 (**0**)
SBPase	1.29 (100)	4.70 (100)	4.397 (100)	4.43 (100)	4.31 (**1**)
PRK	7.64 (100)	7.04 (100)	7.007 (100)	6.38 (100)	7.99 (**22.5**)
ADPGPP	0.49 (100)	2.12 (100)	0.721 (100)	5.09 (100)	1.22 (**0**)
PGCA Pase	85.20 (100)	0.95 (100)	0.325 (100)	0.20 (100)	0.00 (**0**)
Glycerate kinase	6.36 (100)	*6*.36 (100)	0.005 (100)	0.00 (100)	0.00 (*100*)
Glycolate oxidase	4.77 (100)	*4*.77 (100)	0.019 (100)	0.16 (100)	0.00 (*100*)
GSAT	17.30 (100)	*17*.30 (100)	0.027 (100)	0.00 (100)	0.00 (*100*)
Glycer. dehyd.	2.64 (100)	*2*.64 (100)	0.003 (100)	0.00 (100)	0.00 (*100*)
GGAT	21.80 (100)	*21*.80 (100)	0.00005 (100)	0.00 (100)	0.00 (*100*)
GDC	179.00 (100)	0.02 (100)	0.00003 (100)	0.00 (100)	0.00 (*100*)
Cyt. FBP ald.	0.57 (100)	*0*.57 (100)	2.127 (100)	*0*.57 (100)	2.03 (**0**.5)

(continued)

Table 26.1 (continued)

Enzyme name	Initial concentration mg N m^{-1} (the natural leaf)	Optimal concentration of 11 sensitive enzymes mg N m^{-1}	Optimal and robust concentration mg N m^{-1}	Optimal and robust concentration mg N m^{-1} (3 fixed enzymes)	Optimal but *not* robust concentration mg N m^{-1}
Cyt. FBPase	2.24 (100)	*2*.24 (100)	5.554 (100)	*2*.24 (100)	5.27 (**30**.5)
UDPGPP	0.07 (100)	*0*.07 (100)	0.531 (100)	*0*.07 (100)	0.50 (**0**)
SPS	0.20 (100)	*0*.20 (100)	0.034 (100)	0.01 (100)	0.03 (**30**.5)
SPP	0.13 (100)	*0*.13 (100)	0.031 (100)	0.01 (100)	0.03 (**0**)
F26BPase	0.02 (100)	*0*.02 (100)	0.00 (100)	0.00 (100)	0.00 (*100*)
CO_2 uptake $\frac{\mu \text{ mol}}{\text{m}^2 \text{ s}}$	15.486	22.420	20.626	22.156	
(Local R. %, Global R. %)	(100, 81.80)	(81.5, 78.3)	(100, 97.2)	(100, 92.6)	(0, 39.18)

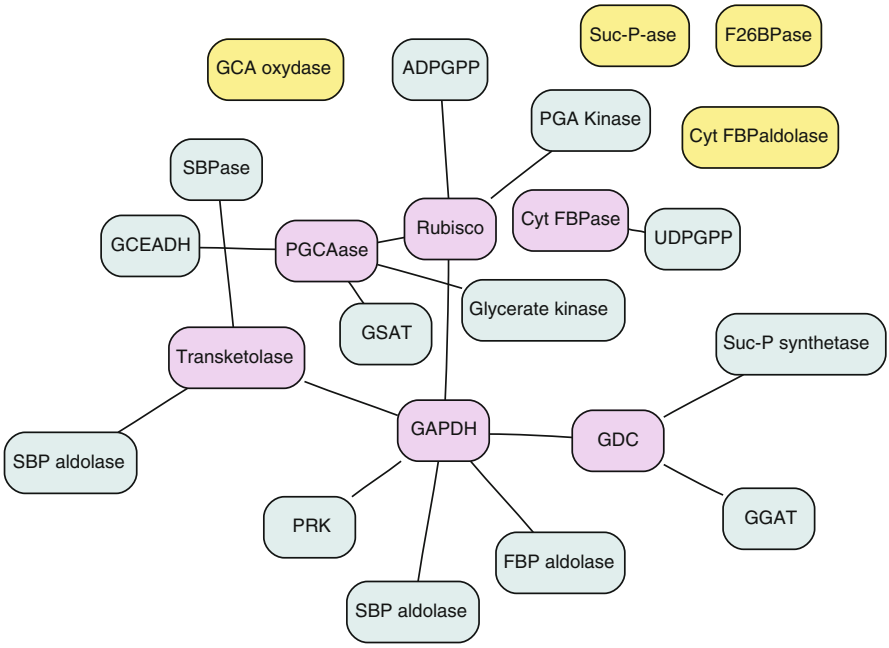


Fig. 26.1 Relationships within the Pathway. Enzyme interactions in the C_3 photosynthetic Carbon metabolism

decarboxylase the enzyme responsible for the loss of the CO_2 fixed by Rubisco. Figure 26.1 presents also 10 enzyme clusters in the C_3 photosynthetic Carbon metabolism pathway:

(1) Phosphoglycolate phosphatase, Glycerate kinase, Ser glyoxylate aminotransferase, Glycerate dehydrogenase; (2) Transketolase, SBPase, SBP aldolase, Enzyme 9; (3) GAP dehydrogenase, GAP dehydrogenase, FBPase, FBP aldolase, Phosphoribulose kinase; (4) Rubisco, PGA Kinase, Enzyme11, ADPGPP; (5) GDC, Glu glyoxylate aminotransferase, Suc-P synthetase; (6) Cytosolic FBPase, UDP-Glc pyrophosphorylase; (7) F26BPase; (8) Suc-P phosphatase; (9) Cytosolic FBP aldolase; (10) Glycolate oxidase.

2.2 *Optimization of Sensitive Enzymes: Single- and Multi-Objective Optimization*

Once these eleven enzymes have been identified as *sensitive*, their optimization has been then evaluated and compared to the optimization of the complete pathway system. This means inspecting a search space in 11 dimensions (sensitive enzymes), instead of inspecting the complete pathway one, that has 23 dimensions. The aim of this is evidently the evaluation of the contribution of these *sensitive enzymes*

in the optimization of the photosynthetic metabolism considered. To achieve this aim we have employed Single- and Multi-objective optimization algorithms; these approaches have considered both the complete domain space ($x \in \mathbb{R}^{23}$) and the "sensitive domain", as cropped by those 11 most sensitive enzymes ($x \in \mathbb{R}^{11}$). Considering a fixed total amount of protein Nitrogen available to the leaf ($1\,\mathrm{g\,m^{-2}}$, i.e., ca. $20.833 \times 10^4\ \mathrm{mg\,l^{-1}}$), we have used Parallel Optimization Algorithm (PAO) [2] to let a pool of solutions evolve in an archipelago fashion. Fixing at their natural value all but sensitive enzymes, PAO has evaluated a number of new enzyme concentration profiles and has associated a CO_2 Uptake to each one of them by computing the system of ODEs mentioned above. The key aspect of PAO is its ability to share *portions* of promising solutions among optimization cores. We adopted two optimization cores (islands): on one PAO runs A-CMA-ES [2] algorithm and on the other, it runs DE [11]; solution portions are exchanged every 200 generations with probability 1/2. On both islands the optimization aim is the same: to find all those sensitive enzyme concentration vectors $\hat{x} = [\mathrm{conc}_1, \mathrm{conc}_2, \ldots, \mathrm{conc}_{11}]$, such that, when $\hat{x}$ is composed with the other enzyme values kept at their nominal value, the resulting CO_2 Uptake function is maximized:

$$\max_{\hat{x} \in \mathbb{R}^{11}} \left(f_1 \ (\hat{x}, x_{\text{non-sensitive}}) \right). \tag{26.1}$$

Relaxing the constraint about the fixed total amount of protein Nitrogen, a new optimization has been performed. The focus here is again on those *sensitive enzymes* and on the comparative evaluation with respect to the rest of the pathway. Parallel Multi-objective Optimization (PMO2) [12] has been adopted to evaluate the contextual optimization of CO_2 Uptake and total Nitrogen needed. Gaining higher CO_2 Uptake rates employing less Nitrogen mean absorbing more CO_2, while consuming less "leaf-fuel", this means, a more efficient metabolism cycle. This means that additionally to the maximization of the CO_2 Uptake function, a new function is taken into account; it is the minimization:

$$\min_{\hat{x} \in \mathbb{R}^{11}} \left(f_2 \ (\hat{x}) \right) = \min_{\hat{x} \in \mathbb{R}^{11}} \left(\sum_{i=1}^{11} \frac{\hat{x}[i] \times \mathrm{WM}_i}{\mathrm{BK}_i} \right), \tag{26.2}$$

where BK_i is the catalytic number or turnover number, and WM_i the molecular weight of the ith enzyme, respectively. Hence, our search for $\hat{x}$ has to accomplish a contextual trade-off between maximal CO_2 Uptake rate and minimal Nitrogen employment. Note that the minimization of f_2 does not take into account those enzymes that are not sensitive: since they are fixed in our search for minima, the second objective is simply shifted by a constant quantity for all of the points evaluated; this obviously does not impact the optimization as all of the point efficiencies are translated as a whole. In PMO2, at the beginning, a random initial

population P_0 is generated and it is sorted based on the non-domination criterion; a fitness proportional to its non-domination level is assigned to each solution. Non-dominated sorting has been introduced in order to rank the population according to its domination level. For each solution, we compute the domination count n_p, that denotes the number of solutions dominated by p, and S_p, that is, the set of solutions dominated by p; obviously, all the solutions belonging to the first front have a domination count set to zero. For each solution with $n_p = 0$, we pick each element of S_p and reduce its domination count by one; if for any member of S_p the domination count goes to zero, it is put into a separate list Q. The process is iterated until each solution is assigned to a front. The algorithm proceeds using binary tournament selection, recombination and mutation to create a population of offspring Q of size N. At each generation g, a population $R_g = P_g \cup Q_g$ is built and, hence, it is sorted according to non-domination; it is important to note that since the parent population is put in R_g, elitism is assured. The selection procedure chooses the individual with $n_p = 0$ and, then, it picks individuals from other domination levels if there are not N non-dominated individuals; this set of dominated solutions is chosen according to a *crowding-comparison* operator.

2.3 Leaf Candidate Robustness

Once single-objective optimization algorithms have found the best solutions to the f_1 maximization problem, we adopt the RA to assess the intrinsic stability of the solution. The definition of robustness here adopted has to be considered as the ability of a system to survive random perturbations [13]. In order to evaluate the robustness of enzymes partitions, the *robustness condition*, ρ, and the *uptake yield*, Γ, have been defined [12]. Let $\bar{x} \in \mathbb{R}^{23}$ an enzyme partitioning and $f : \mathbb{R}^n \to \mathbb{R}$ a function computing the expected CO_2 uptake rate value of $\bar{x}$. Given an enzyme partition $\bar{x}_*$ obtained by perturbing $\bar{x}$, the robustness condition ρ is defined as follows:

$$\rho(\bar{x}, \bar{x}_*, f, \epsilon) = \begin{cases} 1 \text{ if } \mid f(\bar{x}) - f(\bar{x}_*) \mid \leq \epsilon \\ 0 \qquad\quad \text{otherwise} \end{cases}, \tag{26.3}$$

where the robustness threshold ϵ denotes the maximum percentage of variation from the nominal CO_2 uptake value.

Given an ensemble $\mathcal{T}$ of perturbed enzymatic concentrations obtained by perturbing $\bar{x}$, the uptake yield Γ is defined as follows:

$$\Gamma(\bar{x}, f, \epsilon) = \frac{\sum_{\tau \in \mathcal{T}} \rho(\bar{x}, \tau, f, \epsilon)}{|\mathcal{T}|}. \tag{26.4}$$

The ensemble T has been generated using a Monte-Carlo algorithm; mutations occurring on all the enzymes (global RA) and one enzyme at time (local RA) have

been considered [13]. A maximum perturbation of 10% has been fixed for each enzyme concentration, and then it has been generated an ensemble of 5×10^3 trials for the global RA and 200 trials, for each enzyme, for the local RA. All the experiments assume $\epsilon = 5\%$ of the nominal uptake rate value.

3 Results

3.1 *Sensitive Enzymes for the Uptake Objective*

The aim of the present research work is to compare how the exclusive targeting of sensitive enzymes varies either when f_1 (26.1) is maximized or when f_1 is maximized while f_2 (26.2) is minimized. As mentioned above, we want to evaluate if it is really worth moving from the original search space ($x \in \mathbb{R}^{23}$) to the sensitive enzymes search space ($x \in \mathbb{R}^{11}$), without important losses in terms of functional pathway optimization.

In order to suggest correct and minimal biotechnological targets, we present in Table 26.1, four alternative leaf designs, unraveled by PAO algorithm: these solutions represent candidate enzyme concentration whose task is the increase of the CO_2 Uptake rate, while maintaining the actual amount of total Nitrogen contained in the enzymes.

Best solutions obtained on the optimization of *sensitive enzymes* both by PAO (max f_1) and PMO2 (max $f_1 \cap min\ f_2$) have been further inspected and compared to the natural leaf. Figure 26.2 shows how sensitive enzymes changed their concentration in order to maximize the CO_2 Uptake (i.e., the best point found by "PAO 11 Sens" and the end of its convergence, Fig. 26.3). Exclusive targeting of sensitive enzymes brought an optimal uptake rate of $33.317\,\mu\,\mathrm{mol\,m^{-2}s^{-1}}$, that is, only -9% than the most efficient known point; this confirms how these 11 enzymes perform about 91% of the whole photosynthetic optimization. It is also worth noting in Fig. 26.2 histogram how all of the increases and decreases in optimal enzyme concentration are within the range $\times 0.001 - \times 4.3$; this is a plausible biotechnological range, indeed, around a fivefold increase can be achieved by means of enzyme promoters. Afterward, RA has been employed to assess the stability of this solution as well: as reported in Table 26.1, this point has an overall local robustness that is not very high (81.5%) and a global robustness comparable to the natural one (78.3%).

To evaluate the effective contribution of these 11 sensitive enzymes, we have fed into PAO the optimization problem in which the variable enzyme set is extended from 11 to all but those 3 that seemed to play an important role looking at single histograms but did not play an important role according to the SA. These three enzymes are: Cytosolic FBP aldolase, Cytosolic FBPase, and UDPGP. The result of this optimization is reported in Fig. 26.4 (refer to Table 26.1 for single enzyme variation). This configuration registered a CO_2 Uptake rate of $36.197\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$,

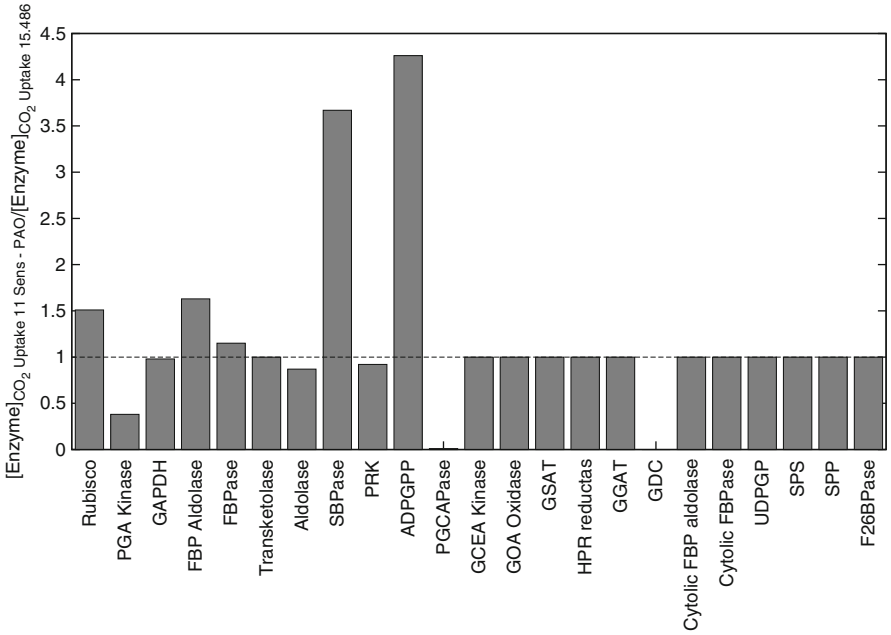


Fig. 26.2 The ratio of the enzyme concentrations optimized by the PAO algorithm ($33.317\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$) at a $c_i = 270\,\mu\,\mathrm{mol\,mol^{-1}}$ compared to the initial concentrations ($15.486\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$). Optimization of CO_2 uptake rate perturbing the 11 most sensitive enzymes only (*Rubisco, PGA kinase, GAP dehydrogenase, FBP aldolase, FBPase, SBP aldolase, SBPase, Phosphoribulose kinase, ADPGPP, Phosphoglycolate phosphatase, and GDC*). These enzymes are the most important enzymes in the studied model of the Carbon metabolism) while the remaining enzymes are maintained at their initial concentration

that is, ca. +8% when compared to the optimization of sensitive enzymes. RA has reported 100% and 92.6% for local and global robustness, respectively, proving the stability of this solution.

We have then decided to target four enzymes of the C_3 metabolic pathway: we all know biotechnological intervention is hard and error prone, then we want to minimize intervention points. This decision came from combining the information gained on different points of our research: (1) promising leaf engineering obtained with the alteration of the 11 most sensitive enzymes, (2) those three enzymes that could have played a crucial role do not seem to affect the Uptake objective, (3) out of those 11 enzymes pointed out by the SA only *six* of them present a change out of the range $0.2\times - 1.5\times$ (Rubisco, FBP aldolase, SBPase, ADPGPP, Phosphoglycolate phosphatase, and GDC). Because of the non-optimal local robustness showed (refer to Table 26.1, 84.5%), Rubisco has been filtered from the analysis to ensure a fair comparison and a more precise identification of robust targets for biotechnological intervention. These five enzymes have been sorted out into three simulations; we have designed each simulation such that there are two enzymes that showed a

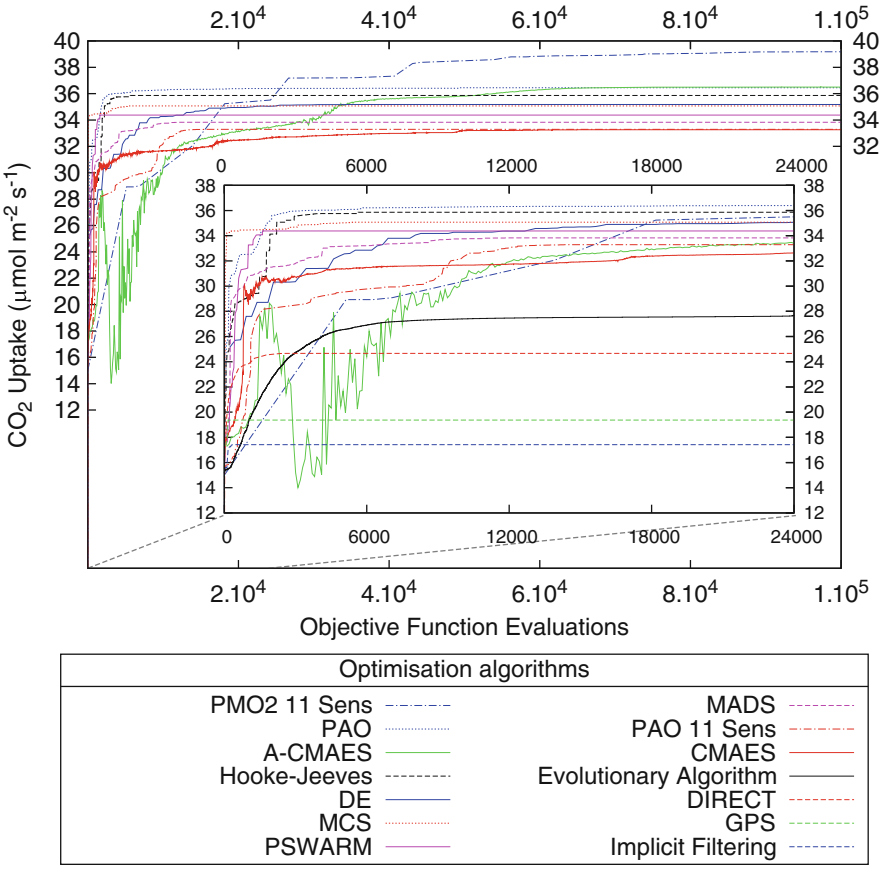


Fig. 26.3 Convergence process of fourteen derivative-free global optimization algorithms; on single objective, the PAO algorithm outperforms all of the other algorithms (from best to worst: Hooke–Jeeves [14], DE [11], MCS [15], PSWARM [16], MADS [17], CMAES [18], Evolutionary Algorithm in [1], DIRECT [19], GPS [20], and Implicit Filtering [21]) in the single-objective optimization of the full problem version (i.e., optimization of all of the enzymes) and then it has been adopted for the optimization of the reduced model which has optimized using the most sensitive enzymes (in the legend "PAO 11 Sens"). This optimization comparison has been performed to maximize light-saturated photosynthetic rate (CO_2 Uptake) at $c_i = 270\,\mu$ mol mol^{-1}, that is, the value characteristic of nowadays CO_2 atmospheric concentration. It is also reported the convergence of the PMO2 algorithm, in terms of non-dominated solutions, when the optimization enzyme set is restricted to the eleven sensitive ones ("PMO2 11 Sens" in the legend)

negative fold-change and other two with a positive one; this has been put into place to ensure a balance with respect to the total Nitrogen partitioning. Indeed, having a fixed amount of protein Nitrogen, it is likely that to allow ADPGPP to grow as in Fig. 26.2, we have to couple it with some of those enzymes that diminished their concentrations (i.e., Phosphoglycolate phosphatase, and GDC). Summarizing, we have further inspected our pathway through three more simulation configurations:

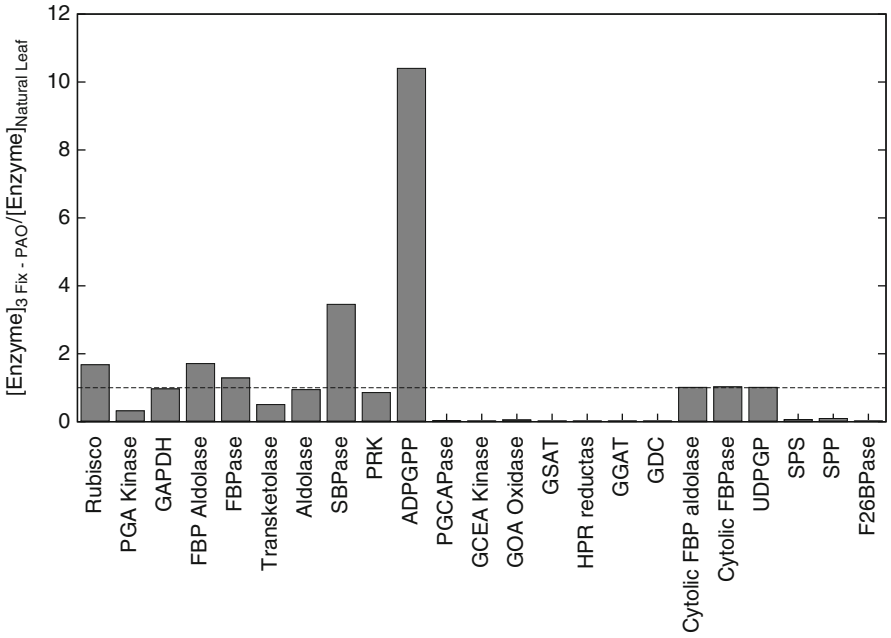


Fig. 26.4 Changes in the concentrations of Carbon metabolism enzymes with respect to their natural values when three metabolites are kept constant: Cytosolic FBP aldolase, Cytosolic FBPase, and UDPGP. The C_i has value $270\,\mu\,\mathrm{mol\,mol}^{-1}$, reflecting nowadays condition. This configuration obtains CO_2 Uptake rate of $36.197\,\mu\,\mathrm{mol\,m}^{-2}\,\mathrm{s}^{-1}$, suggesting that even fixing these three enzymes the uptake performance can be very effective

Var4-1 (that tunes only FBP aldolase, SBPase, Phosphoglycolate phosphatase, and GDC), *Var4-2* (FBP aldolase, ADPGPP, Phosphoglycolate phosphatase, and GDC) and *Var4-3* (that targets only ADPGPP, SBPase, Phosphoglycolate phosphatase, and GDC). Table 26.2 presents these results from a quantitative point of view. It is of note how when the optimization is pushed selectively to the limit varying only 4 enzymes (i.e., *Var4* simulations), we observe enzyme concentrations that readjust their values within the range $0.001\times - 160\times$. Such a step variation has to be considered a strong signal, as targeting such a small set of enzymes we are unraveling how the metabolism can become functional (uptake maximization) without invalidating any other connected pathway.

3.2 Sensitive Enzymes on Uptake Maximization and Nitrogen Minimization

In order to compare the optimization of the whole system with the optimization of the sensitive enzymes, we have compared the Pareto frontiers obtained on

Table 26.2 Concentrations of the enzymes, individual robustness, CO_2 uptake rate (at $c_i = 270\,\mu$ mol mol^{-1}, reflecting current CO_2 atmospheric concentration), global and local robustness values. The first enzyme value column reports touchstone concentrations used in our simulations: the initial/natural leaf. Columns 3–5 present enzyme values obtained as result of Var4-1, Var4-2 and Var4-3 simulations and robustness values associated with each leaf engineering

Enzyme name	Initial concentration mg N m^{-1} (the natural leaf)	Optimal concentration of Var4-1 sensitive enzymes mg N m^{-1}	Optimal concentration of Var4-2 sensitive enzymes mg N m^{-1}	Optimal concentration of Var4-3 sensitive enzymes mg N m^{-1}
Rubisco	517.00 (100)	*517.00* **(99.5)**	*517.00* (100)	*517.00* **(98.5)**
PGA kinase	12.20 (100)	*12.20* (100)	*12.20* (100)	*12.20* (100)
GAP DH	68.80 (100)	*68.80* (100)	*68.80* (100)	*68.80* (100)
FBP aldolase	6.42 (100)	14.76 (100)	10.40 (100)	*6.42* (100)
FBPase	25.50 (100)	*25.50* (100)	*25.50* (100)	*25.50* (100)
Transketolase	34.90 (100)	*34.90* (100)	*34.90* (100)	*34.90* (100)
SBP aldolase	6.21 (100)	*6.21* (100)	*6.21* (100)	*6.21* (100)
SBPase	1.29 (100)	198.05 (100)	*1.29* **(70)**	91.13 (100)
PRK	7.64 (100)	*7.64* (100)	*7.64* (100)	*7.64* (100)
ADPGPP	0.49 (100)	*0.49* (100)	43.52 (100)	46.52 (100)
PGCA Pase	85.20 (100)	63.30 (100)	220.93 (100)	131.79 (100)
Glycerate kinase	6.36 (100)	*6.36* (100)	*6.36* (100)	*6.36* (100)
Glycolate oxidase	4.77 (100)	*4.77* (100)	*4.77* (100)	*4.77* (100)
GSAT	17.30 (100)	*17.30* (100)	*17.30* (100)	*17.30* (100)
Glycer. dehyd.	2.64 (100)	*2.64* (100)	*2.64* (100)	*2.64* (100)
GGAT	21.80 (100)	*21.80* (100)	*21.80* (100)	*21.80* (100)
GDC	179.00 (100)	0.02 (100)	0.49 (100)	22.19 (100)

Cyt. FBP ald.	0.57 (100)	*0.57* (100)	*0.57* (100)	*0.57* (100)
Cyt. FBPase	2.24 (100)	*2.24* (100)	*2.24* (100)	*2.24* (100)
UDPGPP	0.07 (100)	*0.07* (100)	*0.07* (100)	*0.07* (100)
SPS	0.20 (100)	*0.20* (100)	*0.20* (100)	*0.20* (100)
SPP	0.13 (100)	*0.13* (100)	*0.13* (100)	*0.13* (100)
F26BPase	0.02 (100)	*0.02* (100)	*0.02* (100)	*0.02* (100)
CO_2 uptake $\frac{\mu \text{ mol}}{\text{m}^2 \text{ s}}$	15.486	22.420	20.626	22.156
(Local R. %, Global R. %)	(100, 81.80)	(99.5, 91.8)	(70, 69.4)	(98.5, 92.9)

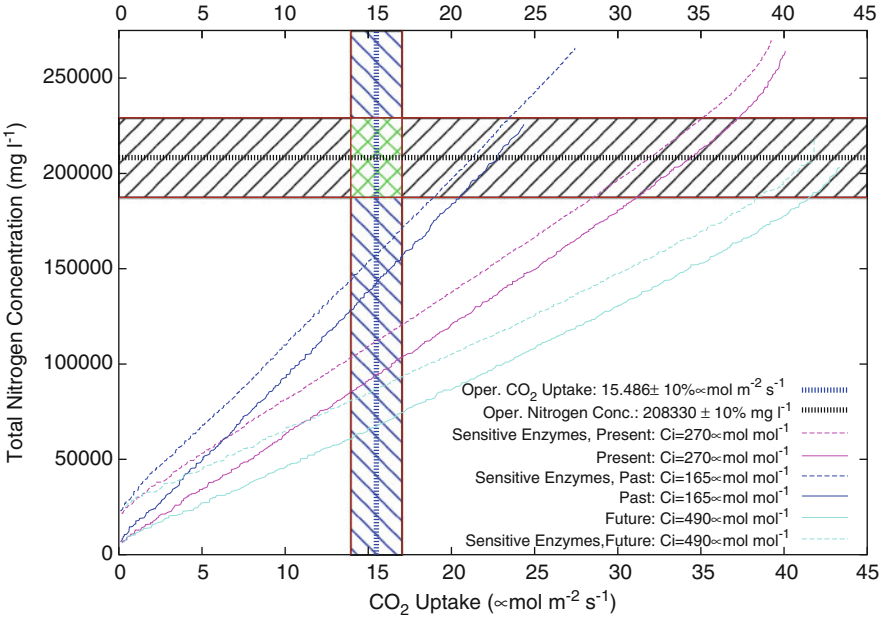


Fig. 26.5 CO_2 uptake and protein–nitrogen concentration trade-off. Maximizing the CO_2 Uptake while minimizing the total amount of protein–nitrogen concentration; the operative area of natural leaves is located in the green checked area. The label "Sensitive Enzymes" indicates the multi-objective optimization using the 11 most sensitive enzymes of the model, the three resulting Pareto Fronts have been dominated by the multi-objective optimization over all enzymes of the model. This trade-off search has been carried out for the three c_i concentration referring to the environmental condition in effect 25 million years ago, nowadays and in 2100 AC

both tasks by PMO2 in three conditions: Present, Past, and Future atmospheric conditions, i.e., $c_i \in \{270, 165, 490\}\ \mu$ mol mol^{-1}. Figure 26.5 presents these Pareto frontiers comparison: in this multi-objective optimization, as we saw in the single-objective one, the optimization of only the sensitive enzymes causes a minor loss in optimization performances. It is interesting, how reducing the search space to less than half of the dimensions (i.e., sensitive optimization), the performances are affected by a factor between 5% and 10%. It is also of note how this difference is consistently kept among all of the atmospheric conditions considered. Having a narrower search space means on one hand the achievement of sub-optimal solutions, but on the other hand, it means that during the biotechnological implementation we will have just half of the variables, compared to the original problem. Functional optimization, versus problem dimensionality, is an intrinsic trade-off that shows how we have to accept slightly lower efficiencies if we want the benefits of dealing with half of the unknowns.

Figure 26.6 shows the comparison between the natural leaf and the best non-dominated solution found by PMO2. This non-dominated solution (i.e., it belongs

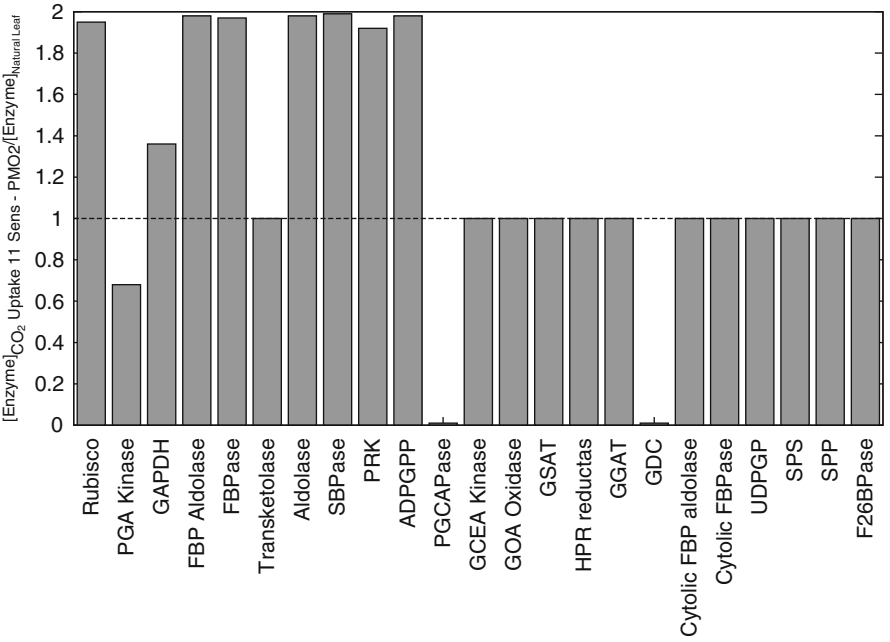


Fig. 26.6 Optimization of CO_2 uptake rate and Nitrogen consumption perturbing the 11 most sensitive enzymes only. The ratio of the enzyme concentrations optimized by the multi-objective optimization algorithm PMO2 ($39.242\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$) at a $c_i = 270\,\mu\,\mathrm{mol\,mol^{-1}}$ compared to the initial concentrations ($15.486\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$). The non-dominated solution here considered is the one with the highest CO_2 uptake rate, which shows a Nitrogen consumption of ca. $269658\,\mathrm{mg\,l^{-1}}$

to the Pareto front, and more in detail, it is the best point found by PMO2 11 Sens and the end of its convergence, Fig. 26.3) is the one with the overall maximal CO_2 Uptake rate ($39.242\,\mu\,\mathrm{mol\,m^{-2}\,s^{-1}}$). It is remarkable how, despite the tremendous increase in uptake rate of ca. 253%, and the relatively high increase in Nitrogen consumption (129%), all of the changes at the enzyme level are within the range $\times 0.01 - \times 2$. From a theoretical point of view, these changes are even easier to implement with the current chemical processing, when compared to the one reported in Fig. 26.2.

4 Conclusion

The statistician George Box said: “All models are wrong, but some are useful”. Nowadays this sentence would reflect many things: the continuous improvement of developing new models in all scientific fields, the different level of

abstractions that a model could express and our difficulties in modeling multi-scale or compartmentalized dynamical systems. In conclusion, we are delighted to report that the modeling of C_3 carbon metabolism is a thriving field of research. It has two immediate and important benefits: the improved understanding of the process that shapes photosynthesis in plants and the possibility to test engineered solutions in silico using a mature single- and multi-objective optimization methodology. We believe that our quantitative findings of a small number of enzymes that concentrate the biotechnology potentialities and our methodological improvements could effectively represent a significant contribution to the community working in this area.

References

1. Zhu XG, de Sturler E, Long SP (2007) Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: A numerical simulation using an evolutionary algorithm. Plant Physiol 145:513–526
2. Stracquadanio G, Umeton R, Papini A, Liò P, Nicosia, G (2010) Analysis and optimization of C3 photosynthetic carbon metabolism. In: Rigoutsos I, Floudas CA (eds) Proc BIBE 2010, 10th IEEE Int Conf Bioinformatics and Bioengineering, May 31–June 3, 2010, Philadelphia, PA, USA, IEEE Computer Society, pp 44–51
3. Papini A, Nicosia G, Stracquadanio G, Lio P, Umeton R (2010) Key Enzymes for the optimization of CO_2 uptake and nitrogen consumption in the C3 photosynthetic carbon metabolism. J Biotechnol 150:525–526
4. Farquhar G, Caemmerer S, Berry J (1980) A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species. Planta 149(1):78–90
5. Wullschleger S (1993) Biochemical limitations to carbon assimilation in C3 plants: a retrospective analysis. J Exp Bot 44:907–920
6. Wingler A, Lea P, Quick W, Leegood R (2000) Photorespiration: metabolic pathways and their role in stress protection. Philos Trans Royal Soc London. Ser B: Biol Sci 355(1402):1517
7. Heber U, Bligny R, Streb P, Douce R (1996) Photorespiration is essential for the protection of the photosynthetic apparatus of C3 plants against photoinactivation under sunlight. Bot Acta 109:307–315
8. Morris M (1991) Factorial sampling plans for preliminary computational experiments. Technometrics 33(2):161–174
9. Saltelli A, Tarantola S, Campolongo F (2004) Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons Inc.
10. Rosvall M, Bergstrom C (2007) An information-theoretic framework for resolving community structure in complex networks. Proc Natl Acad Sci 104(18):7327
11. Storn R, Price K (1997) Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359
12. Umeton R, Stracquadanio G, Sorathiya A, Papini A, Liò P, Nicosia G (2011) Design of robust metabolic pathways. In: Proc 48th design automation conference, DAC 2011, San Diego, CA, USA, June 5–9, 2011, ACM, pp 747–752
13. Stracquadanio G, Nicosia G (2011) Computational energy-based redesign of robust proteins. Comput Chem Eng 35(3):464–473
14. Hooke R, Jeeves TA (1961) "Direct search" solution of numerical and statistical problems. J ACM 8(2):212–229
15. Huyer W, Neumaier A (1999) Global optimization by multilevel coordinate search. J Global Optim 14(4):331–355

16. Vaz A, Vicente L (2007) A particle swarm pattern search method for bound constrained global optimization. J Global Optim 39(2):197–219
17. Audet C, Dennis JE (2007) Mesh adaptive direct search algorithms for constrained optimization. SIAM J Optim 17(1):188–217
18. Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies. Evol Comput 9(2):159–195
19. Jones DR, Perttunen CD, Stuckman BE (1993) Lipschitzian optimization without the Lipschitz constant. J Optim Theor Appl 79(1):157–181
20. Lewis R, Torczon V (1999) Pattern search algorithms for bound constrained minimization. SIAM J Optim 9(4):1082–1099
21. Gilmore P, Kelley CT (1995) An implicit filtering algorithm for optimization of functions with many local minima. SIAM J Optim 5(2):269–285

Chapter 27
Formal Methods for Checking the Consistency of Biological Models

Allan Clark, Vashti Galpin, Stephen Gilmore, Maria Luisa Guerriero, and Jane Hillston

Abstract Formal modeling approaches such as process algebras and Petri nets seek to provide insight into biological processes by using both symbolic and numerical methods to reveal the dynamics of the process under study. These formal approaches differ from classical methods of investigating the dynamics of the process through numerical integration of ODEs because they additionally provide alternative representations which are amenable to discrete-state analysis and logical reasoning. Backed by these additional analysis methods, formal modeling approaches have been able to identify errors in published and widely-cited biological models. This paper provides an introduction to these analysis methods, and explains the benefits which they can bring to ensuring the consistency of biological models.

1 Introduction

Modeling complex systems on a computer allows us to investigate the rich dynamics of phenomena which are difficult or impossible to study at first hand. Making and analyzing models may open the door to understanding but the insights and understanding gained depend crucially on the accuracy of the model and the

A. Clark (✉) • S. Gilmore • M.L. Guerriero • J. Hillston
Centre for Systems Biology at Edinburgh, The University of Edinburgh,
Edinburgh EH9 3JU, Scotland, UK
e-mail: A.D.Clark@ed.ac.uk; S.Gilmore@ed.ac.uk; mguerrie@inf.ed.ac.uk;
Jane.Hillston@ed.ac.uk

V. Galpin
Laboratory for Foundations of Computer Science, The University of Edinburgh,
Edinburgh EH8 9AB, Scotland, UK
e-mail: Vashti.Galpin@ed.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_27,

legitimacy of its assumptions. Accurate modeling of complex systems often leads to difficult computational problems where inherent complexities of the problem such as multi-scale populations and widely-separated reaction rates present genuine technical challenges for robust numerical software. When grappling with these challenges it is important not to lose sight of the fact that the quality of the insights obtained from the modeling depend critically on the quality of the model and that – in addition to carrying the burden of computing robust numerical results – modelers must also shoulder the burden of creating accurate biological models.

Adding to the difficulty of the problem, it is very easy to introduce simple errors which may have subtle effects which are extremely difficult to detect. For example, writing down the wrong variable in a differential equation may give rise to a model whose results look plausible (for example, there are no negative concentrations or other non-physical results) but which are essentially meaningless. From the perspective of a traditional differential equation integrator we have an entirely valid system of equations and a well-posed initial value problem; it is simply that it does not capture the phenomenon under study.

Domain-specific modeling languages such as process algebras and Petri nets specifically tailored for biology are helpful here because they define and enforce rules about the internal consistency of models which can allow simple modeling errors to be detected automatically. In this way, these languages can prevent the computational analysis of non-well-formed models and thereby – in some cases – stop erroneous conclusions being derived from erroneous models.

Evidence for the effectiveness of these methods can be seen when formal modeling is applied retrospectively to published models and at that point a previously-unknown error is discovered. It is possible for these errors to be either in the model itself, or in the computational analysis which was carried out in order to reveal the dynamics of the underlying biological process. An example of an error of the former kind in a model of the TNFα-mediated NFκ-B signal transduction pathway was uncovered using the techniques in [1]. An example of an error of the latter kind was uncovered as reported in [2] where Gillespie simulation is used in co-operation with continuous deterministic simulation to reveal a discrepancy which is traced to an incorrect use of a numerical integrator.

Two of the most important motivations for modeling a biological system are: (1) to identify gaps in the existing knowledge of the system, and (2) to generate new insights and understanding without the need to perform laboratory experiments. In the former case we would work through validation where we try to discover whether the behavior of the model agrees with current biological knowledge. In the latter case we investigate specific hypotheses via computational analysis instead of laboratory work.

Although these aspects are closely interconnected, there are existing computational and mathematical techniques which provide features particularly suitable to tackle one or the other. The process of identifying inconsistencies within models is an important phase which should always be performed before any conclusion is drawn from the results of the analysis of the model. Here, we focus on the use of

analysis techniques which have their roots in formal language theory and program analysis. These range from static analysis and control flow analysis through to invariant generation, graph analysis, and bisimulation. These methods enable us to identify flaws in models: both errors due to unknowns or incorrect hypotheses in the biological knowledge which are then unwittingly encoded in the model, leading to a flawed model; and errors which are introduced during model construction such that the model does not faithfully represent current biological understanding.

Several formal methods have been developed (or adapted) in order to model and analyze biological systems, including Petri nets [3]; rewriting systems such as membrane systems [4], and Kappa [5]; and process algebras such as the biochemical stochastic π-calculus [6], Bio-PEPA [7], and the Continuous π-calculus [8]. Most of these languages are equipped with a discrete stochastic semantics, and some also allow for a continuous deterministic interpretation. The analysis techniques which are available differ for the various formalisms. For some of these languages it is possible to employ verification techniques such as model-checking.

These kinds of computational models can either be analyzed statically via techniques which work at the level of the model structure, or can be dynamically executed via stochastic simulation [9] to produce time-course trajectories of amounts of the participating species. For languages which have a deterministic interpretation, numerical solution of the associated set of ordinary differential equations (ODEs) can be also performed, together with the various mathematical methods available for the analysis of ODE systems such as bistability, bifurcation, and continuation analysis. Existing modeling platforms such as the Bio-PEPA Eclipse Plug-in [10] allow modelers to perform model experimentation including parameter sensitivity analysis, components knock-down, and dose–response experiments.

2 Static Analysis

Viewed as a formal text, a biological model contains *definitions* of constituents of the model such as reaction rate constants, kinetic laws, initial concentrations, and chemical species; and it contains *uses* of these definitions. One simple check of self-consistency in the model is to determine that all definitions are used, and that everything which is used has been defined. This type of checking falls within the domain of *static analysis* because it can be performed without executing the model (via simulation or otherwise). The benefits of static analysis are enormous: a vast range of simple modeling errors can be easily and automatically caught at low computational cost.

Automatic static analysis seems such a simple and sensible check that it may be surprising to learn that not all programing languages enforce a static analysis check. For example, the Python programing language [11] does not and so a biological model implemented in Python has had less thorough automatic checking than a model implemented in Bio-PEPA or Snoopy [12] where a static analysis check is

automatically enforced for every version of every model. Similar remarks apply to biological models coded directly in MATLAB [13]. Both Python and MATLAB have separate, optional static analysis tools (PyLint and M-Lint) which may be used by modelers, but are not required.

Static analysis based on the structure of a model is the first step which allows us to identify a number of errors, ranging from syntax errors such as trivial typos in variable names to more subtle omissions of species behavior. Static analysis can also be used in order to verify the presence or absence of deadlocks in the model behavior and to clarify the causal and temporal relations between events.

Most static analysis checks relate to the internal consistency of a model. They generally do not require quantitative information such as kinetic rates and molecular concentrations. The use of high-level modeling languages helps modelers to reduce potential sources of errors by allowing them to define mnemonic names for system components, reaction kinetic laws, and parameters. In addition to reducing the chance of introducing trivial modeling errors, the use of named definitions instead of numerical vectors for variables and parameters makes it possible to automatically perform a number of internal self-consistency checks as we will see.

In order to be considered valid, a formal model does not have to be only syntactically correct, but it must also satisfy a set of predefined plausible and common constraints. Formal languages are often domain-specific, thus allowing the notion of plausible and common constraints to be tailored to the specific domain. For example, as discussed below, static analysis checks on dependencies of kinetic laws on reactants in Bio-PEPA models will warn that simulations may produce negative results. A general-purpose programing language or numerical computing environment will never warn about this because negative results might be legitimate for some modeling problems in other domains.

Throughout the remainder we illustrate some of the concepts with features of the Bio-PEPA language, as an example of a text-based formalism, and Petri nets, as an example of a graphical formalism.

2.1 The Reagent and Reaction-Centric Views of a Model

A biochemical model can be viewed in one of two orthogonal ways which we call the reagent-centric and the reaction-centric views. In the former, for each reagent we list the set of reactions in which the reagent is involved. Conversely, the reaction-centric view displays, for each reaction, the set of reagents which are involved in that reaction and the associated effect that the reaction has on the population of that reagent. The BIOCHAM language [14] uses the reaction-centric view. In Bio-PEPA models are constructed in the reagent-centric view and the reaction-centric view is generated automatically by the software, in addition to some annotations on the reagent-centric view to be discussed below.

Providing both views of a model is important because they have complementary sets of advantages. The reagent-centric view is appropriate when scrutinizing the

behavior of a particular component. Biologists are often trained to read reaction definitions and this view can assist the detection of errors such as misplaced reactants or products.

This is a strength of high-level modeling languages such as process algebras and Petri nets: they give us more than one view of a model. In contrast, a programing language model such as a system of differential equations implemented in C or Python gives only a single view.

2.2 Missing and Unused Definitions

For a textual modeling language, such as Bio-PEPA, a straightforward static analysis check scrutinizes all definitions of names and considers if any are not subsequently used. There is also a converse check to ensure that any names used have indeed been defined. These checks are fast to perform and catch simple errors made by modelers, commonly misspelt names as well as missing definitions. Whenever a definition is missing, this is considered as an error from which the software cannot recover and evaluation of the model is disallowed. When a definition remains unused, the software can still evaluate the model, but presents the user with a clear warning that something is possibly wrong. For example, in the Bio-PEPA tools these checks are applied to definitions of rate constants, kinetic laws, and biochemical species.

Whenever a rate function or initial concentration uses a constant which lacks a definition this is likely to be an error on the part of the modeler, either they have forgotten to provide such a definition or the constant name has been misspelt at the point of use. Similarly, if a chemical species is given an initial concentration or molecule count but is not involved in any reaction as reactant, modifier, or product then the model is incomplete – perhaps a component definition has been forgotten. Alternatively a name which is misspelled at the point of use will be detected as a missing definition and an unused definition.

An unused constant definition is likely to be caused by an error in either a rate function or an initial concentration or molecule count. An unused rate function is possibly missing behavior in a species definition or perhaps the species definition is missing entirely. Finally, an unused species definition may signal that the modeler has forgotten to set the initial concentration or molecule count for one of the species in the model.

In the context of Petri nets, a graphical notation, a similar problem may arise if the model is not strongly connected, i.e., if there is not a directed path between every pair of nodes in the Petri net. A disconnected Petri net implies that there are two or more independent submodels. If static analysis reports that a model is not connected, this could be the desired model (in which case the distinct submodels can be analyzed independently) but it could be due to reactions or species omitted in error.

2.3 Kinetic Dependency Analysis

Although missing and unused definitions can catch some simple errors made by the user, some similar errors may escape such analysis. There are two basic analyzes which the Bio-PEPA Eclipse Plug-in performs over rate functions to catch further errors. The two analyzes performed check that any species P, whose population affects a given rate function r, has a corresponding behavior for r as a reactant, activator, inhibitor, or general modifier in the definition of P. When this is not the case, the software brings this to the attention of the modeler since it is likely that the modeler has forgotten to add the reaction-behavior to the species definition. Since we cannot know the role in which it should be added (reactant, activator, etc.) the modeler must be notified. The second analysis can be seen as the converse of this, it checks that for every reaction r for which a given species P is defined to be a reactant, activator, or inhibitor then the corresponding rate function for r includes a reference to the population of P. Again, if this is not the case then it is likely that the model erroneously includes reaction r behavior in the definition of P or has an incorrect definition for the rate function for r. These types of errors may lead to the amount of a species P undergoing inappropriate or insufficient updates as the model is exercised.

Kinetic dependency analysis does not guarantee that species are being used effectively in a kinetic law. For example, it is possible to construct pathological functions such as $(k \times E) + (P - P)$ which formally depends on the value of P because the symbol P occurs in the expression, but which uses P in such a way that its current value has no impact on the result of the function – which will always be equal to $k \times E$. It is not possible to detect all such false dependencies statically and so kinetic dependency analysis helps to detect when species have accidentally been omitted from function expressions but it can never provide a guarantee that their values have been used effectively.

2.4 Boundary Nodes, Sources, Sinks, and Input/Output Paths

Simple analysis of the model can determine its boundaries and its interactions with its environment. For example, a Petri net without boundary nodes is a self-contained closed system.

We can consider the interface to a model in terms of both reactions (transitions in a Petri net model) and species (places in a Petri net). In general an input to the model is termed a *source* while an output is termed a *sink*. A source reaction is a reaction which has no reactants and at least one product (e.g., synthesis reaction). A sink reaction is one which has no products and at least one reactant (e.g., degradations). The reaction $r_1 \stackrel{\text{def}}{=} P + S \longrightarrow$ is an example of a sink reaction as mass is consumed without any being produced.

Similarly, a species is considered a source if it is involved in at least one reaction as a consumed reactant and no reactions as a product. Conversely, a species sink is a species which is involved in at least one reaction as a product and no reactions as a consumed reactant. In Bio-PEPA the species S with definition: $S \stackrel{\text{def}}{=} r_1 \downarrow S + r_2 \downarrow S$ is a source species as it can be consumed by the reactions r_1 and r_2 but it is never produced.

The presence of boundary nodes is not in itself an error because these kinds of species and reactions are perfectly valid in general open systems, but they can also be the cause of unintended behavior. For instance, a Petri net with source species cannot be live and might lead to undesired deadlocks once the source node's initial amount is consumed. Similarly, a Petri net with source transitions cannot be bounded because its products could grow unboundedly.

The dual views offered by the Bio-PEPA software allow appropriate source/sink annotations to be added to reaction and species definitions. These annotations can provide useful error detection information to the modeler. Reaction source and sinks in particular denote that mass is not conserved by the model, although this may be intentional. Additionally the user can be warned if a source species has an initial population of zero, since in this case it will never have non-zero population and the reactions associated with it will never occur (assuming that the rate of the reactions do depend in a meaningful way on the population of the source species).

When boundary nodes are not caused by errors, but instead represent inputs or outputs with the environment, they can provide additional insight into the behavior of the model: their identification, supplemented with the minimal sequences of reactions that link a given source/sink combination, illustrates the flow of mass through the model as an input/output behavior. The input/output behavior informs modelers about which sources influence which sink, and what is the effect on the overall model of the sequence of reactions leading from source to sink. For instance, consider a signaling pathway that describes the signaling cascade which, starting from a constant influx of a ligand, leads to the production of one target protein. This will have a source action and one sink species. When dealing with complex interconnected pathways, the input/output behavior can help in understanding causal dependencies and in abstracting from the behavior of part of the system.

3 Structural Analysis

For graphical formalisms, such as Petri nets, a complementary method of investigating internal consistency is to view the model as a graph and consider it in terms of graph-theoretic concepts such as connectedness, reachability, paths, and cycles. An extensive body of work on structural analysis of models comes from Petri net-based techniques so we will discuss this type of analysis in Petri net terms.

Petri nets are graphical models of concurrent systems which contain *places* and *transitions*. Places contain *tokens* and firing a transition moves tokens from

one place to another place. In the context of biological modeling, a Petri net is an automaton whose places represent molecular species and whose transitions represent reactions transforming reactants into products. Places can only ever be connected to transitions, and transitions to places – thus every Petri net defines a *bipartite* graph. Arcs are weighted, and their weights specify the stoichiometric coefficients of reactants. Places contain an arbitrary number of tokens which represent the current molecule count of each biochemical species. The current state of the system, termed the *marking* of the Petri net, is given by the number of tokens on each place.

The behavior of a Petri net is defined by a firing rule, which specifies when transitions are enabled (if there are enough tokens for the involved reactants) and what is their effect (the changes in the number of tokens for the involved species).

Petri nets build on well-established mathematical foundations, which, in addition to the static and structural analysis techniques discussed here, support transient and steady-state analysis of the dynamic behavior. Reachability analysis can be used to identify parts of the model which are not connected; boundedness analysis can be used to ensure uncontrolled growth of molecules is not possible; and invariant analysis can identify violations of the law of conservation of mass. See [15] for more details and examples.

3.1 Structural Concepts in Petri Nets

A number of structural properties have been defined for Petri nets. All of these can be checked statically, because they are based solely on the structure of the Petri net without consideration of the initial marking. Here we give only an informal overview; for more details and their formal definitions see for instance [16]. These common properties are often valid for biological models. If they are not satisfied, it can be an indication of an error in the model specification (though not necessarily). Some concepts, such as, *pure* Petri nets and *ordinary* Petri nets, simply allow models to be categorized – this can provide a validity check to the modeler. For example, an ordinary Petri net is one in which all arcs are equal to 1 (which implies all stoichiometric coefficients are 1). Thus if the model is identified as being an ordinary Petri net but the model should include a homodimerization it is an indication that something is wrong. A pure Petri net is one in which there is no pair of nodes which are connected in both directions. This excludes models in which the same species is both a reactant and a product of a reaction.

Other concepts are related to the possible behaviors of the model when it is executed, i.e., when the Petri net is given a marking. For example a Petri net is considered to be *bounded* if the maximum number of tokens which can be on any place in the net is bounded by a constant. In biological terms this means that the amount of a species cannot grow without limit. A net is said to be *structurally bounded* if it is bounded for any initial marking. In some circumstances this property can be determined without executing the model and exploring its state space. A Petri

net is *conservative* if for each transition the sum of the weights of the incoming arcs is equal to the sum of weights of the outgoing arcs (i.e., the model does not contain any reaction which does not preserve the total number of molecules, such as complex formation reactions). Conservative Petri nets are always structurally bounded.

Another way of characterizing nets is in terms of the conflict and causality structures in operation within the model. For example a Petri net is termed *static conflict free* if there are no two transitions which share an input place. In terms of a biological model this means that there is no competition between reactions for reactants. Again, identifying such properties can be a source of validation – or otherwise – for the model.

3.2 Invariants

For biochemical models it is commonly the case that the modeler wishes to respect conservation of mass, so that the total quantity of matter at any time throughout the simulation is the same as the total quantity at the start. This can be characterized as an *invariant* of the model, a weighted sum of species quantities which remains constant.

Where the model contains source and/or sink reactions, mass will, of course, not be conserved. Such reactions are generally an abstraction, the products of a source reaction do not suddenly materialize from nothing and nor do the reactants of a sink reaction disintegrate into nothing. However the real reactants (of source reactions) or products (of sink reactions) are outside the scope of the model. A sink reaction could also represent a transportation to a place outside the scope of a model. For example, an intracellular model may represent a movement to the extracellular environment by the use of a sink reaction. Similarly, a movement from the extracellular environment into the cell may be represented by a source reaction.

However, even for models in which the entire mass of the system is not conserved, we expect there to be local conservation. When mass is conserved within a set of components we call this an invariant, since the sum of the values of all the members of the invariant – weighted by some suitable coefficients – will remain constant throughout the simulation of the model. Invariant analysis is a well-established technique of structural analysis of Petri nets, and can also be applied to textual notations such as Bio-PEPA.

The stoichiometric information about the reaction network, captured as the *incidence matrix* for a Petri net, defines a system of linear inequalities. *Fourier–Motzkin elimination* can find both real and integer solutions to such a system of linear inequalities. These solutions are the weights which are used in the invariants which hold over the species of the system. These are termed *P-invariants* in the context of Petri nets. We can compute a minimal generating set of invariants with a version of the Fourier–Motzkin method which produces only integer solutions [17]. The algorithm works on a matrix representation of the stoichiometric information

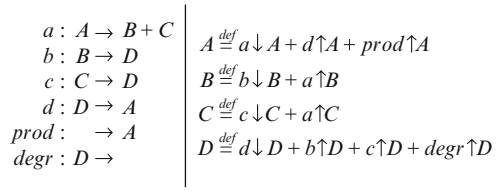


Fig. 27.1 An incorrect model which is not covered by invariants even if the source reaction 'prod' and the sink reaction 'degr' are ignored

for the model. This has as columns the reactions of the model and as rows the species. Each element reflects the change in population of the given species when the given reaction is fired.

When the matrix is transposed, such that, the rows become the reactions and the columns are the species, performing the same Fourier–Motzkin method over the transposed matrix we compute a new set of invariants. This new set of invariants is called the reaction invariants, or reaction loops (termed *T-invariants* in Petri nets). A reaction invariant consists of a set of reaction names with an integer coefficient associated with each reaction name. Such an invariant states that if this exact set of reactions is fired, the number of times indicated by each associated coefficient – in any order – then the model will be returned to the same state it was in at the start of the reaction invariant sequence.

The Bio-PEPA software computes both state and reaction invariants. Additionally the user can temporarily eliminate any of the reactions. This means that the chosen reactions are ignored for the purposes of the invariant analysis. In particular then the modeler may ignore all sink and source reactions in the model. If this is done then the entire set of species in the model should be covered by a list of invariants (which may be summed to create a single invariant which covers all of the species in the model). Where this is not the case, this indicates that somewhere in the model, mass is not conserved by a sequence of (non-source/sink) reactions. This probably indicates an error and the modeler should inspect their model carefully and either be able to repair it or explain why the conservation of mass is not observed.

Consider the model in Fig. 27.1. This model will not be covered by any state invariants, however it does contain two reaction loops: $a + b + c + d +$ degr and $a + b + c + (2 \times \text{degr}) +$ prod. The first of these is suspicious because it includes a sink reaction without a corresponding source reaction. Hence we employ our tactic of ignoring source and sink reactions and computing the set of invariants that this implies. This shows us that there are no invariants and this is highly indicative of an error in modeling.

In this particular case the correction could be that the reaction a should be modified to consume two molecules of A as in: $A + A \rightarrow B + C$ or the reaction d could be modified to consume two molecules of D as in: $D + D \rightarrow A$.

This is not an exhaustive list of possible corrections and in general there are many possible ways to correct this problem. However, the static analysis of invariant coverage has highlighted the possibility of an error and this hopefully will help to ensure that less time is spent analyzing incorrect models.

4 Verification of Behavioral Properties

As valuable as static and structural analysis are, there are some properties of models which cannot be assessed in this way. These instead require the model to be exercised to find all the possible configurations or states that it can reach. These correspond to the possible (although some of them may be unlikely) behaviors of the system. Such properties are termed *behavioral properties*, and generally are able to tell us more about the dynamics of the system.

One point of interest may be the extent to which behavior in the model persists, or whether it reaches a state where no further reactions are possible (often termed *deadlock*). In Petri nets this is known as *liveness*: A Petri net is *live* if, for every transition, it is possible from any state to reach a state where this transition is enabled. A live Petri net is deadlock-free (i.e., the corresponding system does not have any state where no reaction is possible).

Classical techniques for checking the behavior of models over a bounded state-space are considered in theoretical computer science under the heading of *model checking*. These techniques use efficient algorithms and data structures to determine whether logical formulae characterizing desired (or undesired) behavior are satisfied by a model. The presence of desired behavior shows that the model is live, and that sequences of reactions can lead the model to good states. The absence of undesired behavior shows that the model is safe, and that no sequence of reactions can lead the model to bad states. Both liveness and safety are desirable qualities for a model to possess.

Exact discrete-state model-checking where the complete state-space is generated can have applications in the modeling of biological processes (see [18]) but very often the memory needed to store the reachable state-space of a biological model exceeds the memory capacity of any computer system which we can access. Approximate statistical model-checking [19, 20] can be used instead and can give numerical results which are in very good agreement with those which are computed using more expensive techniques [21]. In this method, exhaustive generation of the reachable state-space is replaced by investigation of numerous trajectories over the state-space generated by simulation. Exact numerical solution of the underlying Markov chain is replaced by the execution of an ensemble of sufficiently many Monte Carlo simulations to approximate the measure of interest, and the probability of satisfaction of the logical formulae of interest is reported together with a confidence interval on the result.

4.1 Trace-Based Validation and Model-Checking

Novel techniques for detecting errors in models are now working not on the models themselves, but on their outputs generated through simulation. These complement static and structural methods beautifully because they consider the simulation results

which are the output from a modeling study, not the model used as input. Such a perspective can allow these techniques to detect errors in simulators, as well as errors in models.

The Traviando trace analyzer [22] is a discrete-event simulation trace analyzer which provides graphical techniques to inspect and manipulate simulation trace output and to compute statistical results. These results include counts of reaction events by category and can allow the modeler to discover that their simulation run has not been long enough to allow some reactions to fire, or to discover that they only fire a small number of times. Either of these might indicate an error in the model, caught by trace analysis.

Another novel and promising technique integrates model-checking and trace analysis and can be applied even in the continuous domain to the results of numerical integrators. Fages [23], Donaldson [24], and others describe model checkers which inspect simulation outputs and evaluate quantified logical formulae over a single simulation trace (in contrast to the statistical model-checking approach, which requires an ensemble of traces generated from Monte Carlo simulations conducted in the discrete molecular regime).

When working in the continuous domain the output of a model is a deterministic simulation utilizing continuous sure variables. This contrasts strongly with a stochastic simulation – which uses discrete random variables – because a single stochastic simulation run can be very far from the average-case behavior of the model, and thus conclusions drawn from a single stochastic simulation can rarely give definitive insights into behavior. In contrast, a time-course output from a continuous deterministic simulation returns sure trajectories for each of the chemical species in the model and thus carries more information which can be investigated in model-checking.

4.2 Equivalence Relations

Once we are working at the level of the state space generated by a model as well as verification of behavioral properties, we can also consider whether alternative models of the same system are in some sense equivalent. At the static level this may be carried out by identifying an isomorphism between the constructs of the model – essentially showing that models are equivalent because they are made up of equivalent components. However, at the underlying level of the state space, often termed the *labeled transition system*, much richer and more flexible notions of equivalence can be defined.

A *bisimulation* (and a semantic equivalence, more generally) is a way to assess whether two different labeled transition systems have the same behavior. A bisimulation is a symmetric relation between states of two labeled transition systems that requires that any two states in the relation both have transitions with similar enough behavior, as captured by the labels of transitions, and the states that are a result of a pair of transitions are again in the bisimulation relation [25]. This

captures the idea that the way the two models progress is the same for both models because states have transitions that match, as do the states that are targets of those transitions.

In the context of process algebras such as Bio-PEPA a variety of different bisimulations have been defined depending on what information about transitions is included in the labels [26]. In the original definition of bisimulation [25], equality between labels is required. However, it is possible to relax this requirement to allow for different notions of behavior, as exemplified by the work on g-bisimulation in Bio-PEPA [27]. This applies a function g to the labels on transitions and two labels are determined to be similar enough whenever the function g gives the same value for both labels. This is a powerful mechanism as it allows identification of selected reaction names, and selection of information about the reaction that the transition represents. The function g is chosen by the modeler to express the information of interest when considering behavior of the two systems. For example, a weak form of this equivalence relation, together with invariant analysis, is used to establish the equivalence of two previous models of the MAPK signaling cascade activated by EGF receptors [27–29].

5 Conclusions

As memorably expressed by Box and Draper [30], "all models are wrong, but some are useful". Models are built in order to help us to improve our understanding of systems and processes: if we already had perfect understanding then these models would not be needed. Our current imperfect understanding is necessarily encoded in the model. On top of this, all models simplify and abstract from details which are believed to be inessential in order that they can be tractable and usable for mathematical analysis. The belief that these details were inessential could be misplaced, and simply one part of our imperfect understanding. Because of this, and by their nature, we can never expect models of biological processes to be "correct".

However, we can – and should – expect our models to be consistent. If we intended to make a closed model where mass is conserved then we have an invariant which we expect to hold – computing the species invariants of our model allows us to check that mass is conserved and eliminates one possible source of error in our encoding. Similarly, if we have defined a reaction in our reaction network to involve species E and S then the kinetic law for the rate of that reaction should depend on E and S, and only on E and S. If not, then we have a likely source of error in the model.

Errors such as these may seem like simple carelessness but all of the evidence which we have seen seems to suggest that errors in the construction of formal models are very similar in nature to the errors which occur when writing a computer program. These are very rarely profound misunderstandings and are more likely to be simple mistakes such as a forgotten parameter or a forgotten function [31]. Nevertheless, the impact of such errors should not be underestimated.

Domain-specific modeling languages specialized to biological modeling are much more helpful in enabling simple errors to be caught automatically than are general-purpose numerical computing platforms. By building static analysis techniques into their modeling tools, domain-specific modeling languages can check that models are consistent, for every model, and for every revision of that model. State-of-the-art modeling platforms run inexpensive static analysis procedures every time that a model is saved after an edit has been made. This 'always-on' supervision helps modelers to find flaws in their models early, before analysis results are computed.

At best the methods which we have considered in this paper can help us to produce biological models which are internally consistent, with all parameters, kinetic laws, and species used in the way in which we intended. Internal consistency in our models can never make them right, but lack of consistency will make them wrong.

Acknowledgments Clark, Guerriero, Gilmore, and Hillston are supported by the Centre for Systems Biology at Edinburgh. The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and EPSRC, reference BB/D019621/1. The authors benefited from an introduction to invariant generation by Peter Kemper during his time as a SICSA Distinguished Visiting Fellow.

References

1. Clark A, Gilmore S, Hillston J, Kemper P (2010) Verification and testing of biological models. In: Proc winter simulation conference (WSC), December 2010, pp 620–630
2. Calder M, Duguid A, Gilmore S, Hillston J (2006) Stronger computational modelling of signalling pathways using both continuous and discrete-state methods. In: Proc 4th international conference on computational methods in systems biology (CMSB'06) (Lecture notes in computer science), vol 4210 , pp 63–77
3. Reisig W (1985) Petri nets: an introduction. Springer-Verlag, New York
4. Păun G (2002) Membrane computing: an introduction. Springer-Verlag, New York
5. Danos V, Laneve C (2004) Formal molecular biology. Theor Comput Sci 325(1):69–110
6. Priami C, Regev A, Silverman W, Shapiro E (2001) Application of a stochastic name-passing calculus to representation and simulation of molecular processes. Inf Process Lett 80(1):25–31
7. Ciocchetta F, Hillston J (2009) Bio-PEPA: a Framework for the Modelling and Analysis of Biological Systems. Theor Comput Sci 410(33–34):3065–3084
8. Kwiatkowski M, Stark I (2008) The continuous π-calculus: a process algebra for biochemical modelling. In: *Proc 6th international conference on computational methods in systems biology (CMSB'08)* (Lecture notes in computer science), no 5307. in, Springer-Verlag, Berlin Heidelberg, Germany, pp 103–122
9. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25):2340–2361
10. Duguid A, Gilmore S, Guerriero ML, Hillston J, Loewe L (2009) Design and development of software tools for Bio-PEPA. In: Proc winter simulation conference (WSC'09). IEEE Press, Piscataway, NJ, USA, pp 956–967
11. Python Programming Language Official Website (2011) www.python.org
12. Rohr C, Marwan W, Heiner M (2010) Snoopy, a unifying Petri net framework to investigate biomolecular networks. Bioinformatics 26(7):974–975

13. MathWorks MATLAB (2011) http://www.mathworks.com/.
14. Calzone L, Fages F, Soliman S (2006) BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge. Bioinformatics 22(14):1805–1807
15. Peleg M, Yeh I, Altman R (2002) Modeling biological processes using Workflow and Petri Net models. Bioinformatics 18(6):825–837
16. Heiner M, Gilbert D, Donaldson R (2008) Petri Nets for Systems and Synthetic Biology. In: Proc formal methods for computational systems biology (SFM'08) (Lecture notes in computer science), vol 5016. Springer-Verlag, Berlin Heidelberg, Germany, pp 215–264
17. Martinez J, Manual Silva (1982) A simple and fast algorithm to obtain all invariants of a generalized petri net. In: Proc selected papers from the first and the second european workshop on application and theory of petri nets. Springer-Verlag, London, UK, pp 301–310
18. Guerriero ML (2009) Qualitative and quantitative analysis of a Bio-PEPA model of the Gp130/JAK/STAT signalling pathway. Trans Comput Syst Biol XI 5750:90–115
19. Sen K, Viswanathan M, Agha G (2005) On statistical model checking of stochastic systems. In: Proc computer aided verification (Lecture notes in computer science), vol 3576. Springer-Verlag, Berlin Heidelberg, Germany, pp 266–280
20. Younes HLS, Simmons RG (2006) Statistical probabilistic model checking with a focus on time-bounded properties. Inf Comput 204(9):1368–1409
21. Ciocchetta F, Gilmore S, Guerriero ML, Hillston J (2009) Integrated simulation and model-checking for the analysis of biochemical systems. In: Proc 3rd international workshop on practical applications of stochastic modelling (PASM'08) (Electronic notes in theoretical computer science), vol 232. Elsevier, Oxford UK, pp 17–38
22. Kemper P, Tepper C (2009) Automated trace analysis of discrete-event system models. IEEE Trans Software Eng 35(2):195–208
23. Fages F, Rizk A (2007) On the analysis of numerical data time series in temporal logic. In: Proc computational methods in systems biology (CMSB'07) (Lecture notes in computer science), vol 4695. Springer, Berlin, Heidelberg, Germany, pp 48–63
24. Donaldson R, Gilbert D (2008) A model checking approach to the parameter estimation of biochemical pathways. In: Proc computational methods in systems biology (CMSB'08) (Lecture notes in computer science), vol 5307. Springer, Berlin, Heidelberg, Germany, pp 269–287
25. Milner R (1989) Communication and concurrency. Prentice Hall, Hemel Hempstead, UK
26. Galpin V, Hillston J (2011) A semantic equivalence for Bio-PEPA based on discretisation of continuous values. Theor Comput Sci 412(21):2142–2161
27. Galpin V (2011) Equivalences for a biological process algebra. Theor Comput Sci 412(43):6058–6082, DOI 10.1016/j.tcs.2011.07.006
28. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol 20:270–375
29. Gong Y, Zhao X (2003) Shc-dependent pathway is redundant but dominant in MAPK cascade activation by EGF receptors: a modeling inference. FEBS Lett 554:467–472
30. Box G, Draper N (1987) Empirical model-building and response surfaces. Wiley Series in Probability and Mathematical Statistics, USA
31. Knuth DE (1989) The errors of TEX. Software Pract Exper 19:607–685

Chapter 28
Global Parameter Identification of Stochastic Reaction Networks from Single Trajectories

Christian L. Müller*, Rajesh Ramaswamy*, and Ivo F. Sbalzarini

Abstract We consider the problem of inferring the unknown parameters of a stochastic biochemical network model from a single measured time-course of the concentration of some of the involved species. Such measurements are available, e.g., from live-cell fluorescence microscopy in image-based systems biology. In addition, fluctuation time-courses from, e.g., fluorescence correlation spectroscopy (FCS) provide additional information about the system dynamics that can be used to more robustly infer parameters than when considering only mean concentrations. Estimating model parameters from a single experimental trajectory enables single-cell measurements and quantification of cell–cell variability. We propose a novel combination of an adaptive Monte Carlo sampler, called Gaussian Adaptation (GaA), and efficient exact stochastic simulation algorithms (SSA) that allows parameter identification from single stochastic trajectories. We benchmark the proposed method on a linear and a non-linear reaction network at steady state and during transient phases. In addition, we demonstrate that the present method also provides an ellipsoidal volume estimate of the viable part of parameter space and is able to estimate the physical volume of the compartment in which the observed reactions take place.

1 Introduction

Systems biology implies a holistic research paradigm, complementing the reductionist approach to biological organization [15, 16]. This frequently has the goal of mechanistically understanding the function of biological entities and processes in

*Authors C.L. Müller and R. Ramaswamy contributed equally to this work.

C.L. Müller • R. Ramaswamy • I.F. Sbalzarini (✉)
Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics,
ETH Zurich, CH–8092 Zurich, Switzerland
e-mail: christian.mueller@inf.ethz.ch; rajeshr@ethz.ch; ivos@ethz.ch

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_28,

interaction with the other entities and processes they are linked to or communicate with. A formalism to express these links and connections is provided by network models of biological processes [1, 4]. Using concepts from graph theory [26] and dynamic systems theory [44], the organization, dynamics, and plasticity of these networks can then be studied.

Systems biology models of molecular reaction networks contain a number of parameters. These are the rate constants of the involved reactions and, if spatiotemporal processes are considered, the transport rates, e.g., diffusion constants, of the chemical species. In order for the models to be predictive, these parameters need to be inferred. The process of inferring them from experimental data is called *parameter identification*. If in addition also the network structure is to be inferred from data, the problem is called *systems identification*. Here, we consider the problem of identifying the parameters of a biochemical reaction network from a single, noisy measurement of the concentration time-course of some of the involved species. While this time series can be long, ensemble replicas are not possible, either because the measurements are destructive or one is interested in variations between different specimens or cells. This is particularly important in *molecular systems biology*, where cell–cell variations are of interest or large numbers of experimental replica are otherwise not feasible.

This problem is particularly challenging and traditional genomic and proteomic techniques do not provide single-cell resolution. Moreover, in individual cells the molecules and chemical reactions can only be observed indirectly. Frequently, fluorescence microscopy is used to observe biochemical processes in single cells. Fluorescently tagging some of the species in the network of interest allows measuring the spatiotemporal evolution of their concentrations from video microscopy and fluorescence photometry. In addition, fluorescence correlation spectroscopy (FCS) allows measuring fluctuation time-courses of molecule numbers [23].

Using only a single trajectory of the mean concentrations would hardly allow identification of network parameters. There could be several combinations of network parameters that lead to the same mean dynamics. A stochastic network model, however, additionally provides information about the fluctuations of the molecular abundances. The hope is that there is then only a small region of parameter space that produces the correct behavior of the mean *and* the correct spectrum of fluctuations [31]. Experimentally, fluctuation spectra can be measured at single-cell resolution using FCS.

The stochastic behavior of biochemical reaction networks can be due to low copy numbers of the reacting molecules [10, 39]. In addition, biochemical networks may exhibit stochasticity due to extrinsic noise. This can persist even at the continuum scale, leading to continuous–stochastic models. Extrinsic noise can, e.g., arise from environmental variations or variations in how the reactants are delivered into the system. Also measurement uncertainties can be accounted for in the model as extrinsic noise, modeling our inability to precisely quantify the experimental observables.

We model stochastic chemical kinetics using the chemical master equation (CME). Using a CME forward model in biological parameter identification amounts to tracking the evolution of a probability distribution, rather than just of a single value. This prohibits predicting the state of the system and only allows statements about the probability for the system to be in a certain state, hence requiring sampling-based parameter identification methods. In the stochastic–discrete context, a number of different approaches have been suggested. Boys et al. proposed a fully Bayesian approach for parameter estimation using an explicit likelihood for data/model comparison and a Markov Chain Monte Carlo (MCMC) scheme for sampling [5]. Zechner et al. developed a recursive Bayesian estimation technique [45] to cope with cell–cell variability in experimental ensembles. Toni and co-workers used an approximate Bayesian computation (ABC) ansatz, as introduced by Marjoram and co-workers [25], that does not require an explicit likelihood [43]. Instead, sampling is done in a sequential Monte Carlo (or particle filter) framework. Reinker et al. used a hidden Markov model where the hidden states are the actual molecule abundances, and state transitions model chemical reactions [40]. Inspired by Prediction Error Methods [24], Cinquemani et al. identified the parameters of a hybrid deterministic–stochastic model of gene expression from multiple experimental time courses [7]. Randomized optimization algorithms have been used, e.g., by Koutroumpas et al. who applied a Genetic Algorithm to a hybrid deterministic–stochastic network model [21]. More recently, Poovathingal and Gunawan used another global optimization heuristic, the Differential Evolution algorithm [32]. A variational approach for stochastic two-state systems has been proposed by Stock and co-workers based on Maximum Caliber [41], an extension of Jaynes' Maximum Entropy principle [14] to non-equilibrium systems. If estimates are to be made based on a single trajectory, the stochasticity of the measurements and of the model leads to noisy similarity measures, requiring optimization and sampling schemes that are robust against noise in the data.

Here, we propose a novel combination of exact stochastic simulations for a CME forward model and an adaptive Monte Carlo sampling technique, called Gaussian Adaptation (GaA), to address the single-trajectory parameter estimation problem for monostable stochastic biochemical reaction networks. Evaluations of the CME model are done using exact partial-propensity stochastic simulation algorithms (SSA) [35]. Parameter optimization uses GaA. The method iteratively samples model parameters from a multivariate normal distribution and evaluates a suitable objective function that measures the distance between the dynamics of the forward model output and the experimental measurements. In addition to estimates of the kinetic parameters in the network, the present method also provides an ellipsoidal volume estimate of the viable part of parameter space and is able to estimate the physical volume of the compartment in which the reactions take place.

We assume that quantitative experimental time series of either a transient or the steady state of the concentrations of some of the molecular species in the network are available. This can, for example, be obtained from single-cell fluorescence microscopy by translating fluorescence intensities to estimated chemical

concentrations. Accurate methods that account for the microscope's point-spread function and the camera noise model are available to this end [6, 12, 13]. Additionally, FCS spectra can be analyzed in order to quantify molecule populations, their intrinsic fluctuations, and lifetimes [23, 34, 39]. The present approach requires only a *single* stochastic trajectory from each cell. Since the forward model is stochastic and only a single experimental trajectory is used, the objective function needs to robustly measure closeness between the experimental and the simulated trajectories. We review previously considered measures and present a new distance function in Sect. 5. First, however, we set out the formal stochastic framework and problem description below. We then describe GaA and its applicability to the current estimation task. The evaluation of the forward model is outlined in Sect. 4. We consider a linear cyclic chain and a non-linear colloidal aggregation model as benchmark test cases in Sect. 6 and conclude in Sect. 7.

2 Background and Problem Statement

We consider a network model of a biochemical system given by M coupled chemical reactions

$$\sum_{i=1}^{N} \nu_{i,j}^{-} \mathrm{S}_i \xrightarrow{k_j} \sum_{i=1}^{N} \nu_{i,j}^{+} \mathrm{S}_i \quad \forall\, j = 1, \ldots, M \tag{28.1}$$

between N species, where $\underline{\underline{\nu}}^{-} = [\nu_{i,j}^{-}]$ and $\underline{\underline{\nu}}^{+} = [\nu_{i,j}^{+}]$ are the stoichiometry matrices of the reactants and products, respectively, and S_i is the ith species in the reaction network. Let n_i be the population (molecular copy number) of species S_i. The reactions occur in a physical volume Ω and the macroscopic reaction rate of reaction j is k_j. This defines a dynamic system with integer-valued state $\underline{n}(t) = [n_i(t)]$ and $M + 1$ parameters $\underline{\theta} = [k_1, \ldots, k_M, \Omega]$.

The state of such a system can be interpreted as a realization of a random variable $\underline{n}(t)$ that changes over time t. Every one can know about the system is the probability for it to be in a certain state at a certain time t_j given the system's state history, hence

$$\begin{aligned} &P(\underline{n}(t_j) \,|\, \underline{n}(t_{j-1}), \ldots, \underline{n}(t_1), \underline{n}(t_0))\, \mathrm{d}^N n \\ &\quad = \mathrm{Prob}\{\underline{n}(t_j) \in [\underline{n}(t_j),\, \underline{n}(t_j) + \mathrm{d}\underline{n}) \,|\, \underline{n}(t_i), \quad i = 0, \ldots, j-1\}. \end{aligned} \tag{28.2}$$

A frequently made model assumption, substantiated by physical reasoning, is that the probability of the current state depends solely on the previous state, i.e.,

$$P(\underline{n}(t_j) \,|\, \underline{n}(t_{j-1}), \ldots, \underline{n}(t_1), \underline{n}(t_0)) = P(\underline{n}(t_j) \,|\, \underline{n}(t_{j-1})). \tag{28.3}$$

The system is then modeled as a first-order Markov chain where the state $\underline{n}$ evolves as:

$$\underline{n}(t + \Delta t) = \underline{n}(t) + \underline{\Xi}(\Delta t; \underline{n}, t). \tag{28.4}$$

This is the *equation of motion* of the system. If $\underline{n}$ is real-valued, it defines a continuous–stochastic model in the form of a continuous-state Markov chain. Discrete $\underline{n}$, as is the case in chemical kinetics, amount to discrete–stochastic models expressed as discrete-state Markov chains. The Markov propagator $\underline{\Xi}$ is itself a random variable, distributed with probability distribution $\Pi(\underline{\xi} \mid \Delta t; \underline{n}, t) = P(\underline{n} + \underline{\xi}, t + \Delta t \mid \underline{n}, t)$ for the state change $\underline{\xi}$. For continuous-state Markov chains, Π is a continuous probability density function (PDF), for discrete-state Markov chains a discrete probability distribution. If $\Pi(\underline{\xi}) = \delta(\underline{\xi} - \underline{\xi}_0)$, with δ the Dirac delta distribution, then the system's state evolution becomes deterministic with predictable discrete or continuous increments $\underline{\xi}_0$. Deterministic models can hence be interpreted as a limit case of stochastic models [22].

In chemical kinetics, the probability distribution Π of the Markov propagator is a linear combination of Poisson distributions with weights given by the reaction stoichiometry. This leads to the equation of motion for the population $\underline{n}$ given by

$$\underline{n}(t + \Delta t) = \underline{n}(t) + (\underline{\underline{\nu}}^{+} - \underline{\underline{\nu}}^{-}) \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_M \end{bmatrix}, \tag{28.5}$$

where $\psi_i \sim \mathscr{P}(a_i(\underline{n}(t))\Delta t)$ is a random variable from the Poisson distribution with rate $\lambda = a_i(\underline{n}(t))\Delta t$. The second term on the right-hand side of (28.5) follows a probability distribution $\Pi(\underline{\xi} \mid \Delta t; \underline{n}, t)$ whose explicit form is analytically intractable in the general case. The rates a_j, $j = 1, \ldots M$, are called the *reaction propensities* and are defined as:

$$a_j = \prod_{i=1}^{N} \binom{n_i}{\nu_{i,j}^{-}} \frac{k_j}{\Omega^{1+\sum_{i'=1}^{N} \nu_{i',j}^{-}}}. \tag{28.6}$$

They depend on the macroscopic reaction rates and the reaction volume and can be interpreted as the probability rates of the respective reactions. Advancing (28.5) with a Δt such that more than one reaction event happens per time step yields an approximate simulation of the biochemical network as done in approximate SSA [3, 9].

An alternative approach consists in considering the evolution of the state probability distribution $P(\underline{n}, t \mid \underline{n}_0, t_0)$ of the Markov chain described by (28.5), hence:

$$\frac{\partial P}{\partial t} = \sum_{j=1}^{M} \left(\prod_{i=1}^{N} \mathrm{E}_i^{\nu_{i,j}^{-}} \mathrm{E}_j^{-\nu_{i,j}^{+}} - 1 \right) a_j(\underline{n}(t)) P(\underline{n}, t) \tag{28.7}$$

with the *step operator* $\mathrm{E}_i^p f(\underline{n}) = f\left(\underline{n} + p\hat{\underline{i}}\right)$ for any function f, where $\hat{\underline{i}}$ is the N-dimensional unit vector along the ith dimension. This equation is called the *CME*. Directly solving it for P is analytically intractable, but trajectories of the Markov chain governed by the unknown state probability P can be sampled using exact SSA [8]. Exact SSAs are exact in the sense that they sample Markov chain realizations from the exact solution P of the CME, without ever explicitly computing this solution. Since SSAs are Monte Carlo algorithms, however, a sampling error remains.

Assuming that the population $\underline{n}$ increases with the volume Ω, $\underline{n}$ can be approximated as a continuous random variable in the limit of large volumes, and (28.5) becomes

$$\underline{n}(t + \Delta t) = \underline{n}(t) + (\underline{\underline{\nu}}^+ - \underline{\underline{\nu}}^-) \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_M \end{bmatrix}, \tag{28.8}$$

where $\eta_i \sim \mathcal{N}(a_i(\underline{n}(t))\Delta t, a_i(\underline{n}(t))\Delta t)$ are normally distributed random variables. The second term on the right-hand side of this equation is a random variable, that is, distributed according to the corresponding Markov propagator $\Pi(\underline{\xi} \mid \Delta t; \underline{n}, t)$, which is a Gaussian. Equation (28.8) is called the *chemical Langevin equation* with Π given by:

$$\Pi\left(\underline{\xi} \mid \Delta t; \underline{n}, t\right) = (2\pi)^{-N/2} \left|\underline{\underline{\Sigma}}\right|^{-1/2} \mathrm{e}^{-\frac{1}{2}(\underline{\xi}-\underline{\mu})^{\mathrm{T}} \underline{\underline{\Sigma}}^{-1} (\underline{\xi}-\underline{\mu})}, \tag{28.9}$$

where

$$\underline{\mu} = \Delta t \left(\underline{\underline{\nu}}^+ - \underline{\underline{\nu}}^-\right) \begin{bmatrix} a_1(\underline{n}(t)) \\ \vdots \\ a_M(\underline{n}(t)) \end{bmatrix} \text{ and } \underline{\underline{\Sigma}} = \Delta t \left(\underline{\underline{\nu}}^+ - \underline{\underline{\nu}}^-\right) \operatorname{diag} \underline{a}(\underline{n}(t)) \left(\underline{\underline{\nu}}^+ - \underline{\underline{\nu}}^-\right)^{\mathrm{T}}.$$

The corresponding equation for the evolution of the state PDF is the non-linear Fokker–Planck equation, given by:

$$\frac{\partial P}{\partial t} = \underline{\nabla}^{\mathrm{T}} \left(\frac{1}{2} \underline{\underline{D}}\, \underline{\nabla} - \underline{F}\right) P(\underline{n}, t), \tag{28.10}$$

where

$$\underline{\nabla}^{\mathrm{T}} = \left[\frac{\partial}{\partial n_1}, \ldots, \frac{\partial}{\partial n_N}\right], \tag{28.11}$$

$$F_i = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{+\infty} \mathrm{d}\xi_i \, \xi_i \, \Pi(\underline{\xi} \mid \Delta t; \underline{n}, t), \tag{28.12}$$

and

$$D_{ij} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathrm{d}\xi_i \mathrm{d}\xi_j \, \xi_i \, \xi_j \, \Pi(\underline{\xi} \mid \Delta t; \underline{n}, t) - F_i F_j . \quad (28.13)$$

At much larger Ω, when the population $\underline{n}$ is on the order of Avogadro's number, (28.8) can be further approximated as:

$$\underline{n}(t + \Delta t) = \underline{n}(t) + \left(\underline{\underline{\nu}}^{+} - \underline{\underline{\nu}}^{-}\right) \begin{bmatrix} \phi_1(\underline{n}(t))\Delta t \\ \vdots \\ \phi_M(\underline{n}(t))\Delta t \end{bmatrix}, \quad (28.14)$$

where $\phi_j(\underline{n}) = k_j \Omega^{1-\sum_{i'=1}^{N} \nu_{i',j}^{-}} \prod_{i=1}^{N} n_i^{\nu_{i,j}^{-}} (\nu_{i,j}^{-}!)^{-1}$. Note that the second term on the right-hand side of this equation is a random variable whose probability distribution is the Dirac delta

$$\Pi\left(\underline{\xi} \mid \Delta t; \underline{n}, t\right) = \delta\left(\underline{\xi} - \left(\underline{\underline{\nu}}^{+} - \underline{\underline{\nu}}^{-}\right) \begin{bmatrix} \phi_1(\underline{n}(t))\Delta t \\ \vdots \\ \phi_M(\underline{n}(t))\Delta t \end{bmatrix}\right). \quad (28.15)$$

Equation (28.14) hence is a deterministic equation of motion. In the limit $\Delta t \to 0$ this equation can be written as the ordinary differential equation

$$\frac{\mathrm{d}\underline{x}}{\mathrm{d}t} = \left(\underline{\underline{\nu}}^{+} - \underline{\underline{\nu}}^{-}\right) \begin{bmatrix} \phi_1(\underline{x}(t)) \\ \vdots \\ \phi_M(\underline{x}(t)) \end{bmatrix} \quad (28.16)$$

for the *concentration* $\underline{x} = \underline{n}\Omega^{-1}$. This is the classical *reaction rate equation* for the system in (28.1).

By choosing the appropriate probability distribution Π of the Markov propagator, one can model reaction networks in different regimes: small population $\underline{n}$ (small Ω) using SSA over (28.7), intermediate population (intermediate Ω) using (28.8), and large population (large Ω) using (28.16). The complete model definition therefore is $\mathscr{M}(\underline{\theta}) = \left\{\underline{\underline{\nu}}^{-}, \underline{\underline{\nu}}^{+}, \Pi\right\}$.

The problem considered here can then be formalized as follows: Given a forward model $\mathscr{M}(\underline{\theta})$ and a single noisy trajectory of the population of the chemical species $\hat{\underline{n}}(t_0 + (q-1)\Delta t_{\mathrm{exp}})$ at K discrete time points $t = t_0 + (q-1)\Delta t_{\mathrm{exp}}$, $q = 1, \ldots, K$, we wish to infer $\underline{\theta} = [k_1, \ldots, k_M, \Omega]$. The time between two consecutive measurements Δt_{exp} and the number of measurements K are given by the experimental technique used. As a forward model we use the full CME as given in (28.7) and sample trajectories from it using the partial-propensity formulation of Gillespie's exact SSA as described in Sect. 4.

3 Gaussian Adaptation for Global Parameter Optimization, Approximate Bayesian Computation, and Volume Estimation

Gaussian Adaptation, introduced in the late 1960s by Gregor Kjellström [17,19], is a Monte Carlo technique that has originally been developed to solve design-centering and optimization problems in analog electric circuit design. Design-centering solves the problem of determining the nominal values (resistances, capacitances, etc.) of the components of a circuit such that the circuit output is within specified design bounds and is maximally robust against random variations in the circuit components with respect to a suitable criterion or objective function. This problem is a superset of general optimization, where one is interested in finding a parameter vector that minimizes (or maximizes) an objective function without any additional robustness criterion. GaA has been specifically designed for scenarios where the objective function $f(\underline{\theta})$ is only available in a black-box (or oracle) model that is defined on a real-valued domain $\mathscr{A} \subseteq \mathbb{R}^n$ and returns scalar real-valued output. The black-box model assumes that gradients or higher-order derivatives of the objective function may not exist or may not be available, hence including the class of discontinuous and noisy functions. The specific objective function used here is presented in Sect. 5.

The principle idea behind GaA is the following: Starting from a user-defined point in parameter space, GaA explores the space by iteratively sampling single parameter vectors from a multivariate Gaussian distribution $\mathscr{N}(\underline{m}, \underline{\underline{\Sigma}})$ whose mean $\underline{m} \in \mathbb{R}^n$ and covariance matrix $\underline{\underline{\Sigma}} \in \mathbb{R}^{n \times n}$ are dynamically adapted based on the information from previously accepted samples. The acceptance criterion depends on the specific mode of operation, i.e., whether GaA is used as an optimizer or as a sampler [27,28]. Adaptation is performed such as to maximize the entropy of the search distribution under the constraint that acceptable search points are found with a predefined, fixed hitting (success) probability $p < 1$ [19]. Using the definition of the entropy of a multivariate Gaussian distribution $\mathscr{H}(\mathscr{N}) = \log\left(\sqrt{(2\pi \mathrm{e})^n \det(\underline{\underline{\Sigma}})}\right)$ shows that this is equivalent to maximizing the determinant of the covariance matrix $\underline{\underline{\Sigma}}$. GaA thus follows Jaynes' Maximum Entropy principle [14].

GaA starts by setting the mean $\underline{m}^{(0)}$ of the multivariate Gaussian to an initial acceptable point $\underline{\theta}^{(0)}$ and the Cholesky factor $\underline{\underline{Q}}^{(0)}$ of the covariance matrix to the identity matrix $\underline{\underline{I}}$. At each iteration $g > 0$, the covariance $\underline{\underline{\Sigma}}^{(g)}$ is decomposed as: $\underline{\underline{\Sigma}}^{(g)} = \left(r \cdot \underline{\underline{Q}}^{(g)}\right)\left(r \cdot \underline{\underline{Q}}^{(g)}\right)^{\mathrm{T}} = r^2 \left(\underline{\underline{Q}}^{(g)}\right)\left(\underline{\underline{Q}}^{(g)}\right)^{\mathrm{T}}$, where r is the scalar step size that controls the scale of the search. The matrix $\underline{\underline{Q}}^{(g)}$ is the normalized square root of $\underline{\underline{\Sigma}}^{(g)}$, found by eigen- or Cholesky decomposition of $\underline{\underline{\Sigma}}^{(g)}$. The candidate parameter

vector in iteration $g+1$ is sampled from a multivariate Gaussian according to $\underline{\theta}^{(g+1)} = \underline{m}^{(g)} + r^{(g)}\underline{\underline{Q}}^{(g)}\underline{\eta}^{(g)}$, where $\underline{\eta}^{(g)} \sim \mathcal{N}(\underline{0}, \underline{I})$. The parameter vector is then evaluated by the objective function $f(\underline{\theta}^{(g+1)})$.

Only if the parameter vector is accepted, the following adaptation rules are applied: The step size r is increased as $r^{(g+1)} = f_\mathrm{e} \cdot r^{(g)}$, where $f_\mathrm{e} > 1$ is termed the *expansion factor*. The mean of the proposal distribution is updated as:

$$\underline{m}^{(g+1)} = \left(1 - \frac{1}{N_\mathrm{m}}\right)\underline{m}^{(g)} + \frac{1}{N_\mathrm{m}}\underline{\theta}^{(g+1)}. \tag{28.17}$$

N_m is a weighting factor that controls the learning rate of the method. The successful search *direction* $\underline{d}^{(g+1)} = \left(\underline{\theta}^{(g+1)} - \underline{m}^{(g)}\right)$ is used to perform a rank-one update of the covariance matrix: $\underline{\underline{\Sigma}}^{(g+1)} = \left(1 - \frac{1}{N_\mathrm{C}}\right)\underline{\underline{\Sigma}}^{(g)} + \frac{1}{N_\mathrm{C}}\underline{d}^{(g+1)}\underline{d}^{(g+1)\mathrm{T}}$. N_C weights the influence of the accepted parameter vector on the covariance matrix. In order to decouple the volume of the covariance (controlled by $r^{(g+1)}$) from its orientation, $\underline{\underline{Q}}^{(g+1)}$ is normalized such that $\det\left(\underline{\underline{Q}}^{(g+1)}\right) = 1$.

In case $\underline{\theta}^{(g+1)}$ is not accepted at the current iteration, only the step size is adapted as $r^{(g+1)} = f_\mathrm{c} \cdot r^{(g)}$, where $f_\mathrm{c} < 1$ is the *contraction factor*.

The behavior of GaA is controlled by several strategy parameters. Kjellström analyzed the information-theoretic optimality of the acceptance probability p for GaA in general regions [19]. He concluded that the efficiency E of the process and p are related as $E \propto -p \log p$, leading to an optimal $p = \frac{1}{\mathrm{e}} \approx 0.3679$, where e is Euler's number. A proof is provided in [18]. Maintaining this optimal hitting probability corresponds to leaving the volume of the distribution, measured by $\det\left(\underline{\underline{\Sigma}}\right)$, constant under stationary conditions. Since $\det\left(\underline{\underline{\Sigma}}\right) = r^{2n}\det\left(\underline{\underline{Q}}\,\underline{\underline{Q}}^{\mathrm{T}}\right)$, the expansion and contraction factors f_e and f_c expand or contract the volume by a factor of f_e^{2n} and f_c^{2n}, respectively. After S accepted and F rejected samples, a necessary condition for constant volume thus is: $\prod_{i=1}^{S}(f_\mathrm{e})^{2n}\prod_{i=1}^{F}(f_\mathrm{c})^{2n} = 1$. Using $p = \frac{S}{S+F}$, and introducing a small $\beta > 0$, the choice $f_\mathrm{e} = 1 + \beta(1-p)$ and $f_\mathrm{c} = 1 - \beta p$ satisfies the constant-volume condition to first order. The scalar rate β is coupled to N_C. N_C influences the update of $\underline{\underline{\Sigma}} \in \mathbb{R}^{n\times n}$, which contains n^2 entries. Hence, N_C should be related to n^2. We suggested using $N_\mathrm{C} = (n+1)^2/\log(n+1)$ as a standard value, and coupling $\beta = \frac{1}{N_\mathrm{C}}$ [29]. A similar reasoning is also applied to N_m. Since N_m influences the update of $\underline{m} \in \mathbb{R}^n$, it is reasonable to set $N_\mathrm{m} \propto n$. We propose $N_\mathrm{m} = \mathrm{e}n$ as a standard value.

Depending on the specific acceptance rule used, GaA can be turned into a global optimizer [29], an adaptive MCMC sampler [27,28], or a volume estimation method [30], as described next.

3.1 GaA for Global Black-Box Optimization

In a minimization scenario, GaA uses an adaptive-threshold acceptance mechanism. Given an initial scalar cutoff threshold $c_{\mathrm{T}}^{(0)}$, we accept a parameter vector $\underline{\theta}^{(g+1)}$ at iteration $g+1$ if $f\left(\underline{\theta}^{(g+1)}\right) < c_{\mathrm{T}}^{(g)}$. Upon acceptance, the threshold c_{T} is lowered as $c_{\mathrm{T}}^{(g+1)} = \left(1 - \frac{1}{N_{\mathrm{T}}}\right) c_{\mathrm{T}}^{(g)} + \frac{1}{N_{\mathrm{T}}} f\left(\underline{\theta}^{(g+1)}\right)$, where N_{T} controls the weighting between the old threshold and the objective-function value of the *accepted* sample. This sample-dependent threshold update renders the algorithm invariant to linear transformations of the objective function. The standard strategy parameter value is $N_{\mathrm{T}} = e n$ [28]. We refer to [28] for further information about convergence criteria and constraint handling techniques in GaA.

3.2 GaA for Approximate Bayesian Computation and Viable Volume Estimation

Replacing the threshold acceptance-criterion by a probabilistic Metropolis criterion, and setting $N_{\mathrm{m}} = 1$, turns GaA into an adaptive MCMC sampler with global adaptive scaling [2]. We termed this method *Metropolis-GaA* [27, 28]. Its strength is that GaA can automatically adapt to the covariance of the target probability distribution while maintaining the fixed hitting probability. For standard MCMC, this cannot be achieved without fine-tuning the proposal using multiple MCMC runs. We hypothesize that GaA might also be an effective tool for ABC [43]. In essence, the ABC ansatz is MCMC without an explicit likelihood function [25]. The likelihood is replaced by a distance function – which plays the same role as our objective function – that measures closeness between a parameterized model simulation and empirical data $\mathscr{D}$, or summary statistics thereof. When a uniform prior over the parameters and a symmetric proposal are assumed, a parameter vector in ABC is unconditionally accepted if its corresponding distance function value $f(\underline{\theta}^{(g+1)}) < c_{\mathrm{T}}$ [25]. The threshold c_{T} is a problem-dependent constant that is fixed prior to the actual computation. Marjoram and co-workers have shown that samples obtained in this manner are approximately drawn from the posterior parameter distribution given the data $\mathscr{D}$. While Pritchard et al. used a simple rejection sampler [33], Marjoram and co-workers proposed a standard MCMC scheme [25]. Toni and co-workers used sequential MC for sample generation [43]. To the best of our knowledge, however, the present work presents the first application of an adaptive MCMC scheme for ABC in biochemical network parameter inference. Finally, we emphasize that when GaA's mean, covariance matrix, and hitting probability p stabilize during ABC, they provide direct access to an ellipsoidal estimation of the volume of the viable parameter space as defined by the threshold c_{T} [30]. Hafner and co-workers have shown how to use such viable volume estimates for model discrimination [11].

4 Evaluation of the Forward Model

In each iteration of the GaA algorithm, the forward model of the network needs to be evaluated for the proposed parameter vector $\underline{\theta}$. This requires an efficient and exact SSA for the chemical kinetics of the reaction network, used to generate trajectories $\underline{n}(t)$ from $\mathscr{M}(\underline{\theta})$. Since GaA could well propose parameter vectors that lead to low copy numbers for some species, it is important that the SSA be exact since approximate algorithms are not appropriate at low copy number.

In its original formulation, Gillespie's SSA has a computational cost that is linearly proportional to the total number M of reactions in the network. If many model evaluations are required, as in the present application, this computational cost quickly becomes prohibitive. While more efficient formulations of SSA have been developed for weakly coupled reaction networks, their computational cost remains proportional to M for strongly coupled reaction networks [35]. A reaction network is weakly coupled if the number of reactions that are influenced by any other reaction is bounded by a constant. If a network contains at least one reaction whose firing influences the propensities of a fixed proportion (in the worst case all) of the other reactions, then the network is strongly coupled [35]. Scale-free networks as seem to be characteristic for systems biology models [1, 42] are by definition strongly coupled. This is due to the existence of *hubs* that have a higher connection probability than other nodes. These hubs frequently correspond to chemical reactions that produce or consume species that also participate in the majority of the other reactions, such as water, ATP, or CO_2 in metabolic networks.

We use partial-propensity methods [35, 36] to simulate trajectories according to the solution of the chemical master (28.7) of the forward model. Partial-propensity methods are exact SSAs whose computational cost scales at most linearly with the number N of species in the network [35]. For large networks, this number is usually much smaller than the number of reactions. Depending on the network model at hand, different partial-propensity methods are available for its efficient simulation. Strongly coupled networks where the rate constants span only a limited spectrum of values are best simulated with the partial-propensity direct method (PDM) [35]. Multi-scale networks where the rate constants span many orders of magnitude are most efficiently simulated using the sorting partial-propensity direct method (SPDM) [35]. Weakly coupled reaction networks can be simulated at constant computational cost using the partial-propensity SSA with composition-rejection sampling (PSSA-CR) [37]. Lastly, reaction networks that include time delays can be exactly simulated using the delay partial-propensity direct method (dPDM) [38]. Different combinations of the algorithmic modules of partial-propensity methods can be used to constitute all members of this family of SSAs [36]. We refer to the original publications for algorithmic details, benchmarks of the computational cost, and a proof of exactness of partial-propensity methods.

5 Objective Function

In the context of parameter identification of stochastic biochemical networks, a number of distance or objective functions have previously been suggested. Reinker et al. proposed an approximate maximum-likelihood measure under the assumption that only a small number of reactions fire between two experimental measurement points, and a likelihood based on singular value decomposition that works when many reactions occur per time interval [40]. Koutroumpas et al. compared objective functions based on least squares, normalized cross-correlations, and conditional probabilities using a Genetic Algorithm [21]. Koeppl and co-workers proposed the Kantorovich distance to compare experimental and model-based probability distributions [20]. Alternative distance measures include the Earth Mover's distance or the Kolomogorov–Smirnov distance [32]. These distance measures, however, can only be used when many experimental trajectories are available. In order to measure the distance between a *single* experimental trajectory $\hat{\underline{n}}(t)$ and a *single* model output $\underline{n}(t)$, we propose a novel cost function $f(\underline{\theta}) = f\left(\mathscr{M}(\underline{\theta}), \hat{\underline{n}}\right)$ that reasonably captures the kinetics of a monostable system. We define a compound objective function $f(\underline{\theta}) = f_1(\underline{\theta}) + f_2(\underline{\theta})$ with

$$f_1(\underline{\theta}) = \sum_{i=1}^{4} \gamma_i \ , \qquad f_2(\underline{\theta}) = \sum_{i=1}^{N} \frac{\sum_{l=0}^{z_x} |\mathrm{ACF}_l(\hat{n}_i) - \mathrm{ACF}_l(n_i)|}{\sum_{l=0}^{z_x} \mathrm{ACF}_l(\hat{n}_i)} \ , \tag{28.18}$$

where

$$\gamma_i = \sum_{j=1}^{N} \sqrt{\left(\frac{\mu_i(n_j) - \mu_i(\hat{n}_j)}{\mu_i(\hat{n}_j)} \right)^2} \tag{28.19}$$

with the central moments given by:

$$\mu_i(n_j) = \begin{cases} \sum_{p=1}^{K} n_j \left(t_0 + (p-1)\Delta t_{\mathrm{exp}}\right) & \text{if} \quad i = 1 \\ \left(\left| \sum_{q=1}^{K} \left(n_j \left(t_0 + (q-1)\Delta t_{\mathrm{exp}}\right) - \mu_1(n_j) \right)^i \right| \right)^{1/i} & \text{otherwise} \end{cases} \tag{28.20}$$

and the time–autocorrelation function (ACF) at lag l given by:

$$\mathrm{ACF}_l(n_i) = \frac{n_i(t_0) n_i(t_0 + l\ \Delta t_{\mathrm{exp}}) - (\mu_1(n_i))^2}{\mu_2(n_i)}.$$

The variable z_x is the lag at which the experimental ACF crosses 0 for the first time. The function $f_1(\underline{\theta})$ measures the difference between the first four moments of $\underline{n}$ and $\hat{\underline{n}}$. This function alone would, however, not be enough to capture the kinetics since it lacks information about correlations in time. This is taken into account by $f_2(\underline{\theta})$, measuring the difference in the lifetimes of all chemical species. These lifetimes are systematically modulated by the volume Ω [39], hence enabling volumetric measurements of intra-cellular reaction compartments along with the identification of the rate constants.

The present objective function allows inclusion of experimental readouts from image-based systems biology. The moment-matching part is a typical readout from fluorescence photometry, whereas the autocorrelation of the fluctuations can directly be measured using, e.g., FCS.

6 Results

We estimate the unknown parameters $\underline{\theta}$ for two reaction networks: a weakly coupled cyclic chain and a strongly coupled non-linear colloidal aggregation network. For the cyclic chain we estimate $\underline{\theta}$ at steady state. For the aggregation model we estimate $\underline{\theta}$ both at steady state and in the transient phase. Every kinetic parameter is allowed to vary in the interval $[10^{-3},\ 10^{3}]$ and the reaction volume Ω in $[1,\ 500]$. Each GaA run starts from a point selected uniformly at random in logarithmic parameter space.

6.1 Weakly Coupled Reaction Network: Cyclic Chain

The cyclic chain network is given by:

$$\begin{aligned} \mathrm{S}_i &\xrightarrow{k_i} \mathrm{S}_{i+1} \qquad i = 1, \ldots, N-1, \\ \mathrm{S}_i &\xrightarrow{k_N} \mathrm{S}_1 \qquad i = N. \end{aligned} \tag{28.21}$$

In this linear network, the number of reactions M is equal to the number of species N. The maximum degree of coupling of this reaction network is 2, irrespective of the size of the system (length of the chain), rendering it weakly coupled [35]. We hence use PSSA-CR to evaluate the forward model with a computational complexity of $O(1)$ [37]. In the present test case, we limit ourselves to 3 species and 3 reactions, i.e., $N = M = 3$. The parameter vector for this case is given by $\underline{\theta} = [k_1, k_2, k_3]$, since the kinetics of linear reactions is independent of the volume Ω [39].

We simulate steady-state "experimental" data $\underline{\hat{n}}$ using PSSA-CR with ground truth $k_1 = 2$, $k_2 = 1.5$, $k_3 = 3.2$ (see Fig. 28.1a). We set the initial population of the species to $n_1(t = 0) = 50$, $n_2(t = 0) = 50$, and $n_3(t = 0) = 50$ and sample a single CME trajectory at equi-spaced time points with $\Delta t_{\mathrm{exp}} = 0.1$ between $t = t_0$ and $t = t_0 + (K-1)\Delta t_{\mathrm{exp}}$ with $t_0 = 2000$ and $K = 1001$ for each of the 3 species S_1, S_2, and S_3. For the generated data we find $z_x = 7$.

We generate trajectories from the forward model for every parameter vector $\underline{\theta}$ proposed by GaA using PSSA-CR between $t = 0$ and $t = (K-1)\Delta t_{\mathrm{exp}} = 100$, starting from the initial population $n_i(t = 0) = \hat{n}_i(t = t_0)$.

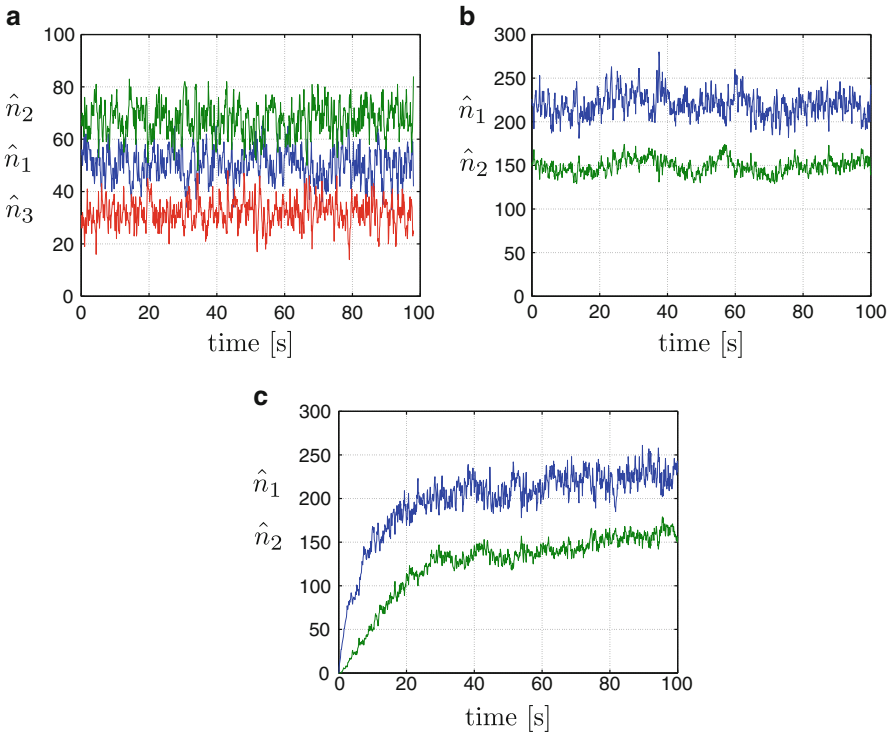


Fig. 28.1 In silico data for all test cases. (**a**) Time evolution of the populations of three species in the cyclic chain model at steady state (starting at $t_0 = 2000$). (**b**) Time evolution of the populations of two species in the aggregation model at steady state (starting at $t_0 = 5000$). (**c**) Same as (**b**), but during the transient phase (starting at $t_0 = 0$)

Before turning to the actual parameter identification, we illustrate the topography of the objective function landscape for the present example. We fix $k_3 = 3.2$ to its optimal value and perform a two-dimensional grid sampling for k_1 and k_2 over the full search domain. We use 40 logarithmically spaced sample points per parameter, resulting in 40^2 parameter combinations. For each combination we evaluate the objective function. The resulting landscapes of $f_1(\underline{\theta})$, $f_2(\underline{\theta})$, and $f(\underline{\theta})$ are depicted in Fig. 28.2a. Figure 28.2b shows refined versions around the global optimum. We see that the moment-matching term $f_1(\underline{\theta})$ is largely responsible for the global single-funnel topology of the landscape. The autocorrelation term $f_2(\underline{\theta})$ sharpens the objective function near the global optimum and renders it locally more isotropic.

We perform both global optimization and ABC runs using GaA. In each of the 15 independent optimization runs the number of objective function evaluations (FES) is limited to MAX_FES $= 1000M = 3000$. We set the initial step size to $r^{(0)} = 1$ and perform all searches in logarithmic scale of the parameters. Independent restarts from uniform random points are performed when the step size r drops below

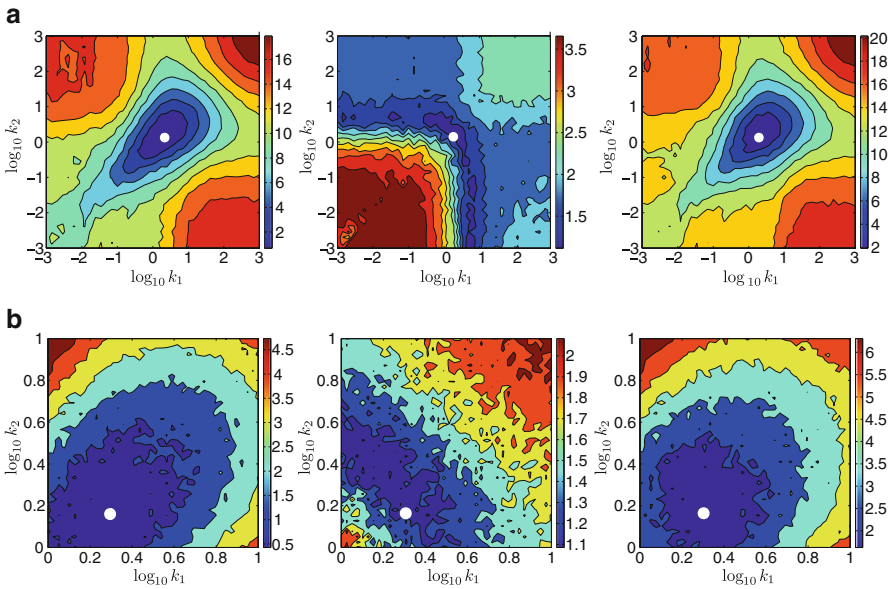


Fig. 28.2 (**a**) Global objective function landscape for the cyclic chain over the complete search domain for optimal $k_3 = 3.2$. The three panels from left to right show $f_1(\underline{\theta})$, $f_2(\underline{\theta})$, and $f(\underline{\theta})$, respectively. (**b**) A refined view of the global objective function landscape near the global optimum. The three panels from left to right show $f_1(\underline{\theta})$, $f_2(\underline{\theta})$, and $f(\underline{\theta})$, respectively. The white dots mark the ground truth parameters

10^{-4} [29]. For each of the 15 independent runs, the 30 parameter vectors with the smallest objective function value are collected and displayed in the box plot shown in the left panel of Fig. 28.3a. All 450 collected parameter vectors have objective function values smaller than 1.6. These results suggest that the present method is able to accurately determine the correct scale of the kinetic parameters from a single experimental trajectory, although an overestimation of the rates is apparent.

We use the obtained optimization results for subsequent ABC runs. We conduct 15 independent ABC runs using $c_T = 2$. The starting points for the ABC runs are selected uniformly at random from the 450 collected parameter vectors in order to ensure stable initialization. For each run we again set MAX_FES $= 1000M = 3000$. The initial step size $r^{(0)}$ is set to 0.1, and the parameters are again explored in logarithmic scale. For all runs we observe rapid convergence of the empirical hitting probability p_{emp} to the optimal $p = \frac{1}{e}$ (see Sect. 3). We collect the ABC samples along with the means and covariances of GaA as soon as $|p_{\text{emp}} - p| < 0.05$. As an example we show the histograms of the posterior samples for a randomly selected run in Fig. 28.3b. The means of the posterior distributions are again larger than the true kinetic parameters. Using GaA's means, covariance matrices, and the corresponding hitting probabilities that generated the posterior

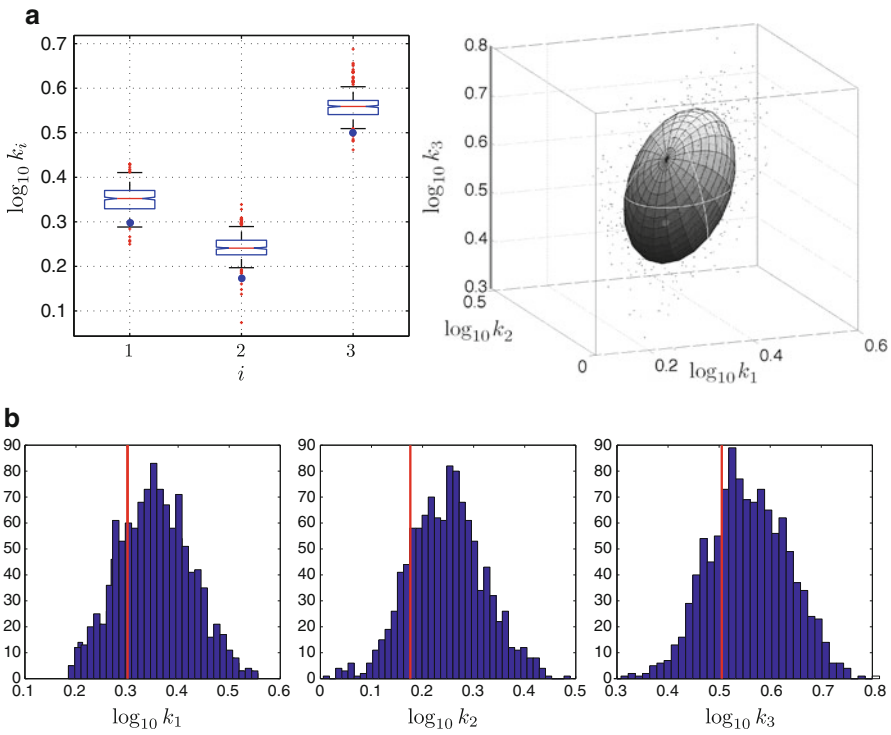


Fig. 28.3 (**a**) *Left panel*: Box plot of the 30 best parameter vectors from each of the 15 independent optimization runs. The blue dots mark the true parameter values. *Right panel*: Ellipsoidal volume estimate of the parameter space below an objective-function threshold $c_{\mathrm{T}} = 2$ from a single ABC run. (**b**) Empirical posterior distributions of the kinetic parameters from the same single ABC run with $c_{\mathrm{T}} = 2$. The red lines indicate the true parameters

samples, we can construct an ellipsoidal volume estimation [30]. This is done by multiplying each eigenvalue of the average of the collected covariance matrices with $c_{p_{\mathrm{emp}}} = \mathrm{inv}\,\chi^2_n(p_{\mathrm{emp}})$, the n-dimensional inverse Chi-square distribution evaluated at the empirical hitting probability. The product of these scaled eigenvalues and the volume of the n-dimensional unit sphere, $|S(n)| = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$, then yields the ellipsoid volume with respect to a uniform distribution (see [30] for details). The resulting ellipsoid contains the optimal kinetic parameter vector and is depicted in the right panel of Fig. 28.3a. It has a volume of 0.045 in log-parameter space. This constitutes only 0.0208% of the initial search space volume, indicating that GaA significantly narrows down the viable parameter space around the true optimal parameters despite the noise in the forward model and in the data.

6.2 *Strongly Coupled Reaction Network: Colloidal Aggregation*

The colloidal aggregation network is given by:

$$\begin{aligned} \emptyset &\xrightarrow{k_1^{\text{on}}} S_1 \\ S_i + S_j &\xrightarrow{k_{ij}} S_{i+j} \qquad i+j = 1,\ldots,N \\ S_{i+j} &\xrightarrow{\bar{k}_{ij}} S_i + S_j \qquad i+j = 1,\ldots,N \\ S_i &\xrightarrow{k_i^{\text{off}}} \emptyset \qquad i = 1,\ldots,N. \end{aligned} \tag{28.22}$$

For this network of N species, the number of reactions is $M = \left\lfloor \frac{N^2}{2} \right\rfloor + N + 1$. The maximum degree of coupling of this reaction network is proportional to N, rendering the network strongly coupled [35]. We hence use SPDM to evaluate the forward model with a computational complexity of $O(N)$ [35]. We use SPDM instead of PDM since the search path of GaA is unpredictable and could well generate parameters that lead to multi-scale networks. For this test case, we limit ourselves to two species, i.e., $N = 2$ and $M = 5$. The parameter vector for this case is $\underline{\theta} = \left[k_{11}, \bar{k}_{11}, k_1^{\text{on}}, k_1^{\text{off}}, k_2^{\text{off}}, \Omega\right]$.

We perform GaA global optimization runs following the same protocol as for the cyclic chain network with MAX_FES = $1000(M+1) = 6000$.

6.2.1 At Steady State

We simulate "experimental" data $\underline{\hat{n}}$ using SPDM with ground truth $k_{11} = 0.1$, $\bar{k}_{11} = 1.0$, $k_1^{\text{on}} = 2.1$, $k_1^{\text{off}} = 0.01$, $k_2^{\text{off}} = 0.1$, and $\Omega = 15$ (see Fig. 28.1b). We set the initial population of the species to $n_1(t=0) = 0$, $n_2(t=0) = 0$, and $n_3(t=0) = 0$ and sample $K = 1001$ equi-spaced data points between $t = t_0$ and $t = t_0 + (K-1)\Delta t_{\text{exp}}$ with $t_0 = 5000$ and $\Delta t_{\text{exp}} = 0.1$.

We generate trajectories from the forward model for every parameter vector $\underline{\theta}$ proposed by GaA using SPDM between $t = 0$ and $t = (K-1)\Delta t_{\text{exp}} = 100$, stating from the initial population $n_i(t=0) = \hat{n}_i(t=t_0)$.

The optimization results are summarized in the left panel of Fig. 28.4a. For each of the 15 independent runs, the 30 lowest-objective parameter vectors are collected and shown in the box plot. We observe that the true parameters corresponding to $\theta_2 = \bar{k}_{11}$, $\theta_3 = k_1^{\text{on}}$, $\theta_4 = k_1^{\text{off}}$, and $\theta_5 = k_2^{\text{off}}$ are between the 25th and 75th percentiles of the identified parameters. Both the first parameter and the reaction volume are, on average, overestimated. Upon rescaling the kinetic rate constants with the estimated volume, we find $\underline{\theta}^{\text{norm}} = [\theta_1/\theta_6, \theta_2, \theta_3\,\theta_6, \theta_4, \theta_5]$, which are the specific probability rates of the reactions. The identified values are shown in the

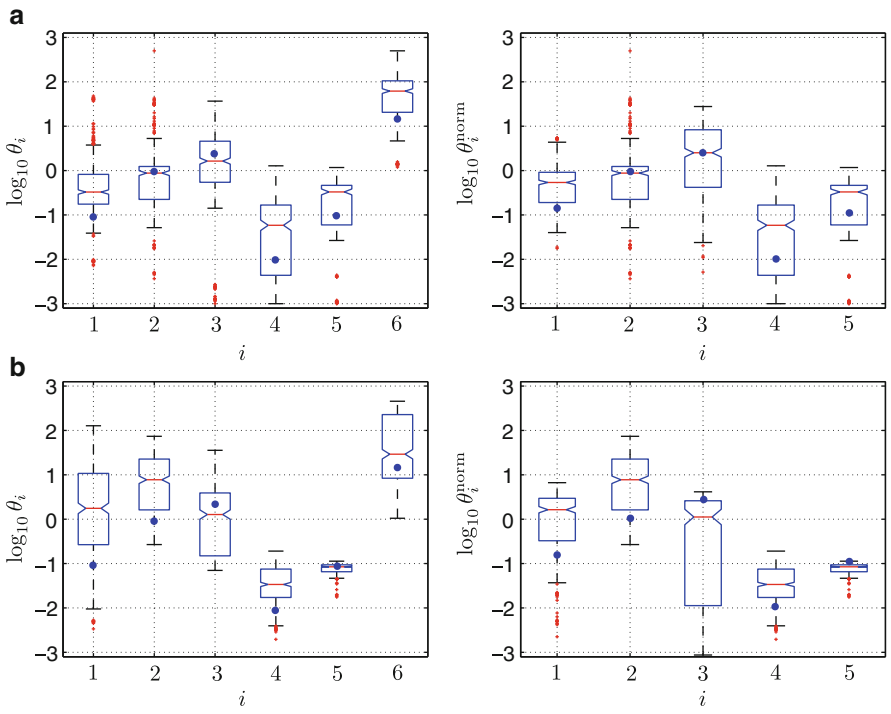


Fig. 28.4 (**a**) *Left panel*: Box plot of the 30 best parameter vectors from each of the 15 independent optimization runs for the steady-state data set. *Right panel*: Box plots of the normalized parameters (see main text for details). (**b**) *Left panel*: Box plot of the 30 best parameter vectors from each of the 15 independent optimization runs for the transient data set. *Right panel*: Box plot of the normalized parameters (see main text for details). The blue dots indicate the true parameter values

right panel of Fig. 28.4a. The median of the identified θ_3^{norm} coincides with the true specific probability rate. Likewise, θ_1^{norm} is closer to the 25th percentile of the parameter distribution. This suggests a better estimation performance of GaA in the space of specific probability rates, at the expense of not obtaining an estimate of the reactor volume.

6.2.2 In the Transient Phase

We simulate "experimental" data in the transient phase of the network dynamics using the same parameters as above between $t = t_0$ and $t = (K-1)\Delta t_{\mathrm{exp}}$ with $t_0 = 0$, $\Delta t_{\mathrm{exp}} = 0.1$, and $K = 1001$ (see Fig. 28.1c). We evaluate the forward model with $n_i(t = 0) = \hat{n}_i(t = t_0)$ to obtain trajectories between $t = 0$ and $t = (K-1)\Delta t_{\mathrm{exp}}$ for every proposed parameter vector $\underline{\theta}$.

The optimization results for the transient case are summarized in Fig. 28.4b. We observe that the true parameters corresponding to $\theta_3 = k_1^{\text{on}}$, $\theta_5 = k_2^{\text{off}}$, and $\theta_6 = \Omega$ are between the 25th and 75th percentiles of the identified parameters. The remaining parameters are, on average, overestimated. In the space of rescaled parameters $\underline{\theta}^{\text{norm}}$ we do not observe a significant improvement of the estimation.

7 Conclusions and Discussion

We have considered parameter estimation in monostable stochastic biochemical networks from single experimental trajectories. Parameter identification from single time series is desirable in image-based systems biology, where per-cell estimates of the fluorescence evolution and its fluctuations are available. This enables quantifying cell–cell variability on the level of network parameters. The histogram of the parameters identified for different cells provides a biologically meaningful way of assessing phenotypic variability beyond simple differences in the fluorescence levels.

We have proposed a novel combination of a flexible Monte Carlo method, the GaA algorithm, and efficient exact stochastic simulation algorithms, the partial-propensity methods. The presented method can be used for global parameter optimization, approximate Bayesian inference under uniform prior, and ellipsoidal volume estimation of the viable parameter space. We have introduced an objective function that measures closeness between a single experimental trajectory and a single trajectory generated by the forward model. The objective function comprises a moment-matching and a time-autocorrelation part. This allows including experimental readouts from, e.g., fluorescence photometry and FCS.

We have applied the method to estimate the parameters of two monostable reaction networks from a single simulated temporal trajectory each, both at steady state and during transient phases. We considered the linear cyclic chain network and a non-linear colloidal aggregation network. For the linear model we were able to robustly identify a small region of parameter space containing the true kinetic parameters. In the non-linear aggregation model, we could identify several parameter vectors that fit the simulated experimental data well. There are two possible reasons for this reduced parameter identifiability: either GaA cannot find the globally optimal region of parameter space due to high ruggedness and noise in the objective function, or the non-linearity of the aggregation network modulates the kinetics in a non-trivial way [10, 39]. Both cases are not accounted for in the current objective function, thus leading to reduced performance for non-linear reaction networks.

We also used GaA as an adaptive MCMC method for approximate Bayesian inference of the posterior parameter distributions in the linear chain network. This enabled estimating the volume of the viable parameter space below a given objective-function value threshold. We found these volume estimates to be stable

across independent runs. We thus believe that GaA might be a useful tool for exploring the parameter spaces of stochastic systems.

Future work will include (1) alternative objective functions that include temporal cross-correlations between species and the derivative of the autocorrelation; (2) longer experimental trajectories; (3) multi-stable and oscillatory systems; and (4) alternative global optimization schemes. Moreover, the applicability of the present method to large-scale, non-linear biochemical networks, and real-world experimental data will be tested in future work.

Acknowledgments RR was financed by a grant from the Swiss SystemsX.ch initiative (grant WingX), evaluated by the Swiss National Science Foundation. This project was also supported with a grant from the Swiss SystemsX.ch initiative, grant LipidX-2008/011, to IFS.

References

1. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(21):4947–4957
2. Andrieu C, Thoms J (2008) A tutorial on adaptive MCMC. Stat Comput 18(4):343–373
3. Auger A, Chatelain P, Koumoutsakos P (2006) *R*-leaping: accelerating the stochastic simulation algorithm by reaction leaps. J Chem Phys 125:084103
4. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
5. Boys RJ, Wilkinson DJ, Kirkwood TBL (2008) Bayesian inference for a discretely observed stochastic kinetic model. Stat Comput 18(2):125–135
6. Cardinale J, Rauch A, Barral Y, Székely G, Sbalzarini IF (2009) Bayesian image analysis with on-line confidence estimates and its application to microtubule tracking. In: Proc. IEEE Int Symp Biomedical Imaging (ISBI). IEEE, Boston, USA, pp 1091–1094
7. Cinquemani E, Milias-Argeitis A, Summers S, Lygeros J (2008) Stochastic dynamics of genetic networks: modelling and parameter identification. Bioinformatics 24(23):2748–2754
8. Gillespie DT (1992) A rigorous derivation of the chemical master equation. Phys A 188:404–425
9. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys 115(4):1716–1733
10. Grima R (2009) Noise-induced breakdown of the Michaelis–Menten equation in steady-state conditions. Phys Rev Lett 102(21):218103 DOI 10.1103/PhysRevLett.102.218103
11. Hafner M, Koeppl H, Hasler M, Wagner A (2009) 'Glocal' robustness analysis and model discrimination for circadian oscillators. PLoS Comput Biol 5(10):e1000534
12. Helmuth JA, Burckhardt CJ, Greber UF, Sbalzarini IF (2009) Shape reconstruction of subcellular structures from live cell fluorescence microscopy images. J Struct Biol 167:1–10
13. Helmuth JA, Sbalzarini IF (2009) Deconvolving active contours for fluorescence microscopy images. In: Proc Int Symp Visual Computing (ISVC) (Lecture notes in computer science), vol 5875. Springer, Las Vegas, USA, pp 544–553
14. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620–630 DOI 10.1103/PhysRev.106.620
15. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
16. Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
17. Kjellström G (1969) Network optimization by random variation of component values. Ericsson Tech 25(3):133–151
18. Kjellström G (1991) On the efficiency of Gaussian Adaptation. J Optim Theor Appl 71(3):589–597

19. Kjellström G, Taxen L (1981) Stochastic optimization in system design. IEEE Trans Circ Syst 28(7):702–715
20. Koeppl H, Setti G, Pelet S, Mangia M, Petrov T, Peter M (2010) Probability metrics to calibrate stochastic chemical kinetics. In: Proc IEEE Int Symp Circuits and Systems, Paris, France, pp 541–544
21. Koutroumpas K, Cinquemani E, Kouretas P, Lygeros J (2008) Parameter identification for stochastic hybrid systems using randomized optimization: a case study on subtilin production by *Bacillus subtilis*. Nonlin Anal Hybrid Syst 2(3):786–802
22. Kurtz TG (1972) Relationship between stochastic and deterministic models for chemical reactions. J Chem Phys 57(7):2976–2978
23. Lakowicz JR (2006) Principles of fluorescence spectroscopy. Springer USA DOI 10.1007/978-0-387-46312-4
24. Ljung L (2002) Prediction error estimation methods. Circ Syst Signal Process 21(1):11–21
25. Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100(26):15324–15328
26. Mason O, Verwoerd M (2007) Graph theory and networks in biology. Syst Biol IET 1(2):89–119
27. Müller CL (2010) Exploring the common concepts of adaptive MCMC and Covariance Matrix Adaptation schemes. In: Auger A, Shapiro JL, Whitley D, Witt C (eds.) Theory of evolutionary algorithms, Dagstuhl Seminar Proceedings, no. 10361. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany, Dagstuhl, Germany URL http://drops.dagstuhl.de/opus/volltexte/2010/2813
28. Müller CL, Sbalzarini IF (2010) Gaussian Adaptation as a unifying framework for continuous black-box optimization and adaptive Monte Carlo sampling. In: Proc IEEE Congress on Evolutionary Computation (CEC). Barcelona, Spain, pp 2594–2601
29. Müller CL, Sbalzarini IF (2010) Gaussian Adaptation revisited – an entropic view on covariance matrix adaptation. In: Proc EvoStar (Lecture notes computer science), vol 6024. Springer, Istanbul, Turkey, pp 432–441
30. Müller CL, Sbalzarini IF (2011) Gaussian Adaptation for robust design centering. In: Proc EuroGen Int Conf Evolutionary and Deterministic Methods for Design, Optimization and Control. Capua, Italy, pp 736–742
31. Munsky B, Trinh B, Khammash M (2009) Listening to the noise: random fluctuations reveal gene network parameters. Mol Sys Biol 5(1):318
32. Poovathingal SK, Gunawan R (2010) Global parameter estimation methods for stochastic biochemical systems. BMC Bioinformatics 11(1):414
33. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16(12):1791–1798
34. Qian H, Elson EL (2004) Fluorescence correlation spectroscopy with high-order and dual-color correlation to probe nonequilibrium steady states. Proc Natl Acad Sci USA 101(9):2828–2833
35. Ramaswamy R, González-Segredo N, Sbalzarini IF (2009) A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. J Chem Phys 130(24):244104
36. Ramaswamy R, Sbalzarini IF (2010) Fast exact stochastic simulation algorithms using partial propensities. In: Proc ICNAAM, numerical analysis and applied mathematics, international conference. AIP, Rhodes, Greece, pp 1338–1341
37. Ramaswamy R, Sbalzarini IF (2010) A partial-propensity variant of the composition-rejection stochastic simulation algorithm for chemical reaction networks. J Chem Phys 132(4):044102
38. Ramaswamy R, Sbalzarini IF (2011) A partial-propensity formulation of the stochastic simulation algorithm for chemical reaction networks with delays. J Chem Phys 134:014106
39. Ramaswamy R, Sbalzarini IF, González-Segredo N (2011) Noise-induced modulation of the relaxation kinetics around a non-equilibrium steady state of non-linear chemical reaction networks. PLoS ONE 6(1):e16045

40. Reinker S, Altman RM, Timmer J (2006) Parameter estimation in stochastic biochemical reactions. IEE Proc Syst Biol 153(4):168
41. Stock G, Ghosh K, Dill KA (2008) Maximum Caliber: a variational approach applied to two-state dynamics. J Chem Phys 128(19):194102
42. Strogatz SH (2001) Exploring complex networks. Nature 410:268–276
43. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6(31):187–202
44. Wolkenhauer O (2001) Systems biology: the reincarnation of systems theory applied in biology? Briefings Bioinform 2(3):258
45. Zechner C, Pelet S, Peter M, Koeppl H (2011) Recursive Bayesian estimation of stochastic rate constants from heterogeneous cell populations. In: Proc 50th IEEE CDC, Conference on Decision and Control. Orlando, Florida, USA

Chapter 29
A Systems Biology View of Adaptation in Sensory Mechanisms

Pablo A. Iglesias

Abstract Adaptation, the desensitization to persistent changes in environmental conditions, is present throughout biological sensory mechanisms. Not surprisingly, it has been an active area of research to systems biologists. Here, we consider some of the models proposed to account for adaptation as well as the experiments used to motivate and validate these models. We discuss some salient features of these models including robustness, deadaptation, transient responses, and the response of these systems to more complex temporal stimuli. While most of these models have been used to study chemoattractant-induced responses in bacteria and amoebae, the system-theoretic issues associated with these systems are of importance in a broad spectrum of biological systems.

1 Introduction

All organisms, from the simplest single-celled species to humans, use sensory mechanisms to monitor and respond to changes in their environment. An important aspect of these systems is the ability to *adapt* – to adjust their sensitivity so as to be able to respond to a wide range of inputs. As an example, the human eye adapts to varying levels of light in approximately 5–30 min, partly by regulating the quantity of light that reaches the retina, but also by changing the sensitivity of rods and cones [1]. In vision, problems with adaptation can be lead to nyctalopia (night blindess) or hemeralopia.

In single-celled organisms, adaptation is probably best understood in the chemoattractant-mediated response of bacteria [2, 3] and amoeba [4, 5]. Fast

P.A. Iglesias (✉)
Department of Electrical and Computer Engineering, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
e-mail: pi@jhu.edu

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_29,

swimming bacteria, such as *Escherichia coli*, direct migration by responding to temporal changes in receptor–ligand binding to chemical chemoattractants and repellents. Two proteins, CheW and CheA bind to the receptor. The latter, a histidine kinase, phosphorylates the response regulator CheY. When phosphorylated, CheY binds to the flagellar motors inducing clockwise rotation which causes the bacteria to tumble. This tumbling stops the cell and reorients it in a more-or-less random direction. In contrast, in the presence of unphosphorylated CheY, the motors rotate in a counter-clockwise direction propelling the cell in a straight run. High receptor–ligand binding inactivates CheA thus maintaining CheY unphosphorylated. This ensures that in the presence of high chemoattractant concentration, tumbling is suppressed. A feedback mechanism, based on receptor methylation, is used by cells to adapt to the level of receptor occupation. Two enzymes, a methyltransferase CheR and a methylesterase CheB regulate adaptation in the chemotactic pathway. When receptor occupancy is high, CheA induces CheB phosphorylation. The subsequent methylation of the receptor returns CheA to its prestimulus levels.

It follows that when a cell is moving up a chemoattractant gradient, the rate of tumbling decreases, maintaining movement in this favorable direction. In this respect, perfect adaptation is a signature of a low-passed filtered temporal differentiator [6–8]. When presented by a constant dose of chemoattractant, the tumbling rate is unaffected by the actual concentration. In contrast, swimming in the direction of increasing concentration leads to a positive response which induces a decrease in tumbling rate. Similarly, a run in the direction of decreasing chemoattractant concentration results in a negative signal, manifested in an increase in tumbling rate.

In contrast to this temporal sensing, larger, slower cells like the amoeba *Dictyostelium discoideum* and human neutrophils employ a spatial sensing mechanism. With chemoattractant receptors uniformly distributed throughout their membrane, they compare the concentration of chemoattractant even when immobilized by suppressors of actin polymerization like latrunculin. Though the method of interpreting the chemoattractant gradient is different, these cells also display perfect adaptation. That is, when presented by a spatially uniform dose of chemoattractant, several signaling events, including actin polymerization and the translocation of PH-domain containing proteins from the cytosol to the membrane respond in a transient manner.

Here, we review some of the models that have been proposed to account for these adaptive responses. We highlight some of the differences and similarities with these models, and also present some open questions.

2 Perfect Adaptation

The property of adaptation can be described by the scheme depicted in Fig. 29.1. We note that there are two important components: the ability to detect the stimulus, referred to as the *sensitivity* of the sensory system [9], and the adaptation step. If we

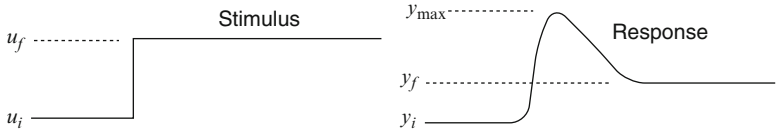


Fig. 29.1 Adaptive response to step changes in stimulus concentration. A sudden change in the stimulus concentration from an initial (u_i) to a final (u_f) value gives rise to a corresponding response from y_i, peaking at y_{max} before settling back to y_f. Perfect adaptation is a return to the initial value: $y_f = y_i$

denote the stimulus (the *input*) by u and the response (the *output*) by y, then the sensitivity is given by:

$$S = \left| \frac{(\max y(t) - y_i)/y_i}{(u_f - u_i)/u_i} \right|, \tag{29.1}$$

and the precision is

$$P = \left| \frac{(y_f - y_i)/y_i}{(u_f - u_i)/u_i} \right|, \tag{29.2}$$

where u_i and u_f are the initial and final input concentrations, respectively, and y_i and y_f are the initial and final output concentrations. Perfect adaptation refers to the property $P \equiv 0$. Note that, in general, these are both nonlinear functions of the stimulus, and hence the precision and sensitivity measures depend on both the initial and final input levels. Nevertheless, we will seek adapting systems where perfect adaptation holds over a wide range of concentrations.

If we consider dynamic models of signaling systems, we assume that the system equations can be expressed in terms of the nonlinear differential equations

$$\frac{dx}{dt} = f(x, p, u), \tag{29.3}$$

$$y = h(x, p, u). \tag{29.4}$$

Here, x is a vector representing internal states (such as the concentrations of different regulatory proteins, etc.), and p includes parameters in the differential equations. The function f is a vector field describing the differential equations. The function h outlines the output of the system, which may be a specific x (for example, if $y = x_1$) or a combination of states.

Perfect adaptation is a steady-state property. Thus, given constant initial and final inputs $u(t) = u_i$ and $u(t) = u_f$, we have

$$0 = f(x_i, p, u_i) \quad \Rightarrow \quad y_i = h(x_i, p, u_i),$$
$$0 = f(x_f, p, u_f) \quad \Rightarrow \quad y_f = h(x_f, p, u_f),$$

And Perfect adaptation requires that $h(x_f, p, u_f) = h(x_i, p, u_i)$.

3 A First Model of Adaptation

We now present a simple model of perfect adaptation based on covalent modification of receptors. The basic model is shown in Fig. 29.2. We assume that unoccupied receptors can be found in two states: R_1 and R_2, where the latter is a modified form of R_2. This receptor is capable of binding the ligand L; bound receptors are also assumed to exist in two states: C_1 and C_2. Two enzymes mediate reversible modification transitions between states. The inhibitory enzyme I catalyzes the modification from R_1 (respectively C_1) to R_2 (respectively C_2), whereas the reverse process is catalyzed by the excitation enzyme E. For example, *E. coli* chemoreceptors can be methylated (R_1, C_1) or not (R_2, C_2), and demethylation is catalyzed by the methylesterase CheB whereas methylation is catalyzed by the methyltransferase CheR.

We now write differential equations describing the concentration of the different species depicted in Fig. 29.2. When the ligand (L) binds to an unbound receptor R, the complex C_1 is formed.

$$R_1 + L \underset{k_{-r}}{\overset{k_r}{\rightleftharpoons}} C_1.$$

In mathematical terms, the differential equation describing this reaction is

$$\frac{\mathrm{d}C_1}{\mathrm{d}t} = k_r R_1 \times L - k_{-r} C_1. \tag{29.5}$$

Similarly, for the modified receptors

$$\frac{\mathrm{d}C_2}{\mathrm{d}t} = k_d R_2 \times L - k_{-d} C_2. \tag{29.6}$$

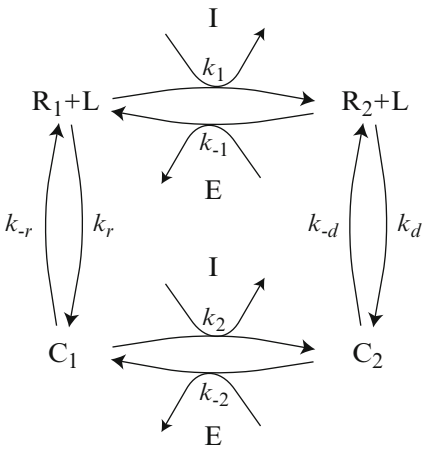


Fig. 29.2 Signaling mechanism used for adaptation. Unmodified receptors can be bound, C_1, or not, R_1, to the ligand L. Receptors also exist in a modified form in both bound, C_2, and unbound, R_2, states. Covalent modification between the two receptor states is mediated by two enzymes, E and I

Using Michaelis–Menten kinetics, a differential equation describing the change in concentration between R_1 and R_2 in the system of Fig. 29.2 is

$$\frac{\mathrm{d}R_1}{\mathrm{d}t} = \frac{k_{-1}E_\mathrm{T}R_2}{k_{M_E} + R_2} - \frac{k_1 I_\mathrm{T} R_1}{k_{M_I} + R_1},$$

and, similarly

$$\frac{\mathrm{d}C_1}{\mathrm{d}t} = \frac{k_{-2}E_\mathrm{T}C_2}{k_{M_E} + C_2} - \frac{k_2 I_\mathrm{T} C_1}{k_{M_I} + C_1}.$$

Combining the effects of ligand binding with the enzymatic reactions leads to the following set of four nonlinear differential equations describing the system

$$\begin{aligned}
\frac{\mathrm{d}R_1}{\mathrm{d}t} &= \frac{k_{-1}E_\mathrm{T}R_2}{k_{M_E} + R_2} - \frac{k_1 I_\mathrm{T} R_1}{k_{M_I} + R_1} - k_r R_1 \times L + k_{-r} C_1,\\
\frac{\mathrm{d}R_2}{\mathrm{d}t} &= -\frac{k_{-1}E_\mathrm{T}R_2}{k_{M_E} + R_2} + \frac{k_1 I_\mathrm{T} R_1}{k_{M_I} + R_1} - k_d R_2 \times L + k_{-d} C_2,\\
\frac{\mathrm{d}C_1}{\mathrm{d}t} &= \frac{k_{-2}E_\mathrm{T}C_2}{k_{M_E} + C_2} - \frac{k_2 I_\mathrm{T} C_1}{k_{M_I} + C_1} + k_r R_1 \times L - k_{-r} C_1,\\
\frac{\mathrm{d}C_2}{\mathrm{d}t} &= -\frac{k_{-2}E_\mathrm{T}C_2}{k_{M_E} + C_2} + \frac{k_2 I_\mathrm{T} C_1}{k_{M_I} + C_1} + k_d R_2 \times L - k_{-d} C_2.
\end{aligned}$$

Conservation of receptors implies that $R_\mathrm{T} = R_1 + C_1 + R_2 + C_2$.

We assume that both enzymes are working in the linear regime ($R_1, C_1 \ll k_{M_I}$ and $R_2, C_2 \ll k_{M_E}$), so that

$$\frac{k_1 I_\mathrm{T} R_1}{k_{M_I} + R_1} \approx \frac{k_1 I_\mathrm{T}}{k_{M_I}} R_1 = k_i R_1, \quad \text{and} \quad \frac{k_1 I_\mathrm{T} C_1}{k_{M_I} + C_1} \approx \frac{k_1 I_\mathrm{T}}{k_{M_I}} C_1 = k_i C_1,$$

with similar expressions for the enzymes' actions on R_2 and C_2. Using the conservation of receptors and normalizing by the total number of receptors, we can rewrite the system, as in (29.3), as follows:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\\
&= \underbrace{\begin{bmatrix} -(k_i + k_r L) & k_e & k_{-r} \\ k_i - k_{-d} & -(k_e + k_d L + k_{-d}) & -k_{-d} \\ k_r L - k_{-i} & -k_{-i} & -(k_{-r} + k_i + k_e) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} 0 \\ k_{-d} \\ k_e \end{bmatrix}}_{f(x,p,u)},
\end{aligned} \tag{29.7}$$

where $x_1 = R_1/R_T$, $x_2 = C_1/R_T$, $x_3 = R_2/R_T$, $x_4 = 1 - x_1 - x_2 - x_3$, $u = L$, and the parameters p represent the different coefficients and total enzyme concentrations.

We assume that the total level of activity, $y(t) = A(t)$, and that this is a linear combination of the four receptor states:

$$y(t) = \underbrace{\alpha_1 R_1(t) + \alpha_2 R_2(t) + \alpha_3 C_1(t) + \alpha_4 C_2(t)}_{h(x,p,u)}.$$

Solving (29.7) at equilibrium reveals that the steady-state concentration has the form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \frac{1}{m_1 + m_2 + L\sum_{i=1}^4 n_i} \left(\begin{bmatrix} m_1 \\ m_2 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{bmatrix} L \right),$$

where the m_i and n_i are terms that depend on the specific kinetic coefficients, though the precise form is not needed for the analysis that follows and is therefore omitted. For example, if there is no chemoattractant, $L = 0$, then $x_3 = x_4 = 0$ and the steady-state activity is given by:

$$A_0 = y(\infty)|_{L=0} = \frac{\alpha_1 m_1 + \alpha_2 m_2}{m_1 + m_2},$$

which implies that

$$A_0(m_1 + m_2) = \alpha_1 m_1 + \alpha_2 m_2.$$

With constant $L \neq 0$, we have that

$$A_L = y(\infty)|_{L\neq 0} = \frac{\alpha_1 m_1 + \alpha_2 m_2 + L\sum_i^4 \alpha_i n_i}{m_1 + m_2 + L\sum_{i=1}^4 n_i},$$

and this is equivalent to

$$A_L \left(m_1 + m_2 + L\sum_{i=1}^4 n_i \right) = \alpha_1 m_1 + \alpha_2 m_2 + L\sum_{i=1}^4 \alpha_i n_i.$$

Of course, perfect adaptation means that $A_0 = A_L$. To achieve this we select coefficients α_3 and α_4 so that $A_0 = A_L$ independent of the value of L. Equivalently, we require that

$$A_0 \sum_{i=1}^4 n_i = \sum_{i=1}^4 \alpha_i n_i.$$

As there are two free parameters and only one equation, this can always be achieved.

The scheme is equivalent to that proposed to account for adaptation [10]. Experiments on the responses of both *E. coli* and *D. discoideum* were used to obtain parameters [11].

4 A Robust Model of Adaptation

A property of a system, such as adaptation, can be classified as to whether it is *robust* or not depending on whether it is preserved under the presence of small perturbations in the components [12, 13]. A problem with the scheme presented above is that the property of perfect adaptation is achieved by fine tuning parameters for activity in a way that depends on the specific kinetic parameters. However, if we perturb these kinetic parameters, the resultant steady-state concentrations will differ and the activity will no longer be independent of L. In this respect the system is said to lack robustness. This *fragility* is seen in other more detailed models of adaptation [14, 15].

In a recent seminal paper, Barkai and Leibler proposed a model of the bacterial chemotactic network that achieved perfect adaptation in a robust manner [16]. In the context of the scheme presented in Fig. 29.2, we can recover their model by making the following assumptions. The first is that only fractions of the unmodified receptors, R_1 and C_1, are active, and that the total activity (the output of the system) can be expressed as a linear combination of these two states:

$$y = \alpha_1 x_1 + \alpha_3 x_3. \tag{29.8}$$

As second assumption is that the inhibitor enzyme, I, acts only on these active states. The kinetic constants for both these states are otherwise the same: $k_2 = k_1$ in Fig. 29.2. It is not necessary to assume that $k_{-1} = k_{-2}$. The final key assumption concerns the regimes in which the two enzymatic reactions are acting. We assume that the forward reaction (catalyzed by I) is occurring in the linear regime, whereas the reverse reaction (catalyzed by E), occurs at saturation. Using these assumptions, the equations above are replaced by:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -(\alpha_1 k_1 I_\mathrm{T} + k_r L) & 0 & k_{-r} & 0 \\ \alpha_1 k_1 I_\mathrm{T} & -k_d L & 0 & k_{-d} \\ k_r L & 0 & -(\alpha_3 k_2 I_\mathrm{T} + k_{-r}) & 0 \\ 0 & k_d L & \alpha_3 k_2 I_\mathrm{T} & -k_{-d} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} k_{-1} \\ k_{-2} \\ -k_{-1} \\ -k_{-2} \end{bmatrix} E_\mathrm{T}.$$

Note that, for simplicity, we have incorporated the Michaelis–Menten constant (k_{M_I}) into k_1.

Because of the third assumption, the differential equations describing the concentrations of the unmodified receptor states R_1 (x_1) and C_1 (x_3) do not involve the concentrations of R_2 and C_2. Thus, these two equations have been decoupled from

the other two. Furthermore, we can write these decoupled equations in matrix form as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} -(\alpha_1 k_1 I_{\mathrm{T}} + k_r L) & k_{-r} \\ k_r L & -(\alpha_3 k_2 I_{\mathrm{T}} + k_{-r}) \end{bmatrix}\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} + \begin{bmatrix} k_{-1} \\ -k_{-1} \end{bmatrix} E_{\mathrm{T}}.$$

We now perform a state-variable transformation. That is, we rewrite the differential equations in terms of two auxiliary states. The first is the total activity (given by (29.8)) which corresponds to the total concentration of active receptors. The second variable, z, is proportional to the total number of receptors in their unmodified form:

$$z = \frac{x_1 + x_3}{k_1 I_{\mathrm{T}}}. \tag{29.9}$$

In the new co-ordinates, the resultant second order differential equation is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} -a_1(L, I_{\mathrm{T}}) & a_0(L, I_{\mathrm{T}}) \\ -1 & 0 \end{bmatrix}\begin{bmatrix} y \\ z \end{bmatrix} + \begin{bmatrix} b_1(I_{\mathrm{T}}) \\ 1 \end{bmatrix} r_0, \tag{29.10}$$

where

$$\begin{aligned} a_0(L, I_{\mathrm{T}}) &= (\alpha_1 k_{-r} + \alpha_3 k_r L + \alpha_1 \alpha_3 k_1 I_{\mathrm{T}}) k_1 I_{\mathrm{T}}, \\ a_1(L, I_{\mathrm{T}}) &= k_r L + k_{-r} + (\alpha_1 + \alpha_3) k_1 I_{\mathrm{T}}, \\ b_1(I_{\mathrm{T}}) &= (\alpha_1 k_{-1} + \alpha_3 k_{-2}) k_1 I_{\mathrm{T}}) / (k_{-1} + k_{-2}), \\ r_0 &= \big((k_{-1} + k_{-2})/k_1\big)(E_{\mathrm{T}}/I_{\mathrm{T}}). \end{aligned}$$

Perfect adaptation follows immediately from the differential equation for z:

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = -y(t) + r_0. \tag{29.11}$$

Thus, whenever a steady-state in concentration is reached, the right-hand side of (29.11) is zero, or

$$A^{\mathrm{st}} = \lim_{t \longrightarrow \infty} y(t) = r_0. \tag{29.12}$$

Note that the steady-state value of activity depends on the constants k_1, k_{-1}, and k_{-2}, as well as the total concentrations of the two enzymes E and I. While not shown above, the time that it takes for the system to adapt also depends on these parameters and concentrations. However, as long as the two assumptions stated above are not violated, adaptation is a robust feature of the system. No matter what these constants and enzyme concentrations are, the system's steady-state activity A^{st} is independent of the ligand concentration.

In theory, a robust model of adaptation is more appealing than one that requires precise tuning of parameters. Alon and co-workers tested this notion of robustness

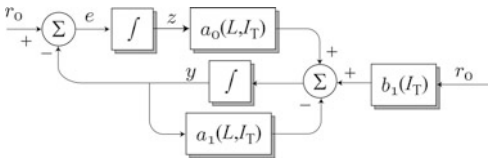


Fig. 29.3 Adaptation through integral control. The regulatory model for adaptation (29.10) can be represented as an integral control mechanism as shown. The difference between the activity, y and the adapted level of activity, r_0, is integrated ensuring that at steady-state, the two equal each other. This integral feedback controller is commonly used in engineering

in the context of the adaptation mechanism in bacterial chemotaxis [17]. They systematically altered the concentrations of various intracellular components of the pathway, including the concentrations of CheB, CheY, and the receptor, and monitored both the property of perfect adaptation, as well as other aspects of this behavior such as adaptation time and steady-state tumbling frequency. As predicted by the model, the property of perfect adaptation was remarkably robust – changes of CheR expression could be altered 50 fold and yet the adaptation precision was unaltered. However, both adaptation time and steady-state tumbling frequency depended on these changes, as suggested by the model.

4.1 Integral Feedback and the Internal Model Principle

The key to achieve robust perfect adaptation comes from the differential equation for the response, (29.11). If we consider the difference

$$e(t) = r_0 - y(t),$$

as an error signal denoting deviations away from the adapted level of activity, this equation is equivalent to the integration of the error signal (Fig. 29.3). In control engineering, it is well known that systems that need to reject constant disturbances require an integral control mechanism [15]. This is a special case of a more encompassing theory, the *internal model principle*, which states that to reject a disturbance robustly, the system must incorporate a model of the disturbance inside the control loop [18]. The engineering context of this is somewhat different than that suggested by the model above. Engineering systems usually consist of a *plant*, a system that is to be controlled, and a separate *controller*, a subsystem that is to be designed. This distinction is somewhat artificial in biological sensory systems. Moreover, in sensory systems a more important requirement is the property of *signal detection*, that is, that the sensitivity of the system, as defined in (29.1), not be zero. It can be shown that if a system adapts to a class of bounded external signals (i.e., achieves perfect adaptation), then the system necessarily contains a subsystem that is capable of generating the signals in that class [19]. Importantly, this holds even in

nonlinear systems. In perfect adaptation, the class of signals is the set of all constant stimuli, and the system that can generate these signals is an integrator.

It is worth asking to what extent this property itself is robust. For example, perfect adaptation in the Barkai–Leibler model relies on basic assumptions about the activity of the enzymes. If these assumptions do not hold exactly, then perfect adaptation is lost, though the system can adapt approximately, that is, $P \approx \varepsilon$. In this case, an *approximate internal model principle* holds [20, 21].

5 Adaptation Through Incoherent Feedforward Loops

The model of Barkai–Leibler is not the only model of a biochemical network that achieves robust adaptation. Here we present a simple model for adaptation, originally due to Koshland, who proposed it as a means of explaining adaptation in bacterial chemotaxis [22] though it is now more widely used to explain adaptation in models of chemotactic amoebae [23].

The central element of network is a response regulator, RR, that is activated by an excitatory process E, and suppressed by an inhibitory process, I (Fig. 29.4). Both of these regulatory processes are themselves regulated by the stimulus, through increases in receptor occupancy, S. A simple ordinary differential equation description of this scheme is given by:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -k_{-e}E + k_e S, \tag{29.13}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -k_{-i}I + k_i S, \tag{29.14}$$

$$\frac{\mathrm{dRR}}{\mathrm{d}t} = -k_{-r}I \times \mathrm{RR} + k_r E. \tag{29.15}$$

The system can be thought of as an *incoherent feedforward loop* in which positive and negative signals come directly from the receptor and act in a complementary, or incoherent, fashion on the response regulator. Incoherent feedforward loops have received considerable attention recently [24–29]. Note that Tyson refers to this topology as a *sniffer*, as it mimics the way that our sense of smell works [30].

Equations (29.13)–(29.15) represent one implementation of the incoherent feedforward loop, whereby the excitation and inhibition activate and inactivate,

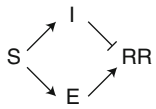


Fig. 29.4 Response regulator model. In this scheme, the external stimulus regulates two intermediate processes which act on a response regulator in complementary fashion

respectively, the response regulator. However, other possibilities exist [27, 31]. For example, the excitation and inhibition processes could inhibit degradation and activation respectively:

$$\frac{\mathrm{dRR}}{\mathrm{d}t} = -\frac{k_{-r}}{E}\mathrm{RR} + \frac{k_r}{I}. \tag{29.16}$$

We can rewrite (29.13)–(29.15) as:

$$\begin{aligned}
\frac{\mathrm{d}\hat{E}}{\mathrm{d}\tau} &= -\hat{E} + \hat{S},\\
\frac{\mathrm{d}\hat{I}}{\mathrm{d}\tau} &= -\alpha(\hat{I} - \hat{S}),\\
\varepsilon\frac{\mathrm{d}\widehat{\mathrm{RR}}}{\mathrm{d}\tau} &= -\hat{I} \times \widehat{\mathrm{RR}} + \hat{E}.
\end{aligned}$$

where the kinetic coefficients have been normalized [23]. Hereafter we assume that this normalization has taken place and drop the $\hat{}$ in the notation. That the system adapts is straightforward to check. In particular, at steady-state: $E = S$, $I = S$, and $RR = E/I = 1$. The transient signal is somewhat more complicated, but can be computed analytically [32]. It is straightforward to check that, if $\alpha = 1$, the system does not detect the change in stimulus. In this case, the receptor signal causes identical increases in both the excitation and inhibition processes, the net effect which is to leave the response regulator unaffected. If the excitation process is faster ($\alpha < 1$) then the response regulator rises in response to the faster increase in excitation. However, as the inhibitory process catches up, the response regulator returns to its steady-state value, and the system adapts perfectly. If $\alpha > 1$, then the stimulus increase causes faster rise in the inhibition leading to a transient decrease in the concentration of the response regulator.

It is also easy to see that this adaptation is completely robust to parameter variations. We might expect this is achieved as a consequence of an integral control mechanism. To demonstrate this we rewrite the equation for the response regulator as:

$$\varepsilon\frac{\mathrm{dRR}}{\mathrm{d}\tau} = -I\,(\mathrm{RR} - E/I)\,.$$

Provided that the stimulus in not zero, the ratio $E/I = 1 +$ "decaying transient" and hence this equation shows that the response regulator is acting as a feedback integral control system, though with a time-varying gain, $I(t)$.

5.1 Deadaptation

In this analysis of the incoherent feedforward loop we have to restrict the stimulus to values, $S \neq 0$. When $S = 0$, the equilibrium of the response regulator is not

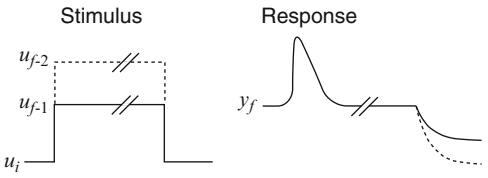


Fig. 29.5 Deadaptation dynamics of an incoherent feedforward loop model. Step increases of equal size give rise to responses that adapt to the same basal value. Once adapted, the stimuli is removed. For the system in which the response is described by (29.17), the cell's response decreases monotonically before settling on a final value that depends on the stimulus, demonstrating a memory of the stimulus concentration

isolated because the right-hand side of (29.15) is zero for all values of RR. This is clearly problematic and has consequences when we consider the cell's response to a removal of stimulus, or *deadapation*. The analysis is somewhat easier if we consider the following simplified model of an incoherent feedforward loop, which we write as:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= u - x, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= u - xy. \end{aligned} \tag{29.17}$$

We will assume that a constant stimulus, u_0, has been applied and that the system has adapted to this value. We select the initial time as the point at which the stimulus has been removed, at which point the differential equations are

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -x, \quad x(0) = u_0 \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -xy, \quad y(0) = 1. \end{aligned}$$

The equation for x is linear and hence it is easy to solve: $x(t) = \mathrm{e}^{-t}u_0$. Replacing this into the equation for the response leads to a time-varying scalar differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\mathrm{e}^{-t}u_0 y, \quad y(0) = 1,$$

whose solution

$$y(t) = \exp(-(1 - \mathrm{e}^{-t})u_0),$$

decays to

$$y(\infty) = \mathrm{e}^{-u_0} < 1.$$

Thus, the system steady-state depends on the original stimulus (Fig. 29.5). This memory can be expected whenever a pulse input is applied [33].

The actual deadaptation mechanism depends on the specific form of the incoherent feedforward loop used. For example, using the mechanism of (29.16), in which we replace the equation for y, which we write as:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = u/x - y$$

leads to the same adapted state, but different deadaptation behavior. In particular, if $y(0) = 1$ and $u(t) = 0$, for $t \geq 0$, then $y(t) = \mathrm{e}^{-t}$, which does not exhibit memory of the previous stimulus, but still does not return to the adapted steady-state in which $y = 1$ is obtained for all $u_0 > 0$.

Experimentally, Devreotes and co-workers studied the recovery of the signaling response in *D. discoideum* cells after adaptation to chemoattractant [34, 35]. As a marker of the response they considered the production and secretion of cAMP. In these experiments, a memory was observed in that reapplication of the stimulus after a period of deadaptation showed an irreversible decrease in responsiveness to the stimulus. More recently, Bodenschatz and co-workers used an elegant combination of microfluidics and photo-activated cAMP to analyze the translocation of GFP-tagged PH-domains in response to a short pulse of cAMP [36]. In this case, the PH-domain returned to pre-stimulus level after application and removal of the stimulus. Reconciling these seemingly contradictory results is not possible at this time, though one must bear in mind that two different assays were used and that different responses were measured. Both the production and secretion of cAMP, as well as the translocation of PH-domains are responses that are downstream of receptor signaling. More importantly, these are now considered to be downstream of the adaptation mechanism in *D. discoideum*. For example, a recent model postulates that the adaptation mechanism biases a second excitable network that controls downstream events like PH-domain translocation [37]. In this model, the return of PH-domains to the cytosol 10–15 s after application of the stimulus may reflect the refractory period of the excitable network rather than the time frame of adaptation, which is in the scale of minutes.

5.2 *Responding to More Complex Stimuli*

So far, all our attention has been on the response to step changes in the concentration of stimulus. While relatively easy to impose in the lab, it is unlikely that cells experience these changes in their native environment. It is worth asking how these systems respond to these more complex temporal stimuli.

We first consider the effect of a temporal ramp, that is, a linear increase in the concentration of ligand over time, as in the experiments carried out by Berg and co-workers [8, 38]. These experiments compensate for a feature of the *E. coli* receptor cluster which responds to logarithmic changes in concentration. Thus, an exponentially increasing ramp increases the chemoreceptor occupancy

linearly [38]. This increase induced a cellular response that reached a steady-state level, which is consistent with the notion that the system acts as a low-pass filtered differentiator [6, 8]. Exponentially varying sinusoidal inputs also gave rise to sinusoidal outputs, further confirming this connection. These responses matched a simplified model of the integral control feedback model of Barkai–Leibler that incorporates logarithmic inputs [7].

Interestingly, the response of the incoherent feedforward model to ramp changes in chemoattractant adapts [39], suggesting that the system behaves as a double integrator.

We note that the use of this frequency-domain analysis can be helpful as a means of studying the adaptive responses of biological circuits that are not as well understood as the bacterial chemotaxis pathway. For example, in a recent study of the high-osmolarity glycerol mitogen-activated protein kinase cascade in the *Saccharomyces cerevisiae*, oscillatory stimuli of different frequencies were applied and the response at these frequencies was used to elucidate a model of the mechanism [40].

6 Transient Response: Fold-Change Detection and Weber's Law

Adaptation is a steady-state phenomenon. Of related interest is transient behavior of sensory mechanisms to changes in the level of stimulus. Perhaps the earliest studies into this behavior are those of Ernst Weber, who noted that the response of sensory mechanisms was proportional to relative changes in the stimulus [41]. These findings have given rise to *Weber's law*, which states that the maximum change in output, that is, the sensitivity, is proportional to a change in input relative to the background level. We write this as:

$$y_{\mathrm{f}} - y_{\mathrm{i}} \propto \frac{u_{\mathrm{f}} - u_{\mathrm{i}}}{u_{\mathrm{i}}}.$$

More recently, Alon and co-workers introduced a related concept: *fold-change detection* (FCD): a response that is completely insensitive to fold changes in the input [27, 42] (Fig. 29.6). Whereas Weber's law refers to the initial or maximal response of the system, fold change detection compares the complete time history of the response. If the system is described by:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, y, u),$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(x, y, u),$$

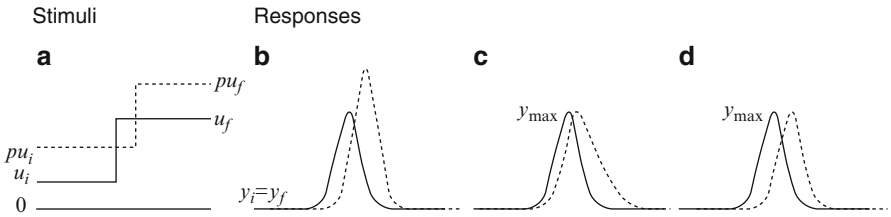


Fig. 29.6 Fold change detection. (**a**) Consider two step changes in stimulus: from u_i to u_f and from pu_i to pu_f. The latter is a p-fold larger change. (It has been delayed so as to highlight the different responses.) (**b**) The observed responses achieve perfect adaptation as they both settle to the same prestimulus levels. (**c**) These responses also adapt perfectly, but also satisfy Weber's law, as the peak is the same. (**d**) These responses achieve fold change detection. Not only are they perfectly adapting and satisfy Weber's law, but in fact the response to each stimulus is the same

then fold-change detection implies that

$$f(px, y, pu) = pf(x, y, u) \quad \text{and} \quad g(px, y, pu) = g(x, y, u).$$

It is worth asking which systems achieve fold-change detection. One example is the incoherent feedforward loop (29.16), which we rewrite as:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = u - x$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = u/x - y.$$

Because the differential equation for x is linear, it follows that changing the input to pu also gives rise to a change in x to px. In the equation for the response, y, the ratio

$$\frac{u}{x} = \frac{pu}{px},$$

and hence this equation is invariant with respect to p-fold changes in the input. We contrast this with a simplified version of the sniffer incoherent feedforward loop (29.15), in which the output equation is replaced by:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = u - xy.$$

Here, a p-fold change in u with corresponding p-change in x gives rise to:

$$\frac{\mathrm{d}\bar{y}}{\mathrm{d}t} = (pu) - (px)\bar{y},$$

which implies that

$$\bar{y}(t) = y(t/p).$$

Note that because the peak in $\bar{y}$ equals that of y, this sniffer system satisfies Weber's law, but does not achieve fold change detection.

Similarly, the integral control feedback loop of (29.11) does not achieve fold change detection. However, if the input to the system is log-transformed, as is the case in the chemotactic pathway [7, 8], then fold change detection is recovered.

7 Discussion and Open Questions

We have presented two general classes of robustly adapting networks, the integral control feedback network and the incoherent feedforward loop. While it is reasonable to expect that biological systems will employ robust mechanisms to achieve adaptation [13], this has only been shown experimentally in the bacterial chemotactic network [17]. Similar experimentation is needed in other systems. A difficulty, however, is that the details of these other mechanisms are not as well understood, and hence experimental perturbation of the networks is more difficult to address. For example, though the adaptation mechanism in the chemoattractant-mediated response in *D. discoideum* has been studied for over three decades and the response is relatively well characterized, to date the actual mechanism by which adaptation is achieved remains an open question.

It is worth pointing out that the two schemes presented here are essentially the only two motifs that achieve perfect adaptation [9]. Nevertheless, within these two general classes, there are numerous choices that can give rise to different behavior. For example, different incoherent feedforward loops can give rise to fold change detection, but others do not. Similarly, some of these networks can give rise to a biphasic response in the presence of feedback, but others cannot [31]. These differences may be useful as a means of discriminating amongst different putative networks based on experimental observations. This will require extensive experimentation using different temporal stimuli, including steps of different sizes, both positive and negative.

As highlighted by the discrepancy between the experimental results studying deadaptation in *D. discoideum*, matching of experimental data to models, is complicated unless there exists a strong connection between the adaptation mechanism and the observed variable in the experimental assay.

In describing the response regulator model, Koshland noted that this would serve as an input to a threshold detector [22]. In the bacterial chemotactic network, this happens at the level of the flagellar motor. The need to amplify the adapted signal has also been well documented in *D. discoideum*, though the precise method of amplifying the response is less clear [23, 37].

Despite these difficulties, we believe that studies of adaptation provide an ideal vehicle for the systems biology community combining modeling, theory, and experimentation in the study of a ubiquitous and fundamental cellular process.

References

1. Hood DC (1998) Lower-level visual processing and models of light adaptation. Ann Rev Psychol 49:503–535
2. Vladimirov N, Sourjik V (2009) Chemotaxis: how bacteria use memory. Biol Chem 390:1097–1104
3. Roberts MA, Papachristodoulou A, Armitage JP (2010) Adaptation and control circuits in bacterial chemotaxis. Biochem Soc Trans 38:1265–1269
4. Swaney KF, Huang CH, Devreotes PN (2010) Eukaryotic chemotaxis: a network of signaling pathways controls motility, directional sensing, and polarity. Ann Rev Biophys 39:265–289
5. Wang Y, Chen CL, Iijima M (2011) Signaling mechanisms for chemotaxis. Dev Growth Differ 53:495–502
6. Andrews BW, Yi TM, Iglesias PA (2006) Optimal noise filtering in the chemotactic response of *Escherichia coli*. PLoS Comput Biol 2:e154
7. Tu Y, Shimizu TS, Berg HC (2008) Modeling the chemotactic response of *Escherichia coli* to time-varying stimuli. Proc Natl Acad Sci USA 105:14855–14860
8. Shimizu TS, Tu Y, Berg HC (2010) A modular gradient-sensing network for chemotaxis in *Escherichia coli* revealed by responses to time-varying stimuli. Mol Syst Biol 6:382
9. Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. Cell 138:760–773
10. Segel LA, Goldbeter A, Devreotes PN, Knox BE (1986) A mechanism for exact sensory adaptation based on receptor modification. J Theor Biol 120:151–179
11. Knox BE, Devreotes PN, Goldbeter A, Segel LA (1986) A molecular mechanism for sensory adaptation based on ligand-induced receptor modification. Proc Natl Acad Sci USA 83:2345–2349
12. Francis BA (1980) On robustness of the stability of feedback systems. IEEE Trans Autom Control 25(4):817–818
13. Csete ME, Doyle JC (2002) Reverse engineering of biological complexity. Science 295:1664–1669
14. Spiro PA, Parkinson JS, Othmer HG (1997) A model of excitation and adaptation in bacterial chemotaxis. Proc Natl Acad Sci USA 94:7263–7268
15. Yi TM, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc Natl Acad Sci USA 97:4649–4653
16. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913–917
17. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168–171
18. Francis BA, Wonham WM (1975) The internal model principle for linear multivariable regulators. Appl Math Optim 2(2):170–194
19. Sontag ED (2003) Adaptation and regulation with signal detection implies internal model. Syst Control Lett 50(2):119–126
20. Andrews BW, Sontag ED, Iglesias PA (2006) Signal detection and approximate adaptation implies an approximate internal model. In: Proc 45th IEEE conference on decision and control, art. no. 4177419, pp 2364–2369
21. Andrews BW, Sontag ED, Iglesias PA (2008) An approximate internal model principle: applications to nonlinear models of biological systems. In: Proc 17th IFAC world congress 17, DOI:10.3182/20080706-5-KR-1001.0568
22. Koshland DE (1977) A response regulator model in a simple sensory system. Science 196:1055–1063
23. Levchenko A, Iglesias PA (2002) Models of eukaryotic gradient sensing: application to chemotaxis of amoebae and neutrophils. Biophys J 82:50–63
24. Ma'ayan A, Jenkins AL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong EJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309:1078–1083

25. Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. J Mol Biol 356:1073–1081
26. Cournac A, Sepulchre JA (2009) Simple molecular networks that respond optimally to time-periodic stimulation. BMC Syst Biol 3:29
27. Goentoro L, Shoval O, Kirschner MW, Alon U (2009) The incoherent feedforward loop can provide fold-change detection in gene regulation. Mol Cell 36:894–899
28. MacGillavry HD, Stam FJ, Sassen MM, Kegel L, Hendriks WT, Verhaagen J, Smit AB, van Kesteren RE (2009) NFIL3 and cAMP response element-binding protein form a transcriptional feedforward loop that controls neuronal regeneration-associated gene expression. J Neurosci 29:15542–15550
29. Osella M, Bosia C, Cora D, Caselle M (2011) The role of incoherent microRNA-mediated feedforward loops in noise buffering. PLoS Comput Biol 7:e1001101
30. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles, and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
31. Yang L, Iglesias PA (2006) Positive feedback may cause the biphasic response observed in the chemoattractant-induced response of *Dictyostelium* cells. Syst Control Lett 55:329–337
32. Krishnan J, Iglesias PA (2003) Analysis of the signal transduction properties of a module of spatial sensing in eukaryotic chemotaxis. Bull Math Biol 65:95–128
33. Sontag ED (2010) Remarks on feedforward circuits, adaptation, and pulse memory. IET Syst Biol 4:39–51
34. Devreotes PN, Steck TL (1979) Cyclic 3',5' AMP relay in *Dictyostelium discoideum*. II. Requirements for the initiation and termination of the response. J Cell Biol 80:300–309
35. Dinauer MC, Steck TL, Devreotes PN (1980) Cyclic 3',5'-AMP relay in *Dictyostelium discoideum* IV. Recovery of the cAMP signaling response after adaptation to cAMP. J Cell Biol 86:545–553
36. Beta B, Wyatt D, Rappel WJ, Bodenschatz E (2007) Flow photolysis for spatiotemporal stimulation of single cells. Anal Chem 79:3940–3944
37. Xiong Y, Huang CH, Iglesias PA, Devreotes PN (2010) Cells navigate with a local-excitation, global-inhibition-biased excitable network. Proc Natl Acad Sci USA 107:17079–17086
38. Block SM, Segall JE, Berg HC (1983) Adaptation kinetics in bacterial chemotaxis. J Bacteriol 154:312–323
39. Krishnan J (2011) Effects of saturation and enzyme limitation in feedforward adaptive signal transduction. IET Syst Biol 5:208
40. Mettetal JT, Muzzey D, Gomez-Uribe C, van Oudenaarden A (2008) The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. Science 319:482–484
41. Ferrell JE (2009) Signaling motifs and Weber's law. Mol Cell 36:724–727
42. Shoval O, Goentoro L, Hart Y, Mayo A, Sontag E, Alon U (2010) Fold-change detection and scalar symmetry of sensory input fields. Proc Natl Acad Sci USA 107:15995–16000

Chapter 30
Leveraging Modeling Approaches: Reaction Networks and Rules

Michael L. Blinov and Ion I. Moraru

Abstract We have witnessed an explosive growth in research involving mathematical models and computer simulations of intracellular molecular interactions, ranging from metabolic pathways to signaling and gene regulatory networks. Many software tools have been developed to aid in the study of such biological systems, some of which have a wealth of features for model building and visualization, and powerful capabilities for simulation and data analysis. Novel high-resolution and/or high-throughput experimental techniques have led to an abundance of qualitative and quantitative data related to the spatiotemporal distribution of molecules and complexes, their interactions kinetics, and functional modifications. Based on this information, computational biology researchers are attempting to build larger and more detailed models. However, this has proved to be a major challenge. Traditionally, modeling tools require the explicit specification of all molecular species and interactions in a model, which can quickly become a major limitation in the case of complex networks – the number of ways biomolecules can combine to form multimolecular complexes can be combinatorially large. Recently, a new breed of software tools has been created to address the problems faced when building models marked by combinatorial complexity. These have a different approach for model specification, using reaction rules and species patterns. Here we compare the traditional modeling approach with the new rule-based methods. We make a case for combining the capabilities of conventional simulation software with the unique features and flexibility of a rule-based approach in a single software platform for building models of molecular interaction networks.

M.L. Blinov (✉) • I.I. Moraru
Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA
e-mail: blinov@uchc.edu; moraru@neuron.uchc.edu

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_30,

1 Models of Reaction Networks

Modelers usually create a model of cellular processes by explicit specification of a reaction network consisting of molecular species and reactions. This is currently the most common paradigm implemented in model building and simulation software such as VCell ([1, 2], http://vcell.org), CellDesigner ([3], http://celldesigner.org), Copasi ([4], http://copasi.org), ECell (http://e-vell.org), MCell (http://www.mcell.cnl.salk.edu), and others. Each species has to be created, named, and reactions specified and assigned to the appropriate compartment within the cell. For each interaction described in the model, a user chooses the appropriate kinetic formalism and inputs relevant parameter values. A model usually includes molecular species corresponding to experimentally identified or hypothesized events, such as ligand–receptor binding, phosphorylation events, etc. Such models can fit experimental data and provide useful predictions. After a reaction network is specified, it can be simulated in order to identify time-courses for species.

Figure 30.1a illustrates a simplified version of seminal model of signaling by epidermal growth factor (EGF) receptor (EGFR) developed by Kholodenko et al. [5]. The major assumptions and elements of reaction network used in this model were later reused is some other modeling studies [6,7]. This reaction network (which is typical to many studies of receptor-initiated signal transduction) is comprised of reactions for binding of extracellular ligand (L) to cell-surface receptor (R), ligand-induced (LR) dimerization of receptors (D), phosphorylation of cytoplasmic

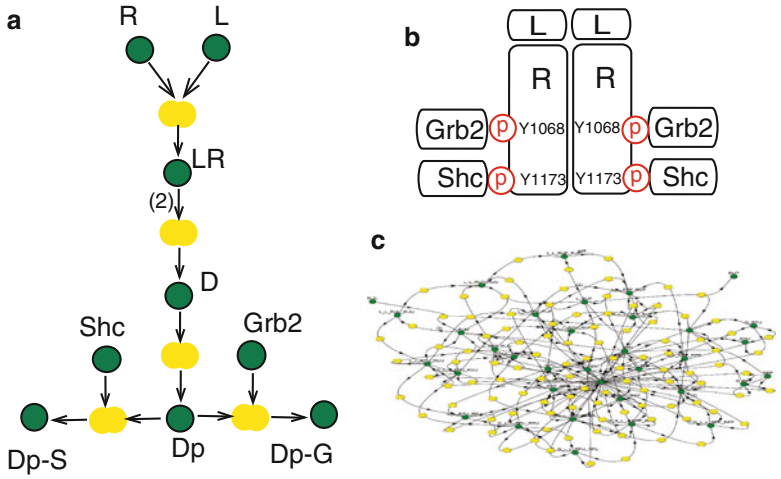


Fig. 30.1 (**a**) A bipartite graph (VCell notations: *green nodes* correspond to species, *yellow* nodes – to reactions) representing a model of initial events in EGFR signaling. (**b**) The potential protein complex arising during EGFR dimer signaling. (**c**) The Vcell model corresponding to the model describing interactions among all protein complexes looks like a maze consisting of *green* and *yellow dots*

receptor tyrosines (Dp) by the intrinsic protein tyrosine kinase, and competitive binding of adapter proteins Grb2 and Shc to a dimer (at most one protein per dimer).

The main paradigm in the reaction network approach is that every species represents a *pool*, a set of "things" that are indistinguishable from the standpoint of the processes (reactions) in which they participate. It provides a convenient way to visualize the molecular interactions as a graph where each species is a node. It also provides a unique and convenient mapping of the model to a mathematical description, where concentration of each species is a variable in time, which can be used to simulate the evolution of the system over time.

1.1 Limitations of the Reaction Network Approach

This approach of describing in detail all elements of a reaction network has several obvious limitations. Whether or not the reaction network is specified manually by the modeler, or through some computer-aided process (such as automated import from a pathway database [8]), it may include only a limited number of species and reactions. Thus, the model is usually based on mechanistic assumptions that limit the size of the reaction network.

Consider, for example, the EGFR signaling network described above. It includes many simplifications, such as omitting events like: ligand binding to cytosolically modified (phosphorylated at some combination of residues) receptor, ligand dissociation from the receptor in a dimer, dissociation of phosphorylated receptors in a dimer, multiple proteins bound to distinct phosphorylated receptors residues at the same time, etc. Additionally, lumping tyrosines of both receptor molecules in a dimer means that in the model they are phosphorylated and dephosphorylated simultaneously, and excludes the possibility that modification of individual tyrosines during signaling may affect the signaling outcome.

One might surmise that molecular complexes are often not considered in detail simply because of Occam's razor concept. As long as the model predictions match experimental observations, one can use simplifying assumptions and omit many details deemed unnecessary. However, these simplifications are often not motivated by experimental considerations. In [9], we reviewed the evidences contradicting some of these assumptions. Individual tyrosines of EGFR may have distinct temporal patterns of phosphorylation, Grb2 and Shc may be simultaneously associated with a single copy of EGFR, which is consistent with the observed nucleation of large heterogeneous protein complexes in other systems, and receptor monomers may be responsible for the spatial spread of receptor phosphorylation observed in response to localized EGF stimulation and therefore involved in signaling. Thus, it is conceivable that distinct combinations of phosphorylated receptor complexes may have distinct functions. Despite this evidence, simplified assumptions preventing formation of receptor dimers with multiple adaptor proteins bound to both receptors and receptor monomers have been used in most modeling studies of EGF receptor signaling.

The main reason to include simplifying assumptions is that without them, many more molecular species and reactions must be considered. For example, if we allow adapter proteins to simultaneously bind to distinct phosphorylated binding sites on receptors and allow receptors in dimers to dissociate, we need to include into the model all multiple forms of monomeric and dimeric EGFR complexes (like the one shown in Fig. 30.1b). Thus, we need to account for 93 species (12 monomeric and 78 dimeric receptor complexes, and proteins EGF, Grb2, and Shc). One can see (Fig. 30.1c) that such model is not easily tractable in a regular reaction editor (here shown in VCell). Furthermore, note that the number of species corresponding to multiple phosphoforms of receptor stacks up explosively: to track phosphorylation of the nine tyrosines of EGFR, one needs to account for the $2^9 = 512$ different phosphorylation states of an individual receptor and the 131,328 distinct combinations of phosphorylation states of receptors in a dimer. Of course, the true scope of such complexity is uncertain and may lie well below these theoretical maximal numbers due to various constraints, such as steric clashes, that might play a role in limiting the combinatoric possibilities in signal-transduction systems. But one would need to have the capability to handle very large number of species and reactions in order for models to be able to capture critical features of variability in signaling [10].

Another practical problem is that such models based on simplified assumptions usually involve species that lump together entities that correspond to specific experimental measurements, and a new model may be required to describe each new set of measurements. A related issue is the fact that even if one would like to include all molecular details of protein complexes in the model, there often is a lack of knowledge of the required detailed mechanisms of interactions and related kinetic parameters. However, new high-throughput flow cytometry and mass spectrometry measurements provide a wealth of information about interactions and activities of proteins [11], which now can be (and should be) included in some models.

2 Rule-Based Models

An alternative to the conventional modeling approach, that has been gaining increasing acceptance, is to attempt to create a model description that is capable of accounting for all the potential molecular complexes and interactions among them that can be generated during a response to a signal. A feasible strategy to implement this is a rule-based approach [12]. In this approach, protein–protein interactions and their effects are represented in the form of reaction rules that serve as generators of chemical species and reactions. This method, discussed in more detail in [12–14], provides an opportunity to consider the whole nomenclature of potential protein complexes, including their phosphoforms, modifications, and interactions that can potentially be generated during the response to signaling. Moreover, this approach also allows exclusion of those species and reactions that cannot be realized, e.g., because of cooperativity and steric clashes.

The rule-based approach has been initially developed based on the modularity of protein domains [15]. A model is specified as a set of reaction rules, which are associated with specific rate laws. Given a set of species, a reaction rule identifies those species that have the features required to undergo the transformation from reactants to products specified in the reaction rule. Interactions represented in a reaction rule do not depend on features not explicitly indicated. Thus, multiple species may qualify as reactants in a type of reaction defined by a reaction rule.

The modeler can define which components and modifications of a molecule or molecular assembly affect a particular chemical transformation, and which do not. Furthermore, the modeler has the ability to account for steric clashes, cooperativity, and any other factors that might influence the rate of a reaction. A reaction rule can state, for example, that "any cell-surface monomeric receptor having an available extracellular binding site and any free extracellular ligand can interact and form a ligand–receptor complex; the probability of this interaction depends only on the total numbers of cell-surface monomeric receptors and extracellular ligands and does not depend on the specific state of the receptor cytosolic portion." In this example, we assume that the cytoplasmic state of a receptor does not affect ligand–receptor binding, which implies that to parameterize all reactions specified by the ligand–receptor interaction rule, we need just two rate constants: on and off rates. Biophysicists would argue that any cytosolic modification will definitely affect all portions of a receptor, and hence must affect these rates. However, as we have shown [16] the model read-outs are relatively robust to parameter variations within the same reaction rule. Thus, in practice, the number of reaction rules (and rate constants) that the user must provide to specify a model is comparable to the number of assumptions about interactions among molecular domains considered in the model, which is usually much less than the total number of actual reactions. Moreover, the number of rate constants that the model is built upon can be limited to those that come directly from experiments, such as in vitro binding affinities for multiple SH2/PTB domains and tyrosines.

Moreover, when the user changes a model to include new assumptions about mechanisms of molecular interactions (such as replacing competitive binding of proteins to a scaffold with cooperative binding), these rate constants remain unchanged. This is in a contrast with a reaction network model where variables representing lumped entities often require adjusted kinetic laws.

2.1 Simulation Methods for Rule-Based Models

A model where the reaction network is explicitly specified always allows one to directly derive a unique mathematical formulation as a set of differential equations in variables corresponding to species concentrations/population numbers. The same is not true for rule-based models – they do not immediately provide the formalism required for simulation and they need pre-processing. In some cases, the rule-based model can be expanded into an explicit reaction network, such as by using

an iterative algorithm for processing reaction rules [13, 14] (e.g., the algorithm implemented in the BioNetGen software – which is, in fact, the origin of the name: Biological Network Generator). The iterations of rule application are halted when specified termination conditions (like reaching a predefined size of an oligomer) are satisfied or all possible reactions have been generated. The exact size of the generated reaction network depends, in general, on the entire set of reaction rules and also on the set of species to which reaction rules are initially applied.

Sometimes a reaction network can be of potentially unlimited size, such as when reaction rules provide a way for infinite elongation of molecular chains, e.g., while specifying actin filaments. In this scenario, "on-the-fly" network generation can be used. Reaction rule evaluation is embedded in a discrete-event Monte Carlo simulation of reaction kinetics, and reactions are generated only when a species is first populated during a simulation [14, 17].

However, the on-the-fly method still requires network generation: a product of a reaction generated by a reaction rule has to be identified either as a new species or as the species that was previously generated and already in the reaction network. This becomes a serious computational problem when the generated reaction network contains species that have complicated topological structures, such as species representing branched actin filaments. To deal with it, a "network free simulation" approach [18–20] was recently introduced. In this method, a model does not have to have its reaction network generated prior to and/or during simulation steps. Individual instances of possible species and interactions are accounted for and reaction rules are evaluated directly during the simulation. Depending on the size and complexity of the system, network-free simulation can be much more effective from a computational standpoint, although it is limited to using discrete simulation algorithms.

2.2 Graphical Representations for Rule-Based Models

One obstacle to the acceptance of rule-based modeling is the unusual way such models are being specified. Modelers tend to think of a model as a pathway or a reaction sequence, where the product of one reaction is used as a reactant (or a catalyst) in another reaction. Reaction rules are intrinsically disconnected, because not all of the multiple species that are products of a certain reaction rule participate in another reaction. Products and reactants of the reaction rule are no more pools of identical things, like species nodes in reaction networks. For example, let Rule 1 be the ligand binding to a receptor with three intracellular binding sites. Let Rules 2–4 be the independent binding of different adapter proteins to each of the three receptor binding sites. Thus, all product species of a ligand–receptor binding rule are divided into three intersecting subsets that can participate in three reaction rules of protein binding (Fig. 30.2).

To describe rule-based models, several non-network-based representations are used. One way of model specification is using a specially designed scripting

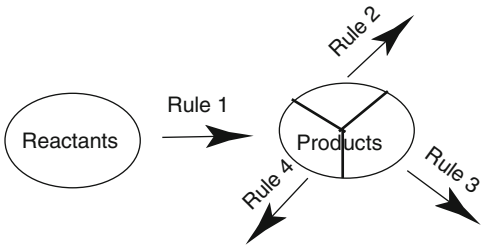


Fig. 30.2 Visualization of rule-based modeling as a reaction network is difficult. Each node is not a pool of identical things but corresponds to a set of species, each of which can participate in different rules

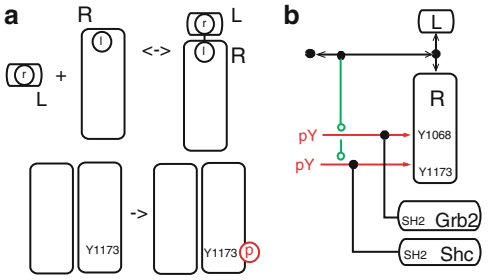


Fig. 30.3 Graphical representations of rule-based models. (**a**) Two rules are represented as cartoons that show only relevant features for each interaction. In the first rule, a ligand L can bind any receptor R provided binding sites l and r or R and L, respectively, are unbound. The second rule says that tyrosine Y1173 can be phosphorylated provided there is another receptor in proximity. (**b**) Molecular interaction map (MIM) representation of all interactions in Fig. 30.1a. Although compact, it might be ambiguous and the temporal order of interactions is difficult to infer

language, such as BioNetGen language (BNGL, [21]) or Kappa language [22] This approach requires intimate knowledge of such specialized languages, and thus is typically used only by advanced users.

Another approach is to use a graphical specification of all reaction rules following certain conventions, as described in [23, 24]. In this approach each molecular entity (protein, receptor, DNA, etc.) in a model is specified as a box containing components that denote features of a molecule. Several boxes can be joined to form a species or species pattern by connecting components. Each reaction rule is specified as a separate cartoon describing reactants and products (Fig. 30.3a).

Yet another approach to specify rule-based models is using cartoons representing interactions among molecular entities and their components, such as entity relationship diagrams in SBGN [25], molecular interaction maps [26] (Fig. 30.3b), or extended contact maps [27]. However, specification of rule-based models in this way is often ambiguous as the temporal order of interactions is difficult to infer [28]. This approach is used mostly to supplement the model-building process using a scripting language [29].

As a conclusion, all current approaches for specification of rule-based models are distinct from the usual model specification as a reaction network and thus require special training in rule-based modeling in order to be used. However, with all these

shortcomings, and despite being relatively new, the rule-based modeling approach has been used to develop a wide range of models [9, 30–34]. Several software tools with some rule-based modeling capabilities have been developed in recent years, including BioNetGen [13, 21], STOCHSIM [35], Moleculizer [17], K-factory [22], and Simmune [36].

3 Merging Reaction Network and Rule-Based Models

The reaction network and rule-based modeling approaches are complementary and have distinct representation schemas. The reaction network approach has the advantage of often being in one-to-one correspondence with cartoons representing signaling pathways and a defined mathematical representation. However, these advantages fade away as more and more details are included into the model, as very large reaction networks become cluttered and difficult to deal with. The rule-based approach has the advantage of being able to account for all details of molecular activities and interactions, but the complete overview of the biological system evolution may not be apparent until network generation or model simulation is performed.

It would be of enormous advantage to the modeling community if these two approaches would seamlessly work together. We have recently developed several prototype methodologies to use rule-based modeling alongside reaction network modeling, and to implement these two techniques into a common modeling and simulation interface.

3.1 Extending a Reaction Network by Adding Species Features

There are several classes of use cases where reaction rules can be organically used to extend existing models. Consider, for example, a model of a reaction network where the user wants to add a fluorescent tag to some features. For a minimal reaction network like $A + B \rightarrow C$, $C + D \rightarrow E$, if species A is fluorescently labeled, then the result of interaction of A with B, species C, will be fluorescently labeled as well. The fluorescence will be then passed to species E. If we now want to model a mixture of fluorescent and non-fluorescent species, we will need to double the size of original reaction network by adding extra reactions $Af + B \rightarrow Cf$, $Cf + D \rightarrow Ef$ (Fig. 30.4). This information makes the reaction network more cluttered and does not provide any new information, since often the kinetic behavior of fluorescent and non-fluorescent species is the same. Such an extension of the model can be easily described by introducing a "fluorescence" feature of molecules A, C, and E, and specifying that the value of this feature (fluorescent or non-fluorescent) is preserved when participating in reactions. Now each former reaction node becomes a reaction rule node, as it describes the same interaction for two different species: fluorescent

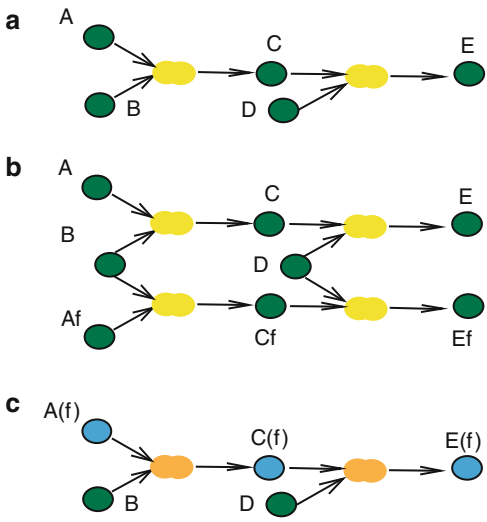


Fig. 30.4 An example of reaction network being extended by introducing fluorescent labeling. (**a**) The simple reaction network. (**b**) The network with species A, C, and E being fluorescently labeled. (**c**) The same reaction network displayed with two types of nodes – *green nodes* for individual species and *blue nodes* for species templates that have modifiable feature f (fluorescence). Reaction nodes become rule nodes and change color to *orange*

and non-fluorescent. Thus, the reactions become now reaction rules. This is a natural way to introduce rule-based modeling into a regular network.

Using this approach, a rule-based model can be build atop of a regular reaction network by converting some species and reactions into species patterns and reaction rules. In this scenario, each species can be extended into a species type by adding a set of features (attributes) and specifying allowable and default values of these features. Thus, a species is converted to a species template that defines a set of species. A second step is the conversion of a reaction to a reaction rule. A reaction where some of reactants or products are species templates becomes a reaction rule, as it is now applied to a set of species selected by a species template.

3.2 Extending a Reaction Network by Specifying Multimolecular Species

A more complicated case is to extend a reaction network that includes multimolecular species – for example the complex depicted in Fig. 30.1b where a transmembrane receptor R can bind extracellular ligand L, associate with another transmembrane receptor R, and bind two intracellular proteins Grb2 and Shc. To extend a reaction network into a rule-based model, the user can start from extending species L and R into species types. A ligand L becomes a species type by introducing a single feature "binding to R." A receptor R becomes a species type by introducing features "binding site for L," "phosphosite Y1," and "phosphosite Y2." Each feature may have several possible values, e.g., phosphosites can have values of "phosphorylated" or "unphosphorylated." Note that in BNGL species types are called molecules, and

features are called components. When all features are uniquely specified (for example, values of phosphosites are set to "unphosphorylated"), a species type includes "things" of the same kind, e.g., it represents a species. Thus, species type is more than species, in this example it includes at least four species representing different phosphoforms (those are species with sites Y1 and Y2 being unphosphorylated, phosphorylated, and pairwise different).

When the user follows a reaction network graph and sees a reaction arrow starting at two species extended into species type, the user has to convert the reaction into a reaction rule, and a product into a species template. Indeed, the species LR becomes a species template containing two species types L and R. To proceed, the user must specify a bond between components of L and R, by setting values of "binding site for R" to "bound to R" and "binding site for L" to "bound to L".

By traversing the reaction graph, each species node in a reaction diagram can be converted into a species template. While doing that, new features are introduced for each species type. For example, when converting dimerization reaction into a reaction rule, the user must specify how two ligand–receptor complexes are connected into the larger dimeric complex. To do it, a new feature "binding to R" must be specified for the species type R. Now each species type in the reaction of ligand–receptor binding becomes a species template, as receptor species type now has two features ("binding to L" and "binding to R"), and the second feature was not specified (ligand could potentially bind to a receptor connected to another receptor). As we noted in Fig. 30.2, the reaction network graph with nodes representing species templates does not represent a reaction network, as nodes do not represent identical "things" anymore. However, we still can use the reaction network graph with composite nodes in place of species templates. These nodes contain all species types used in a species template (Fig. 30.5). Here we follow the notations of [21] where the *center* and *content* of a reaction rule were introduced. The rule center contains all species types that have features changed during the interaction (for example, a binding site becomes bound or a phosphosite becomes phosphorylated), while content contains molecular entities that affect the interaction but remain unmodified. For example, in a dimer transphosphorylation rule, one receptor subject to phosphorylation belongs to a rule center, and another receptor that acts as a kinase and thus remains unmodified belongs to content. In a graphical presentation, a composite node for a reactant contains only reaction center.

Note that the reaction rule is valid only if it can be uniquely converted into a set of reactions. This is often not trivial. Consider the case where there is one reactant species template and one product species template, but each has a different number of unspecified features. The number of reactant species defined by the reactant species template will be different from the number of product species generated by the product species template. This makes one-to-one mapping impossible. Thus, conversion of a reaction to a reaction rule must be done with caution: selected reactant and product species must be converted to species templates, and a mapping between reactant species template and product species template has to be established.

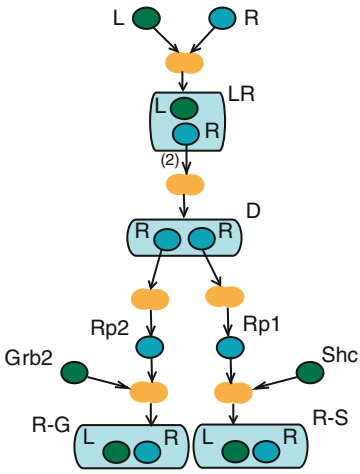


Fig. 30.5 The combination of reaction network view with rule-based modeling. *Green nodes* correspond to species, *blue nodes* to species templates. *Orange nodes* correspond to rules. *Arrows* connect molecular entities that are modified during interactions (rule centers)

3.3 Prototype for a Unified Modeling Interface

We are implementing the approach discussed above into the VCell modeling and simulation framework [1, 2]. By combining a reaction network specified in the VCell editor (explicitly specifying individual species and reactions) with multicomponent species and rules of interactions, a user should be able to use rule-based specifications within the familiar "look and feel" environment of the physiology editor. The new VCell interface introduced in version 5.0 (public beta release as of this writing) includes dual views for a reaction network: a set of tables describing all model elements, and a bipartite graph with species and reaction nodes. Both views can be extended to support rule-based features.

The tabular view is a good interface for extending an existing reaction networks into a rule-based model, or for creating a new rule-based model. The table can be used to specify features and possible feature states for species types, and drop-down menus are a good way to select feature states for species templates and specify bonds. Reaction rules can be specified in two tables that represent reactant and product parts of a rule.

The reaction network view can be used for visualizing a composite rule-based and network model. A mix of species and species patterns, reactions and rules, is illustrated in Figs. 30.4 and 30.5. The reaction network can be also "flattened," when all species and reactions are generated. Flattening is possible and provides essential information only when network generation is possible. In the flattened view, regular species and species generated from reaction rules are displayed as a usual bipartite graph, where each species carries all the features inherited from reaction rule specification. Thus, many modes of model visualization are possible – collapsing all species corresponding to certain reaction rule, displaying sub-network consisting only of species with a certain feature (like fluorescence), etc.

We aim to provide an expert system guiding users in building a rule-based model, providing suggestions on what features have to be introduced for each molecular entity and each interaction. Significant efforts are still required to introduce a convenient way to mix rule-based and network models.

4 Conclusions

Quantitative modeling studies have rapidly spread across many domains of biology in recent years and the scientific community has been putting a great deal of efforts into standardization. These efforts are crucial for more efficient and accurate transmission of biological knowledge between different communities in research, education, publishing, and more. Standards like Systems Biology Markup Language (SBML, [37]), Systems Biology Graphical Notations (SBGN, [25]), and Biological PAthway eXchange (BioPAX, [38]) are already widely used to provide exchange of models, visualization schemas, and pathway data, respectively. The goal is to provide interoperability between various methods and tools, as it has become clear that there is no single strategy and platform that can cover all needs.

Initially, all of these standards have been developed based on the conventional approach to model building. However, the rule-based modeling approach is gaining momentum, and the old paradigm that every species in a model, every node in a graph description and every entity in a database consist of identical "things" is phasing out. This has been recognized by the community, and each of the standards mentioned above has some capabilities to describe "generic" entities and elements of rule-based modeling. SBML has a package ("L3 multi") under development to enable description of multistate and multicomponent species and rules of interactions among them. SBGN and BioPAX have proposals to introduce generic entities that describe sets of species that may participate in multiple interactions.

The VCell simulation and modeling framework always strives to be on the leading edge of new technologies, be user-friendly, and be compatible with the community standards. It is often not easy, as in the case of rule-based modeling. However, adoption of the new standards will hopefully facilitate the development of more tools with mixed capabilities, just like our prototype of VCell-BioNetGen integration.

References

1. Slepchenko BM, Schaff JC, Macara I, Loew LM (2003) Quantitative cell biology with the virtual cell. Trends Cell Biol 13(11):570–576
2. Moraru II, Schaff JC, Slepchenko BM et al (2008) Virtual cell modelling and simulation software environment. IET Syst Biol 2(5):352–362

3. Funahashi A (2003) The ERATO systems biology workbench and systems biology markup language: an integrated environment and standardization for systems biology. Tanpakushitsu Kakusan Koso 48(7):810–816
4. Hoops S, Sahle S, Gauges R et al (2006) COPASI – a COmplex PAthway SImulator. Bioinformatics 22(24):3067–3074
5. Kholodenko BN, Demin OV, Moehren G, Hoek JB (1999) Quantification of short term signaling by the epidermal growth factor receptor. J Biol Chem 274(42):30169–30181
6. Hatakeyama M, Kimura S, Naka T et al (2003) A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. Biochem J 373(Pt 2):451–463
7. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol 20(4):370–375
8. Blinov ML, Ruebenacker O, Schaff JC, Moraru II (2010) Modeling without borders: creating and annotating VCell models using the web. Lecture Notes Bionform 6053:3–17
9. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. Biosystems 83(2–3):136–151
10. Mayer B, Blinov M, Loew L (2009) Molecular machines or pleiomorphic ensembles: signaling complexes revisited. J Biol 8(9):81
11. Schulze WX, Deng L, Mann M (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. Mol Syst Biol 1:2005.0008
12. Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W (2006) Rules for modeling signal-transduction systems. Sci STKE 2006(344):re6
13. Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2004) BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. Bioinformatics 20(17):3289–3291
14. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS (2005) Rule-based modeling of biochemical networks. Complexity 10:22–41
15. Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300(5618):445–452
16. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS (2005) Combinatorial complexity and dynamical restriction of network flows in signal transduction. Syst Biol 2(1):5–15
17. Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Moleculizer 1.0. Nat Biotechnol 23(1):131–136
18. Yang J, Monine MI, Faeder JR, Hlavacek WS (2008) Kinetic Monte Carlo method for rule-based modeling of biochemical networks. Phys Rev E: Stat Nonlinear Soft Matter Phys 78(3 Pt 1):031910
19. Colvin J, Monine MI, Faeder JR, Hlavacek WS, Von Hoff DD, Posner RG (2009) Simulation of large-scale rule-based models. Bioinformatics 25(7):910–917
20. Colvin J, Monine MI, Gutenkunst RN, Hlavacek WS, Von Hoff DD, Posner RG (2010) RuleMonkey: software for stochastic simulation of rule-based models. BMC Bioinform 11:404
21. Faeder JR, Blinov ML, Hlavacek WS (2009) Rule-based modeling of biochemical systems with BioNetGen. Meth Mol Biol 500:113–167
22. Danos V, Feret J, Fontana W, Krivine J (2007) Scalable simulation of cellular signaling networks. Lect Notes Comput Sci 4807:139–157
23. Blinov ML, Yang J, Faeder JR, Hlavacek WS (2006) Graph theory for rule-based modeling of biochemical networks. Trans Comput Syst Biol Vii 4230:89–106
24. Blinov ML, Yang J, Faeder JR, Hlavacek WS (2006) Depicting signaling cascades. Nat Biotechnol 24(2):137–138
25. Le Novere N, Hucka M, Mi H et al (2009) The systems biology graphical notation. Nat Biotechnol 27(8):735–741
26. Kohn KW (2001) Molecular interaction maps as information organizers and simulation guides. Chaos 11(1):84–97

27. Chylek LA, Hu B, Blinov ML et al (2011) Guidelines for visualizing and annotating rule-based models. Mol Biosyst 7(10):2779–2795
28. Kohn KW, Aladjem MI, Kim S, Weinstein JN, Pommier Y (2006) Depicting combinatorial complexity with the molecular interaction map notation. Mol Syst Biol 2:51
29. Xu W, Smith AM, Faeder JR, Marai GE (2011) RuleBender: a visual interface for rule-based modeling. Bioinformatics 27(12):1721–1722
30. An GC, Faeder JR (2009) Detailed qualitative dynamic knowledge representation using a BioNetGen model of TLR-4 signaling and preconditioning. Math Biosci 217(1):53–63
31. Faeder JR, Hlavacek WS, Reischl I, Blinov ML, Metzger H, Redondo A, Wofsy C, Goldstein B (2003) Investigation of early events in Fc epsilon RI-mediated signaling using a detailed mathematical model. J Immunol 170(7):3769–3781
32. Lipniacki T, Hat B, Faeder JR, Hlavacek WS (2008) Stochastic effects and bistability in T cell receptor signaling. J Theor Biol 254(1):110–122
33. Mu F, Williams RF, Unkefer CJ, Unkefer PJ, Faeder JR, Hlavacek WS (2007) Carbon-fate maps for metabolic reactions. Bioinformatics 23(23):3193–3199
34. Nag A, Monine MI, Blinov ML, Goldstein B (2010) A detailed mathematical model predicts that serial engagement of IgE-FcepsilonRI complexes can enhance Syk activation in mast cells. J Immunol 185(6):3268–3276
35. Le Novere N, Shimizu TS (2001) STOCHSIM: modelling of stochastic biomolecular processes. Bioinformatics 17(6):575–576
36. Meier-Schellersheim M, Klauschen F, Angermann B (2009) Computational modeling of signaling networks for eukaryotic chemosensing. Meth Mol Biol 571:507–526
37. Hucka M, Finney A, Sauro HM et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4):524–531
38. Demir E, Cary MP, Paley S et al (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28(9):935–942

Part V
Applications of Systems Biology in Medicine, Biotechnology and Pharmaceutical Industry

Chapter 31
Why and How to Expand the Role of Systems Biology in Pharmaceutical Research and Development

Robert D. Phair

Abstract Seen from the perspective of funding organizations, investors, and the general public, the productivity of our world-wide biomedical research enterprise is declining despite increased investment. This opinion piece suggests a cause and a solution. The cause is the enormous complexity of human biology and pathophysiology. The unsolved human diseases involve so many interacting variables that single research laboratories headed by skilled principal investigators doing innovative experimental work cannot be expected to assemble the reductionist pieces into an integrated working model. Systems biology offers a solution, but it will require teamwork. Co-equal teams of experimental and computational biologists can construct multiscale differential equation models and test them against experimental data. A successful model provides actionable evidence-based guidance to the entire research and development team. These integrative biology teams may, for historical and cultural reasons, be unsustainable in academia, but they seem naturally suited to modern pharmaceutical research and development. One way to organize such teams and their workflow is described in detail.

This opinion piece must begin with a disclaimer. The author is not and never has been an employee of a major pharmaceutical firm. What follows are views of a keenly interested physiologist/biomedical engineer/systems biologist who began his biological modeling work decades before the phrase "systems biology" entered the lexicon of modern science. We systems biologists emphatically do not share a single vision of how best to proceed with integrative biology. A good idea of the diversity of vision can be found in a sampling of recent reviews [1–10]. Nevertheless, we are united in our conviction that translating basic biomedical discovery into an actionable understanding of human disease will require the mathematical and computational tools of the physical and engineering sciences.

R.D. Phair (✉)
Integrative Bioinformatics Inc., Los Altos, CA 94024, USA
e-mail: rphair@integrativebioinformatics.com

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_31,

1 Complexity

Complexity: this single word motivates every sub-field of systems biology. No professional biomedical scientist can seriously maintain that 21st century biology will, without recourse to computation, achieve what the public expects of us. The complexity of biochemistry and molecular biology alone, much less cell biology, organ system physiology, and pathophysiology, demands computational tools.

Two primary approaches to complexity comprise the computational arms of systems biology. They are statistical pattern recognition and mechanistic modeling. We need both, but we also need serious efforts at communication across this boundary. Too often, statistical, unbiased, high throughput approaches are seen as incompatible with mechanistic modeling when, in fact, such methods often yield unimagined novel hypotheses that deserve mechanistic tests. Similarly, mechanistic models are too often seen as "not scalable" to 100s or 1000s of variables when, in fact, models of this size are now commonplace. Both views are at odds with the facts: statistics and modeling are humankind's premier tools for making complex hypotheses testable.

Hypotheses regarding complex human diseases are at the top of our list. There are thousands upon thousands of contributions to the biomedical research literature and each proclaims a key protein, a central gene, an essential signaling pathway, or a pivotal physiological control system. Complex diseases – diseases like atherosclerosis, stroke, autoimmune disease, cancer, metabolic syndrome, neurological disease, diabetes, obesity, infectious diseases, and heart failure – are polygenic and responsive to a multitude of environmental inputs. How should we organize the research enterprise so that all those thousands of individual contributions can best be leveraged to produce an actual therapeutic benefit?

2 Specialization

Complexity forces specialization. Specialization fosters depth. Depth in research is not only revered, it is essential. Expertise in any scientific field requires a view that spans from the big picture all the way to the details. The only way to achieve this level of expertise in our era is to focus. In other words, we have to know "more and more about less and less."

Unfortunately, this necessity to specialize – this fact of modern scientific life – erects walls between disciplines; it blocks cross-fertilization and integration. Almost nothing (perhaps only the joy of science) motivates a scientist to attend a meeting where he or she is not an expert. Specialization, though absolutely essential, is the polar opposite of integration. How are we, as specialists, going to assemble the pieces of a complex pathophysiological puzzle?

3 Synthesis

Everyone agrees on the importance of synthetic, integrative work. This is why a thoughtful review article is so important to and so widely read by a community of scholars. But the review literature is rarely actionable. First, the authors have, by necessity, chosen a subset of the relevant literature – usually the papers published since the last roughly similar review.

Second, the summary diagram – perhaps the most widely studied portion of any review – is limited to a single (hopefully) consistent perspective. Other members of the same community of scholars regularly advance different summary diagrams based on a different reading of the same literature. Hence, the review literature is not actionable because each reviewer's synthesis is, in effect, a large-scale hypothesis. It represents an honest, often brilliant, effort to imagine a system that is consistent with all the reviewed published data plus the enormous store of unpublished information that each scientist uniquely accumulates.

Probably no single expert laboratory has the resources to test all the implications of such a large-scale hypothesis. Yet such tests are the essential prerequisite for an actionable working model (AWM) of human biology – an evidence-based model that can be used as the basis for pharmaceutical development. So we are faced with an apparent systemic design flaw. Scientific expertise requires depth. Depth requires specialization. Specialization is the antithesis of synthesis and integration, but synthesis is vital to the 21st century mission of publicly funded biomedical research and privately funded pharmaceutical development. Once again, the question devolves to this: how should we manage synthesis of all the reductionist results extracted from nature by legions of tenacious experts?

4 Teams

Teamwork is, I think, the only way to get both expertise and coverage – both depth and breadth. We need principal investigator/lab chief-level expertise in all aspects of the integrative biology endeavor. Paradoxically, this may not be possible in academia. Universities and colleges and research institutes are consciously designed to promote individual genius, not team genius. It has never been easy to build teams in academia. Ask any academic dean or any imaginative NIH program director. Convincing professors to work together toward a shared goal and a shared reward, they will say, is like herding cats. There have been exceptional teams in the academic world, but their rarity suggests the need for an expansion of the classic (and extremely successful) Vannevar Bush academic entrepreneurial business model. Our universities are ideally organized for reductionist discovery. Their track record over the past millennium is a testament to what humankind can accomplish. There is no sense in changing something that works so well.

Who, then, has both the motive and the human resources to nurture teams of integrative or systems biologists? A public organization like the newly proposed National Center for Advancing Translational Science (NCATS) at the United States National Institutes of Health might undertake this challenge, but I think the world's pharmaceutical companies are the obvious candidates.

5 Two Cultures Redux

Snow's thesis [11], though popularly defined by the unfortunate gulf that now separates the sciences from the humanities, was actually a critique of specialization. He particularly lamented our increasing inability to communicate across disciplinary boundaries. A modern pharmaceutical firm, however, has the profit motive to design and drive adoption of a 21st century synthesis-driven approach. To an outsider, promoting cooperation between systems and experimental biologists may seem a simple task, but the two-cultures division between those for whom mathematics is the natural language of nature and those for whom, as I was once admonished by a famous senior scientist, "Physiology (or substitute your favorite biomedical discipline) is an experimental science and mathematics has no place in it," is still in evidence.

This is changing, but the pace of change is only slightly faster than evolution. To speed the pace, we need better communication across the divide. This will require cross-education; the systems biologists must understand the experimentalist's results at least at the level of a current review article and ideally at the level of experimental methods. Experimentalists, on the other hand, need to understand more and more clearly how statistics and modeling work with complexity. And everyone on the team should know basic physiology.

One approach to this communication problem that is gaining momentum is the increased availability of both commercial and academic software tools that take advantage of diagrams as a common language – a language that can be understood by both experimental and systems biologists. So important is this approach that multiple standards for the creation of such diagrams are being promulgated [12–15]. Cooperation among these software teams has been laudable and there seems little doubt that a diagram standard will emerge that satisfies the systems biologists. If there is any weakness in this joint effort, it is that few experimental biologists, or their professional societies, have been involved. But within a single pharmaceutical firm, it should be possible for internal debate to identify and adopt one diagram language company-wide. This would be a major step toward building a cohesive integrative biology team.

6 Integrate Pharma M&S into Line Management

Typically, large pharmaceutical firms already employ a nucleus of modeling and simulation (M&S) experts. Often these organizational units emerged from the paramount importance of pharmacokinetics in establishing a dosing regimen and in the regulatory drug approval process. Increasingly, however, M&S input is sought in early discovery and development efforts. From the perspective of building a 21st century integrative biomedical development team, M&S must be elevated from the position of a service organization to co-equal status with the biology and medicine-based discovery and development teams.

This would necessitate a dramatic change in mindset and substantially greater responsibility. In order to participate in the firm's most pivotal decisions, M&S leadership will need to understand in substantial detail the biochemical and physiological rationale for each drug target and each drug candidate. Indeed, a co-equal M&S organizational unit will have contributed substantially to that rationale.

Profound increases in organizational responsibility are always attended by increased reward and increased risk. Line managers become highly valued and rewarded employees when products succeed. Line managers must change jobs when products fail. It is the nature of the pharmaceutical industry that unanticipated clinical outcomes can change your career overnight. Building integrative biology teams – teams that are managed so that both experimental and computational expertise is leveraged – is arguably the best strategy for a 21st century pharmaceutical firm whose target therapeutic areas are the major diseases. These diseases are enormously complex. Co-equal teams of experimental and computational biologists are going to be essential.

It might be argued that academic collaborations are just as likely to succeed. Teamwork, however, is different from collaboration. Collaboration is a negotiated bargain; teamwork spreads responsibility and credit equally. Credit is the coin of the realm in academia and credit is diluted as the number of PIs on the team increases. Success is the coin of the realm in industry. Integrative biology teams, by necessity, must be bigger than any academic laboratory. The capacity for size is yet another reason the modern pharmaceutical firm is ideally positioned to undertake this essential work.

Importantly, there are innovative small companies in and around the pharmaceutical industry whose business models are centered on M&S. Small businesses in this category can be excellent partners for firms adopting the integrative biology approach. Medium-sized M&S-centric companies will probably choose to remain independent and could also opt for an integrative biology team approach, but they would have to elevate their experimental work from outsourced contracts to line management. Neither experiment nor modeling, alone, is enough. These teams can only thrive on trust and mutual respect. To achieve this, they need principal investigator/lab chief-level expertise at all positions on the team. Academia cannot,

and arguably should not, be reorganized in this way; we need to sustain individual genius. But pharmaceutical firms have the managerial covenant, the experimental skill, the modeling expertise, and the organizational size to discover what team genius and shared leadership can accomplish.

7 How Integrative Biology Teams Could Work

How would integrative biology teams function? What would be the value added for the pharmaceutical development enterprise? One view of this process is diagrammed in Fig. 31.1.

In the lower left of the diagram there are two sources of experimental data: proprietary data collected by the company and public data from the open scientific literature. The first task of an integrative biology team is to identify those

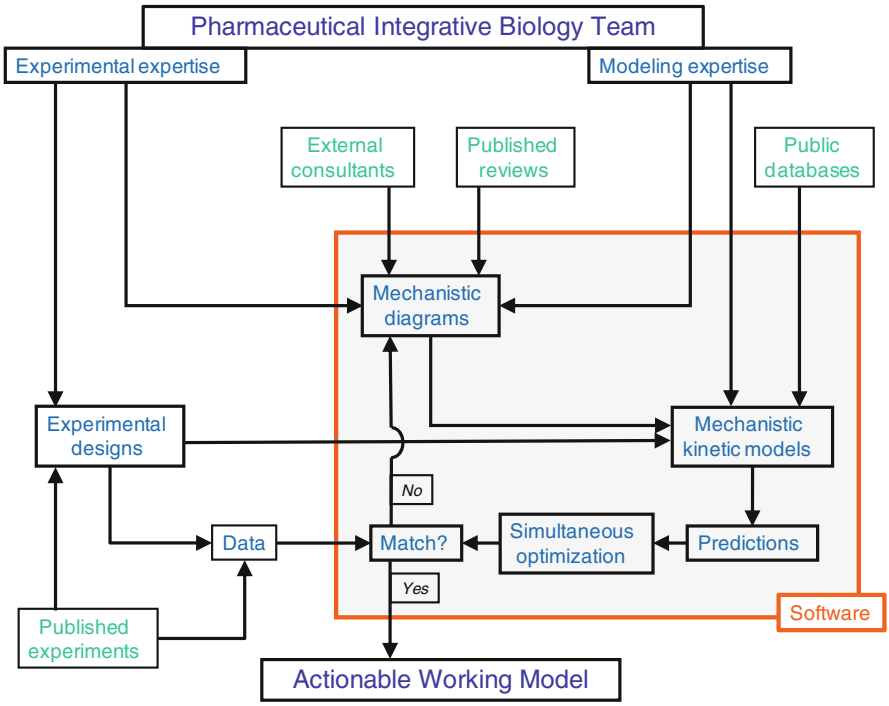


Fig. 31.1 Integrative biology team workflow. An approach, detailed in the text, by which pharmaceutical integrative biology teams can combine experimental and modeling expertise to leverage systems biology productively. This method allows input from all internal stakeholders plus all information in the public domain. External consultants are more effectively employed. This workflow succeeds by providing an objective synthesis of these inputs that can be rigorously tested and then used to guide development

experiments that define the system for which an actionable working model is sought. For candidate drugs and candidate drug targets, the data will likely be proprietary, but much of the current paradigm in any research field will depend on the published literature. Extracting the key defining experiments from that literature is an important task that requires PI-level expertise. Outside consultancy may have an advisory role here, but the in-house integrative biology team must shoulder the final responsibility. Doing this well on an ongoing basis will require team members to stay current with the literature in their fields. Advanced text mining, database, and search technology could contribute greatly to each team's success.

The second task facing an integrative biology team is constructing a mechanistic diagram that the team agrees might account for all the key experimental findings. For this purpose, the team may wish to combine its own views with those expressed in current review articles. The process of diagram building can provide a productive means of leveraging investments in outside consultants, because the team will be seeking a consultant's mechanistic insight in order to test that insight against the accumulated data. The choice of working models thus becomes more objective and less subjective.

With defining experiments and mechanistic diagram in hand, the third task is testing the diagram against the defining experiments. At this stage, software would play a central role. Mechanistic diagrams can be automatically converted to corresponding mechanistic kinetic models. Team members with modeling expertise can supply rate laws for processes where default mass-action rate laws are deemed insufficient. Relevant processes, or even entire sub-models, may be imported from public databases such as BioModels.net [16] or the CellML repository [17] because innovative software teams have seen and met the need for standardized model exchange languages at both biochemical [18] and cell physiological [19] levels of biological organization. Next, experimental designs or protocols from the defining experiments would be applied to each mechanistic kinetic model. Numerical solution of the model's differential equations would yield predictions for each of the key experiments. Parameters would be optimized simultaneously over all experiments and the resulting optimized solutions would be compared directly to the experimental data. Initially there will be significant failures; model predictions will not match the data especially if the team has chosen a large, comprehensive assembly of key experiments for testing. Mismatch will usually signify some significant failure of the proposed mechanistic diagram, although it is always possible that simultaneous testing will uncover fundamental data inconsistencies. Such inconsistencies, viewed by the experimental members of the team, will immediately suggest new experiments to fill gaps in mechanistic understanding. Alternative diagrams can be systematically tested and any diagram can be revised by the team and re-tested against the defining experiments. One extremely valuable source of new mechanisms that might account for uncovered discrepancies is unbiased, high-throughput "-omics" studies of the system of interest. This synergy between the two halves of systems biology, mechanistic and statistical, has been incompletely leveraged to date. Formulation of new diagrams is an exercise in scientific imagination and is always improved by polling as many stakeholders as possible.

This process is repeated until a diagram is found that adequately accounts for all of the defining experimental data. The successful diagram and its corresponding kinetic model become the team's actionable working model.

Value added by integrative biology teams would take many forms only some of which can be listed here. First, confidence in the team's working model is dramatically improved because it has been explicitly and quantitatively tested against the available data. Such a model is evidence-based instead of opinion-based. Testing against all the data simultaneously protects the pharmaceutical development program from unrecognized implications of datasets that might be glossed over in the standard qualitative assessment. The systems biology approach will not "forget" one dataset while formulating a mechanistic explanation for another dataset. Second, an actionable working model can check new experimental designs, before experimental resources are committed, to ensure that they are capable of answering the questions posed. This has enormous potential to save both time and money. Third, such models establish a new kind of corporate memory that insulates the enterprise against inevitable changes in personnel and provides a tool for bringing new team members up to speed quickly.

It might be argued that such tests are suspect because the defining experimental results are known by the team when the mechanistic diagram is formulated, but decades of experience have convinced our field that discovery of a single model that accounts for any extensive collection of different experiments is a major challenge to any team. There are so many constraints requiring simultaneous satisfaction and only so many molecular, cellular, and physiological mechanisms available to work with that success will depend on the very best efforts of all members of the integrative biology team. Successful diagrams and their corresponding kinetic models represent valuable intellectual property and would provide their owners with an enviable competitive advantage.

Others might argue that this approach ignores the principle of model validation. Advocates of validation generally withhold some dataset(s) from the original panel of experiments used to develop the model. Then, they argue, if the model subsequently accounts for the validation dataset, it can be promoted to a new and more august stature known as "validated." Another interpretation, though, is that the withheld experiments simply contain no new information.

Models, theories, and hypotheses are, on the view expressed in this article, never validated, only corroborated. We test models against experimental data and either reject or corroborate them. Thus, the modeling process is simply a quantitative version of classical Popperian hypothesis testing [20]. Whether or not one agrees with Popper's logic of scientific discovery, it must be granted that a "validated" model could be invalidated by next week's experiment. For all these reasons "validation" reduces rather than improves a modeler's credibility among his or her experimental colleagues.

In other words, an actionable working model is never final or validated in any absolute sense, but it has unique value because it has actually passed the stiffest tests the integrative biology team can devise based on a combination of proprietary and public data. Because its details were assembled within the company, they are

known and can be questioned and checked by any employee with the need to know. A proprietary actionable working model not only integrates a company's experimental and systems biology teams but also serves as a platform for testing new biological ideas and new mechanisms of action against the totality of what the company sees as the essential primary data. Actionable working models thus become tangible and extraordinarily valuable intellectual property.

8 Conclusion

Declining productivity in the worldwide biomedical research and development enterprise is a consequence of the enormous complexity of the unsolved human diseases. This complexity will not yield to experimental biology and pharmaceutical development unless we build co-equal teams of expert experimental and expert systems biologists. Academic institutions will always be major sources of new insights and data because they are ideally structured to promote individual genius. But specialization and competition – both essential features of this enormously successful publically funded research engine – have left our universities and research institutes structurally ill-positioned to establish and nurture high-level teams for integrative biology. A potent combination of size, teamwork, expertise, resources, and profit motive strongly suggests that the world-wide pharmaceutical industry is where 21st century integrative biology should be done.

Success will require visionary corporate leadership and skillful management. It will require experimental biologists and physicians who are willing to give up a portion of their historical responsibility for success and failure. It will require M&S leadership that is able and willing to share both the responsibility for scientific decision making and the risks of failure in a for-profit environment where successful patient outcomes are, ultimately, the bottom line. It will not be easy, but, in my view, it is the only way forward.

References

1. Beard DA, Kushmerick MJ (2009) Strong inference for systems biology. PLoS Comput Biol 5(8):e1000459. doi:10.1371/journal.pcbi.1000459
2. Phair RD, Misteli T (2001) Kinetic modelling approaches to in vivo imaging. Nat Rev Mol Cell Biol 2(12):898–907
3. Teusink B, Westerhoff HV, Bruggeman FJ (2010) Comparative systems biology: from bacteria to man. Wiley Interdiscip Rev Syst Biol Med 2(5):518–532. doi:10.1002/wsbm.74
4. Lelandais G, Devaux F (2010) Comparative functional genomics of stress responses in yeasts. OMICS 14(5):501–515
5. Chuang H-Y, Hofree M, Ideker T (2010) A decade of systems biology. Annu Rev Cell Dev Biol 26:721–744
6. Greene CS, Troyanskaya OG (2010) Integrative systems biology for data-driven knowledge discovery. Semin Nephrol 30(5):443–454

7. Kohl P, Crampin EJ, Quinn TA, Noble D (2010) Systems biology: an approach. Clin Pharmacol Ther 88(1):25–33
8. Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
9. Auffray C, Imbeaud S, Roux-Rouquié M, Hood L (2003) From functional genomics to systems biology: concepts and practices. C R Biol 326(10–11):879–892
10. Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD, Mendes P (2010) Further developments towards a genome-scale metabolic model of yeast. BMC Syst Biol 4(1):145. doi:10.1186/1752–0509–4–145
11. Snow CP (1959) Two cultures. Science 130(3373):419
12. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H (2009) The systems biology graphical notation. Nat Biotechnol 27(8):735–741
13. Freeman TC, Raza S, Theocharidis A, Ghazal P (2010) The mEPN scheme: an intuitive and flexible graphical system for rendering biological pathways. BMC Syst Biol 4:65–65. doi:10.1186/1752–0509–4–65
14. Kohn KW, Aladjem MI, Weinstein JN, Pommier Y (2006) Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. Mol Biol Cell 17(1):1–13. doi:10.1091/mbc.E05–09–0824
15. Wimalaratne SM, Halstead MDB, Lloyd CM, Cooling MT, Crampin EJ, Nielsen PF (2009) A method for visualizing CellML models. Bioinformatics 25(22):3012–3019
16. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34(Database issue):D689–D691
17. Lloyd CM, Lawson JR, Hunter PJ, Nielsen PF (2008) The CellML model repository. Bioinformatics 24(18):2122–2123
18. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, and Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531
19. Lloyd CM, Halstead MDB, Nielsen PF (2004) CellML: its future, present and past. Prog Biophys Mol Biol 85(2–3):433–450
20. Popper KR (1965) The logic of scientific discovery, vol 479. Harper & Roy, New York

Chapter 32
Multiscale Mechanistic Modeling in Pharmaceutical Research and Development

Lars Kuepfer, Jörg Lippert, and Thomas Eissing

Abstract Discontinuation of drug development projects due to lack of efficacy or adverse events is one of the main cost drivers in pharmaceutical research and development (R&D). Investments have to be written-off and contribute to the total costs of a successful drug candidate receiving marketing authorization and allowing return on invest. A vital risk for pharmaceutical innovator companies is late stage clinical failure since costs for individual clinical trials may exceed the one billion Euro threshold. To guide investment decisions and to safeguard maximum medical benefit and safety for patients recruited in clinical trials, it is therefore essential to understand the clinical consequences of all information and data generated. The complexity of the physiological and pathophysiological processes and the sheer amount of information available overcharge the mental capacity of any human being and prevent a prediction of the success in clinical development. A rigorous integration of knowledge, assumption, and experimental data into computational models promises a significant improvement of the rationalization of decision making in pharmaceutical industry. We here give an overview of the current status of modeling and simulation in pharmaceutical R&D and outline the perspectives of more recent developments in mechanistic modeling. Specific modeling approaches for different biological scales ranging from intracellular processes to whole organism physiology are introduced and an example for integrative multiscale modeling of therapeutic efficiency in clinical oncology trials is showcased.

L. Kuepfer (✉) • J. Lippert • T. Eissing
Systems Biology and Computational Solutions, Bayer Technology Services GmbH, Building 9115, 51368 Leverkusen, Germany
e-mail: lars.kuepfer@bayer.com; joerg.lippert@bayer.com; thomas.eissing@bayer.com

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_32,

1 An Introduction to Drug Development

Development of novel drugs is a laborious, longsome, and risky process. This is even though biology has seen the advent of many new high-throughput techniques in recent years. Even worse, despite the fundamentally new insights in the regulatory processes underlying biology at different scales, both cost and expenditure of time to market for new drugs have constantly been increasing [17, 28]. The average time to develop a new drug is currently more than 10 years involving costs in the order of 1 billion US dollars, which is to a large extent spend in the late clinical phases [7]. Hence, the translation of mechanistic insights in fundamental and preclinical research towards clinical applications remains challenging [14].

When considering current drug development, the tardily proof of therapeutic efficiency in late phases of clinical development clearly is the largest drawback. This is because in case of failure, large amounts of money allocated for development of a novel drug candidate have already been spent, while in turn expected revenues after approval by marketing and sales are inevitably lost. Hence, any kind of precocious indication of later failure or even withdrawal after market launch will be very beneficial due to a large leverage effect, the earlier the better. Failure in proving therapeutic efficiency, however, might not be the only reason for withdrawal of a novel drug candidate, since occurrence of adverse effects may be an even more serious incident.

With regard to above challenges, novel approaches to generate a mechanistic understanding of drug action and toxicity are important and the wealth of experimental data nowadays available in molecular and cellular biology is not sufficient on its own. The still largely isolated representation of the various layers of biological organization such as the transcriptome, proteome, or kinome [16, 24, 25] reflects the rather reductionist approach resulting from historical development of experimental techniques. Not unexpectedly, this is where systems biology comes in. While contextualization of experimental data in a comprehensive, interpretive framework is the avowed goal of this highly interdisciplinary field, integration across multiple physiological scales still remains an ambitious long-term objective. Moreover, despite an increasing number of targeted studies combining experiment and computational models, systems biology has hardly left fundamental research. Urging questions for the upcoming role and contribution in drug research and development (R&D) hence remain largely unanswered as of now. This becomes even more important as both therapeutic success and occurrence of unwanted side-effects frequently show a large inter-individual variability, which can only be resolved based on a profound mechanistic understanding of the governing processes. This in-depth understanding is also a prerequisite for personalized medicine in order to overcome the current “one size fits all” paradigm in therapeutic design.

To put the various modeling approaches in a wider context, we will briefly review the usual workflow in the development of novel drug candidates in the following (Fig. 32.1). Target identification and compound screening are the prerequisites for the overall development process to start. This still is, by and large, a trial

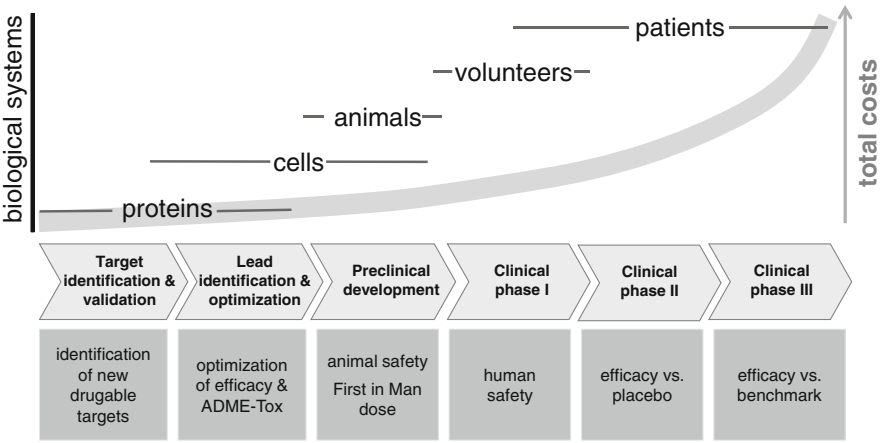


Fig. 32.1 Phases, costs, and biological systems used in drug development

and error process, which, however, can be supported by chem- or bioinformatics database mining. Once a lead compound and a corresponding target have been identified, in vitro assays are performed to optimize affinity and analyze causal dose–response relationships at a cellular scale. During preclinical development, the substance is then tested in different animal models to provide information on the distribution and action including dose–response relations as well as toxicity in order to estimate a therapeutic window. Toxicity assessments are in fact most important to design a first in man dose suggestion. Nevertheless, translation of preclinical results into a clinical settings remains a major challenge and failure of a drug candidate may still occur at each of the three phases of clinical research, which are (1) assessment of general safety in healthy volunteers, (2) evaluation of efficacy in patients, and (3) randomized trials in large groups of patients to achieve a high level of confidence that the drug is safe and efficient. As outlined above, drop out of novel drug candidates in these late stages represents the worst case scenario where large amounts of money are inevitably lost. For these reasons, any kind of comprehensive evaluation of the to be expected therapeutic potential of novel drug candidates supports the establishment of a more favorable risk-benefit profile and bears, moreover, a huge economic potential (Fig. 32.1).

2 Computational Systems Biology in Pharma R&D

Computational models are an effective way to integrate knowledge, information, and assumptions with experimental data in one unified representation. This is in particular important in complex systems, such as animal models or human patients, because experimental data generated on different levels of biological organization

may easily become too complex for non-formal, intuition-based analyses. Here, computational models represent a comfortable way for data processing, integration, and subsequent analyses and they allow the inclusion of experimental data into a (mathematically) rigorous framework. Ideally, computational models may represent the accumulated expert knowledge generated over years of research. This allows simulation of system behavior in the face of perturbations and the generation of testable hypotheses for further experimental planning. However, so far the role of computational models in drug development is restricted to mostly pharmacokinetic (PK) and pharmacodynamic modeling (see below), thus accompanying late stage (pre-)clinical research. This again is in sharp contrast to the wealth of highly specific modeling approaches developed in computational systems biology over recent years, which largely remain unused in the context of pharmaceutical development. Moreover, the main focus of model-based investigations is on the retrospective analysis of experimental data rather than on future design of preclinical and clinical studies such that existing possibilities remain unused. One reason is that the acceptance of modeling and the belief in the predictive power is still limited.

In principle, computational simulations in drug development can provide mechanistic insights into both PK, i.e., the processes governing the absorption, distribution, metabolization, and elimination (ADME) behavior of substances within the body [13, 26, 34, 46] and pharmacodynamics (PD), i.e., the modes of action at the specific target site and its physiological effect on the organism [5, 6, 38]. Model-based analyses may significantly contribute to the generation of a truly mechanistic understanding of drug action such that fundamental analyses may be performed in silico in a prospective rather than in a retrospective manner. Ideally, computational models can be used for an exhaustive integration of experimental data into mathematical frameworks at all stages of drug development. Thereby, data can be processed, represented, and analyzed within the context of the current systems understanding. Likewise, this understanding can in turn constantly be put to the test by comparing simulation results with new experimental findings. Computational models may hence accompany all phases of drug development by structural knowledge management ranging from early data mining and target identification to planning of clinical trials in phases one to three (Fig. 32.1).

Since mechanistic computational models include a high level of prior information, they are particularly well-suited for analyses of the underlying network structure. Moreover, the comprehensive representation of existing knowledge allows an efficient extrapolation to new scenarios. This is a valuable tool at various stages of drug development, especially when it comes to crossing physiological scales and areas of application, for example, when special sub-populations need to be investigated or new indications are to be identified [10, 44]. In this regard, in vitro to in vivo extrapolation is a typical example, where the mechanistic information included in the basic model structure may help to transfer the predictive capabilities of wet-lab assays to animal studies and further to trials in humans. Mechanistic PK/PD models can be used to summarize the experimental information generated for one animal species and extrapolate existing knowledge to other species, thereby reducing significantly the number of animal sacrifices. In clinical phases,

computational models may be valuable tools to allow a mechanistic consideration of the therapeutic window in order to maximize the efficiency and safety of a drug. In this respect, computational models may ideally be used for the establishment of a mechanistic understanding of occurrence of potentially severe adverse effects, which thus can be avoided in a prospective way. Mechanistic computational models may also be used in between clinical studies to extrapolate to new patient subgroups [10] including pediatric scaling [9].

3 Multiscale Modeling in Pharma R&D

Above examples describe the current role of computational models in pharma R&D also outlining future fields of application. In the face of the various preclinical and clinical phases involved, however, it becomes clear that computational modeling in support of development of novel drug candidates is inevitably a multiscale problem. This is because (1) cellular, (2) tissue and organ, (3) whole-body, and (4) population scale, respectively, ideally need to be addressed in one integrative modeling framework (Fig. 32.2). Given the fact that an exhaustive mechanistic representation of human physiology, however, requires to cover a factor of 10^9 in a spatial scale (size of a molecule wrt the human body) and 10^{15} in a time scale (Brownian motion wrt human lifespan) [19], it is obvious that this endeavor is largely out of reach of current modeling possibilities. Moreover, even though such an aggregate model might certainly be beneficial for an unforeseeable number of future applications in clinical medicine, it is highly questionable if it is actually needed or helpful to support current drug development. This is because any computational model

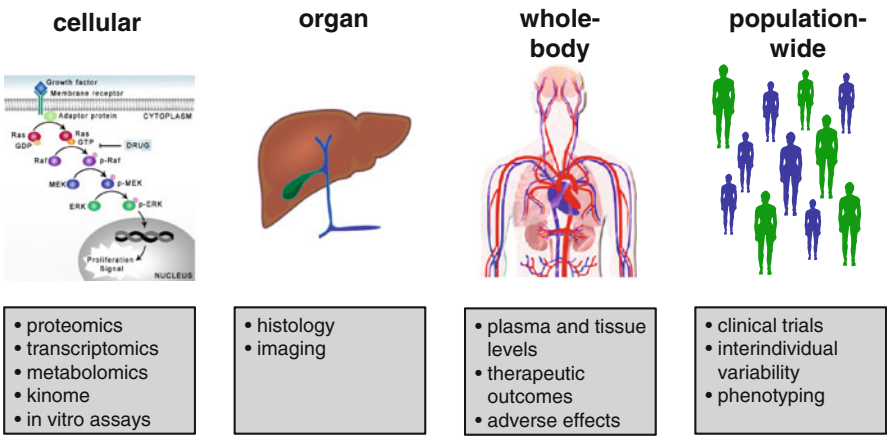


Fig. 32.2 Scales of biological organization that are important for drug development. Experimental techniques and methods needed for model establishment and subsequent validation are indicated in *grey boxes*

generally should focus on the specific questions to be investigated instead of aiming for general-purpose applicability. With regard to drug development it is, however, mandatory to always take the molecular level into account, either in terms of drug ADME or PD, such that the fundamental cause–response relationships governing a therapeutic outcome can be mechanistically described and ideally investigated on a mechanistic level.

We therefore have to arrive at a reasonable level of abstraction and simplification in order to consider different relevant scales within one model. As of now, however, even reduced multiscale models are currently far from being standard in drug development, which is both due to heterogeneity of experimental data and limitations in adequate modeling approaches, which have just started to be developed [12]. Computational models rather follow reductionist approaches thus representing the different levels of biological organization such as molecular or cellular biology. In the face of the different questions to be addressed at these scales, the models frequently use very specific mathematical formalisms for the structural representation. In the following, we will discuss the different approaches with respect to the various scales for which they were developed before outlining possible applications for pharma R&D in the future.

3.1 Cellular Scale

3.1.1 Stoichiometric Models at Cellular Scale

Stoichiometric models of metabolism represent the fundamental inventory of the cell for maintenance and fueling (Fig. 32.3a). Metabolic network models are constructed based on genome annotation and summarize biochemical knowledge within the stoichiometric matrix. They inherently assume steady state of the intracellular components at the expense of abandoning any kinetic information. The overall model structure is generally linear and represents an underdetermined system of algebraic equations in which intracellular fluxes are the unknown variables. Such flux distributions can be seen as functional endpoints of upstream regulation and ultimately specify cellular modes of operation. Metabolic network models have been used previously for the investigation of metabolic principles in microbial cells [20, 35], but recent works have started to consider genome-scale models of human metabolism in a generic [8] or tissue-specific way [15].

A special advantage of stoichiometric models lies in the wealth of analytical algorithms that have been developed in recent years. Human metabolic models have, for example, been used for the analysis of gene expression data thus revealing disease-related enzymes in different organs [36] or for the identification of biomarkers of metabolic diseases [37]. Stoichiometric models provide an analytical framework for the contextualization of experimental data and structural analysis of network disorders. Limitations of metabolic network models are the negligence of regulatory constraints, possible non-identifiability of flux solutions and the

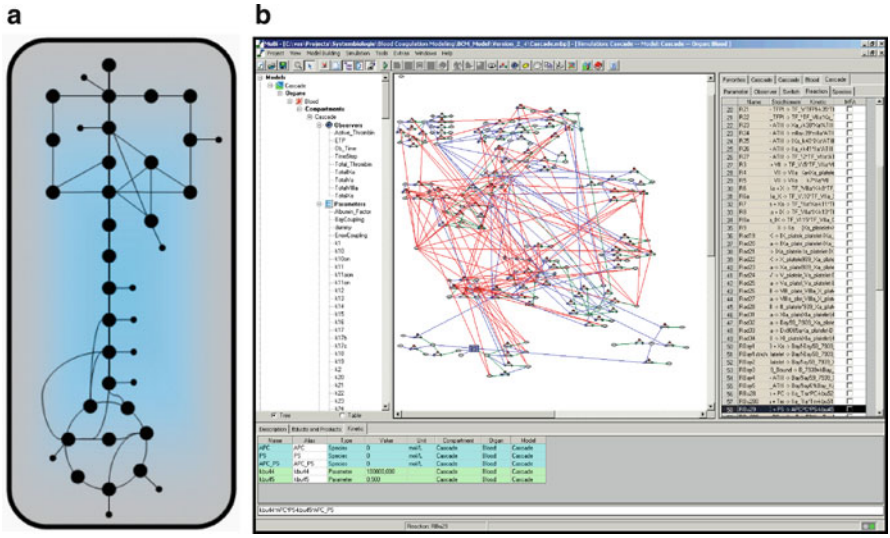


Fig. 32.3 Computational models at cellular scale: a schema of a stochiometric model for the central carbon metabolism (**a**) and a dynamic model of a signaling cascade network (**b**) implemented in MoBi® are exemplarily shown

fundamental steady state assumption, which severely hampers embedding of such models into dynamic models considering physiological processes on a higher scale. Since these network models, however, can represent cellular metabolism at genome scale, they provide in turn a unique possibility to correlate genetic predisposition with clinical observations making them a valuable tool for model-based analysis of genotype–phenotype correlations [22].

3.1.2 Dynamic Models at Cellular Scale

Most of the dynamic models at cellular scale represent metabolic pathways or intracellular signaling (Fig. 32.3b) in the face of specific triggers or perturbations thus addressing fundamental biochemistry of the cell [3, 11, 23]. In case of intracellular signaling cascades, the models describe protein–protein interactions which result in modifications such as protein phosphorylation status or protein complex formation. Such models thus describe intracellular information trafficking in the face of extracellular perturbations and thereby directly complement cell biological in vitro assays. In the context of pharma R&D, dynamic models could generally complement PK/PD models and simulations as they follow similar mathematical formalisms such that an incorporation of submodels is technically feasible given an adequate modeling framework [12]. Dynamic models are moreover a valuable tool for mechanistic investigation at an in-depth level of detail since they enable

simulation of intracellular systems at rapid time scales, which in addition frequently display non-linear behavior. Mechanistic models are particularly important for system identification since they enable simultaneous consideration of multiple sets of experimental data. Adequate algorithms for both structural and parameter identification, respectively, are in turn mandatory.

Dynamic models in drug development can be used to explain cause–response relationships in cellular networks and to quantify system dynamics, respectively [11]. They can furthermore be used to investigate cellular responses in the face of extracellular stimuli such as binding of exogenous and endogenous ligands [3,23]. They have direct applications in pharmacodynamics, where they can be used to investigate specific modes of drug action at cellular scale and dose–response relationships intertwined in the basic structure of the signaling network. Since dynamic models at cellular scale can be used to process and analyze a large level of various kinds of in vitro data, they may significantly support early preclinical research. In this regard, incorporation of the generated data into an integrative modeling framework can help to establish a mechanistic assessment of drug action both on and off-target-wise, and, equally important, detect gaps in the current systems understanding.

While dynamic models in computational systems biology are mostly based on ordinary differential equations (ODEs), some cellular events can only be explained with different modeling approaches. This may, for example, become important if some proteins of a signaling cascade are firmly integrated into a structure of the cell while others may move freely in a particular compartment [21]. To explain resulting changes in signaling dynamics, protein–protein interactions can be described by spatio-temporal models which additionally take into account spatial coordinates. Likewise, molecular diffusion and migration may need to be described by reaction–diffusion equations also requiring partial differential equations. However, parameterization of spatio-temporal models is often difficult, time-consuming, and computationally demanding. Further, many relevant phenomena for whose description such approaches are used might also be modeled by simpler model formulations, e.g., compartmental approximations of spatial effects.

Both ordinary and partial differential equations are deterministic ignoring stochastic effects being at the very basis of molecular interactions [30]. Especially if only a very low number of a certain molecule is strongly influencing a process, intrinsic noise arises. For example, gene expression is variable and consequently even neighboring cells of the same type will have different molecule numbers for most proteins expressed. Different modeling techniques have evolved to exactly or approximately describe stochastic processes and ignoring stochasticity will lead to simplifications, sometimes making it impossible to understand basic processes. It may even be necessary to consider both stochastic and partial effects as, for example, diffusion itself is also stochastic. However, stochastic simulations are computationally demanding, adding random-walk aspects increases this demand further. Also, for many investigations, average models are a good approximation and certain sources of noise can alternatively be reflected in a deterministic framework by executing population simulations that use parameter sets derived

from distributions describing the observed variability. For example, the intrinsic noise leading to a variable gene expressing might not be considered explicitly, but the consequence of variable amounts of proteins can be reflected.

3.2 Tissue and Organ Scale

While dynamic models at the cellular scale represent a structural description of fundamental biochemistry, computational models at the tissue and organ scale largely focus on physical conservation laws such as mass, momentum, and charge in order to describe organ morphology and function [19]. Computational models at that scale, therefore, directly complement experimental data such as imaging techniques and histology. To account for a particular spatial shape of an organ which ultimately contributes to a specific function, computational organ models are frequently based on partial differential equations. Thus, movement of different fluids within the human body can be described in the context of the surrounding tissue and its time resolved motility. While a large level of organ systems have been investigated with such so-called continuum methods, the virtual heart model is by far the most advanced representation of a human organ taking into account both cellular scale and macroscopic anatomy [27]. The heart model has been used to generate a mechanistic understanding of the various processes underlying myocardial contraction ranging from electrophysiology to laws of mechanics. It was, for example, used successfully to analyze drug-induced occurrence of cardiac arrhythmias which has important implication for in silico testing of toxicity [31].

Further examples for models of organ systems at a macroscopic scale are models of kidney [40] or lung [39]. These models are mainly driven by mechanistic consideration of specific organ functions such as urinary secretion which ultimately drive the structure of the underlying modeling framework. The models are valuable tools for the investigation of the physiological functionality in healthy or even diseased individuals. Their applicability in the context of drug development, however, is rather limited since these models often do not consider the cellular or molecular scale, such that drug action and distribution, respectively, cannot be represented at a mechanistic level of detail. Many computational organ models are moreover based on finite-element methods, which again hampers their linking to models at others scales.

An interesting application of organ models at a macroscopic scale offer methods which consider cells at discrete entities. Such models describe the actions and interactions of single cells based on governing equations which are assumed to represent fundamental global rationales such as cellular growth or differentiation. Such discrete models have, for example, been used to describe proliferation of tumor cells and their subsequent migration within tissue [1]. Following a similar approach, agent-based models were successfully used to describe regeneration of liver tissue in mice after toxin-induced injury [18]. Again, the integration of ODE-based models in such discrete models is not trivial.

The number of computational models at organ scale has constantly been increasing in recent years. Macroscopic models, however, demand for elaborate computational approaches such as spatio-temporal modeling or agent-based methods. While such detailed descriptions of specific processes are inevitably necessary for mechanistic investigations of particular physiological functions, linking of advanced organ models to models at cellular scale or to support PK/PD simulations is currently still challenging. To nevertheless consider models at organ scale for the description of drug action and distribution, it is hence necessary to rather focus on a reduced, purpose-driven representation. A simple solution is the consideration of spatial effects in a compartmental model where each component is assigned to a particular location. This approach, which allows the consideration of PDE models as much simpler ODE models will be discussed in the next paragraph.

3.3 Whole-body Scale PK and PD Models

While the site of action of a pharmacological substance might be restricted to certain tissues or cells, first of all a quantitative estimation of the administered substance available at the site of action is required. These questions underlie the subject of PK and different modeling techniques are well-established in pharmaceutical research to support their investigation. So far, the most widely used approach is to establish descriptive and comparably simple compartmental PK models that can be well-identified based on available data (Fig. 32.4a). Such PK models can be extended to include compartments or a descriptive relation to effects (PD), for example, in the form of a simple hyperbolic concentration–effect relation. In contrast to the rather phenomenological consideration of drug PK of compartmental models, physiology-based pharmacokinetic (PBPK) models aim for a detailed representation of physiological processes (Fig. 32.4b). In the next two sections both approaches will be introduced in more detail.

3.3.1 Compartmental PK Models

Compartmental PK models are kinetic models to describe the concentration–time curve of a substance [2, 32, 41]. The simplest form is a one-compartment model for an intravenous bolus corresponding to a linear first order differential equation, which considers that a given amount of substance is homogenously distributed in an unknown volume from where it is eliminated via a first order process. Concentration–time measurements from blood plasma taken in a clinical study are then used to identify the elimination rate constant and the volume. The volume to be identified can indicate the distribution between blood plasma and the rest of the body. Whereas a very large substance that cannot cross the endothelial barrier of the blood vessels will have a small volume of distribution corresponding to the blood

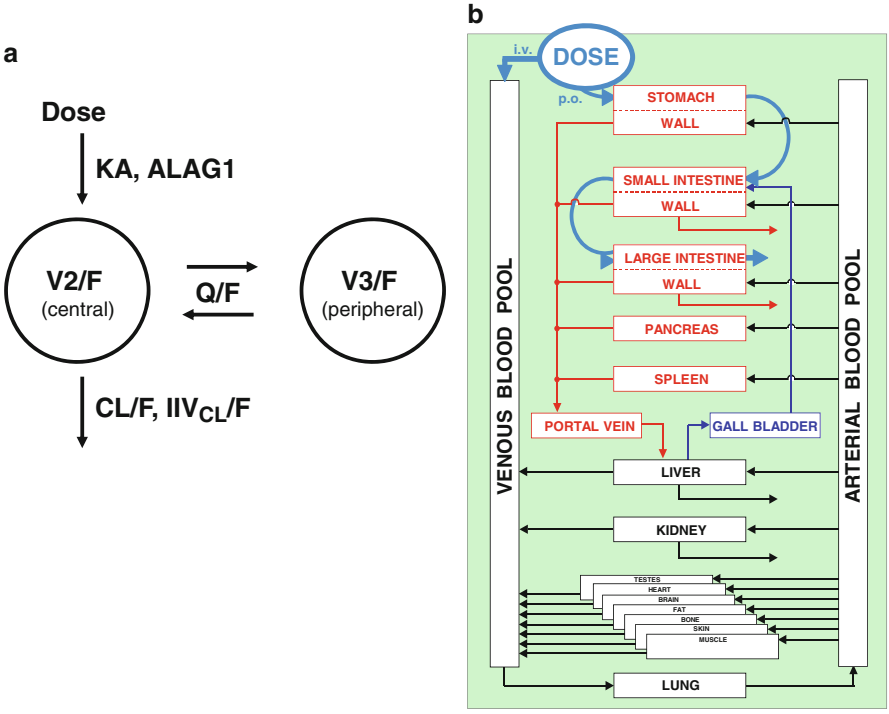


Fig. 32.4 Compartmental PK (**a**) versus PBPK model (**b**). The compartmental non-linear mixed effect model scheme (**a**) outlines a depot compartment from where the drug is absorbed into the central compartment with the absorption rate constant KA and the lag time of absorption ALAG1. The drug is eliminated from the central compartment with the apparent clearance CL/F (subjected to inter-individual variability, IIV) and distributes into the peripheral compartment with the apparent inter-compartmental clearance Q/F. The apparent volumes of distribution of the central and peripheral compartments are V2/F and V3/F, respectively. The PBPK model scheme (**b**) outlines organs/compartments relevant for oral absorption (*red*) and systemic distribution (*black*). *Arrows* indicate mass-transport via organ-specific blood flow rates [47]

volume, a small lipophilic substance will quickly distribute into most parts of the body and accumulate in fat giving rise to a large volume of distribution that can become larger than the physiological volume of the whole organisms [32].

More elaborate compartmental models consider more than one compartment, and the different compartments are generally interconnected by exchange rates (see Fig. 32.4a). Structural model identification is guided by identifiability measures and quality of fit. Besides the model parameters themselves that already provide a first characterization of the substance, the model can be further used to determine descriptive concentration–time curve characteristics such as the maximal concentration, area under the plasma concentration–time curve, or the half-life time. This holds true even if the experimental data are sparse making a direct assessment difficult.

Because compartmental models are often rather simple, sophisticated identification techniques can be efficiently applied. This allows to not only identify typical parameters for an average individual but also to quantify variability between subjects in a population or between occasions in order to investigate sources of variation, and provide individual estimates for the different subjects in a populations using non-linear mixed effect models. Such a model-based data analysis can also be used to evaluate the statistical significance of a certain influence or status (covariate) which is to be included in the model in a hierarchical manner. For example, an analysis can show how much of the inter-individual variability can be explained by a covariate such as body-weight or disease status.

Consequently, non-linear mixed effect models are a powerful tool in order to analyze population data as obtained in clinical studies. However, such models are generally drug or even dataset-specific and therefore do not provide the ideal platform for knowledge integration. Furthermore, since the model parameters are empiric and generally have no true physical or physiological correspondence, it is difficult to translate these models to new situations thereby limiting the predictive power. Certain concepts including allometric scaling of parameters have been adapted over time, but these have their limits as they only consider selected aspects [33,43]. Also, there are approaches to extend compartmental PK models to consider important aspects of physiology. While these can provide some mechanistic insight into the PK, the approach often remains phenomenological. In order to enable truly comprehensive analyses of the processes governing the distribution and subsequent metabolization and excretion of a substance, a much more elaborate model structure is clearly needed. This approach is used by PBPK models which are presented in the following.

3.3.2 Physiologically-Based Pharmacokinetic Models

PBPK models are a mechanistic approach to describe the ADME behavior of a substance based on substance-specific properties and human physiology, which include a large level of prior biological information for model building [13,26,34,45,47–49].

In order to provide a physiological framework model, the (human) body is divided into containers representing relevant organs or tissues as well as arterial and venous blood pools connecting the different organs through the blood flow (see Fig. 32.4b). Organs are further sub-divided into several sub-compartments considering, for example, the vascular space divided into plasma and (red) blood cells as well as the avascular space divided into interstitial and cellular space. Such a model framework corresponds to a large compartmental model and provides the basis to describe the ADME behavior of a substance, while all free parameters can be identified independent of substance knowledge and PK measurements. In addition, information on compartment composition, e.g., in terms of volume fractions of water, proteins, and lipids can be implemented independent of the substance. In order to further describe active transport processes and enzyme-catalyzed metabolization, the basic model can be extended accordingly.

Based on generic distribution models, only a few basic physico-chemical parameters of a substance such as molecular weight, lipophilicity, and protein binding are necessary to describe the ADME behavior in this framework including permeabilities across membranes and partition coefficients between compartments. To start with, such a model will only consider passive processes, which are primarily the distribution based on blood flow and diffusion as well as absorption for orally administered substances. Active processes including metabolization, transportation, or binding can be added as needed. In this regard, substance knowledge and in vitro data can often guide decision making between structural model alternatives. Especially for antibodies target binding or receptor-mediated clearance or protection from clearance might also be important. With adequate PK data at hand, selected parameters of the model can be fine-tuned to achieve a good fit.

For establishing a PBPK model in humans, physiological knowledge is not restricted to average individuals, but for many parameters their distribution within different populations is known in an age-dependent or patient-specific manner allowing population PK predictions rather than fits only. PBPK models can also be established for different animal species. With a predictive animal PBPK model at hand, for example, the physiological parameters can then be substituted to make a first prediction for humans. The physiological correspondence of parameters allows both a good interpretation of results as well as a translation to new scenarios. Consequently, PBPK models are well-established in the environmental toxicology and risk assessment fields and are becoming increasingly popular also in pharmaceutical research. In addition, PBPK models automatically provide exposure estimates at the side of action and therefore provide a natural basis to build multiscale models PK/PD models as exemplified below and thereby provide a good platform for knowledge integration along the pharmaceutical R&D process [42].

4 An Exemplary Multiscale Model

In the previous chapters, different modeling approaches specific for the various scales of biological organization were presented and discussed. To support pharma R&D, a model-based presentation of expert knowledge and experimental data in integrative multiscale models is clearly desirable. In the following, we will discuss such a modeling concept, which, despite being generic in large parts, clearly outlines our vision of computational models in drug development. The exemplary multiscale model describes a virtual patient with a pancreatic tumor and the treatment by a generic chemotherapeutic agent (Fig. 32.5) [12].

The model (Fig. 32.5a) considers the PK of the parent prodrug and its activation through metabolization by a polymorphic enzyme (cytochrome P450 2D6, CYP2D6) generating the active drug itself. The coupled PBPK model thus provides drug concentration–time information at the site of action, which is the pancreatic

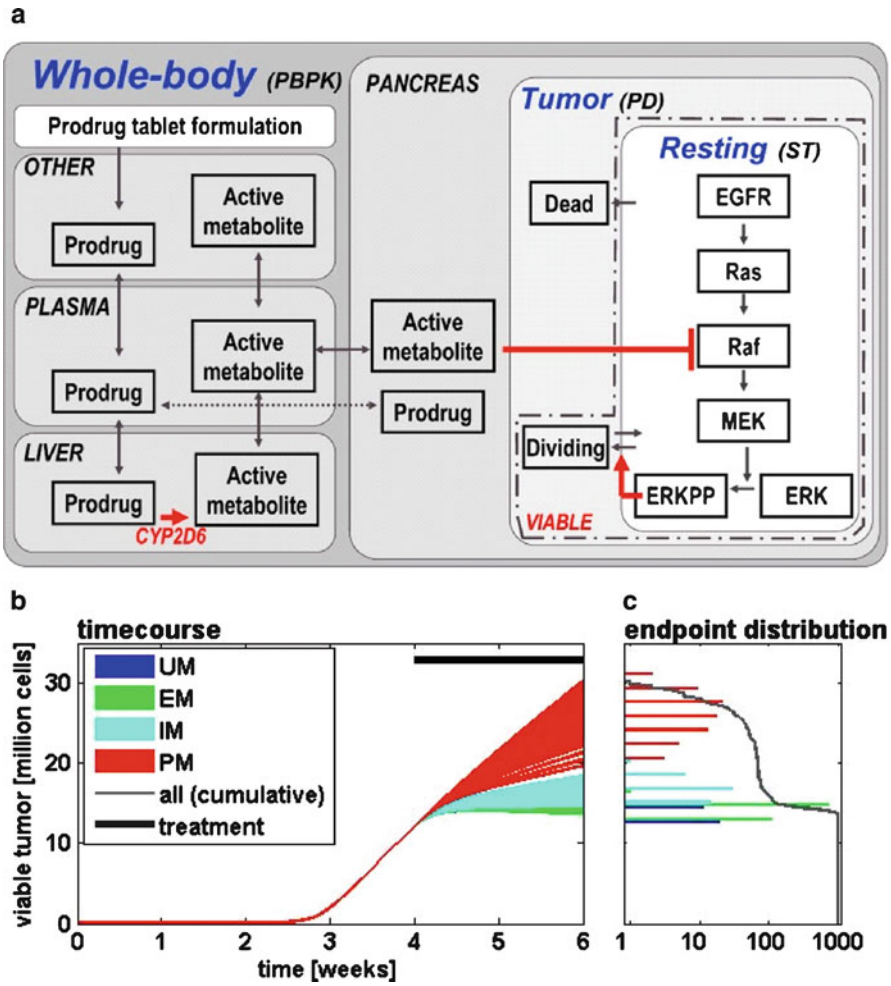


Fig. 32.5 Multiscale model outline (**a**) and simulation results (**b** and **c**) according to [12] as discussed in the text

tumor modeled to in a larger level of detail to include the signal transduction pathway the drug is interfering with trough a competitive binding to the Raf kinase, and a tumor model whose growth is driven by a transcription factor activated in the signal transduction pathway.

In a virtual clinical study, the individual therapeutic outcome of the chemotherapeutic intervention is simulated for a large population with heterogeneous genomic background. Apart from normal physiological variability, e.g., in organ sizes and blood flow rates, also the phenotypic CYP2D6 activity differences are considered as indicated by the different colors in Fig. 32.5b (ultra-rapid, extensive, intermediate, and poor metabolizers abbreviated as UM, EM, IM, and PM, respectively). The

model is simulated without drug treatment for four weeks resulting in an initial exponential growth phase, followed by a linear growth phase during the first four weeks of untreated tumor growth. Thereafter, two weeks of bi-daily treatment (indicated by the black bar on the top) in the form of orally administered tablets are simulated. Simulation results are shown for the viable tumor mass (Fig. 32.5b), which was also chosen as a simple endpoint (Fig. 32.5c). As can be seen, the endpoint distribution shows a significant variability with some virtual patients who are hardly responding to treatment, whereas for other patients the treatment even leads to a weak tumor regression.

More detailed analysis into the sources of variability shows that treatment success is only weakly influence by general physiology, but strongly dependent on the CYP2D6 phenotype, which would likely also have fundamental implications for optimal therapeutic dosing. Comparable findings have been reported for, e.g., the chemotherapeutic prodrug tamoxifen [4, 29]. Further investigations on the influence of GAP-insensitive Ras mutations indicate that this mutation mainly impacts on the linear tumor growth phase in the current setting, potentially reflecting the important role of growth-factor signaling during this phase to further promote cell division although limiting environmental conditions do not support exponential growth any more. The results of such analyses for a validated model can provide a rational basis for the planning of clinical studies and pave a road to personalized medicine.

5 Conclusions and Outlook

Ethical and medical constraints and the high level of investments required for the development of novel drugs demand rational decision making even beyond the typical level in other industries. At the same time, the biological system "patient" as well as the preclinical model systems ranging from individual target proteins through cellular systems to animal models provide a level of complexity unraveled by any technical system. As a consequence, human intuition or classical "management reasoning" alone cannot cope with the wealth of information accumulated in the course of a pharmaceutical R&D project.

Pharmaceutical innovator companies and public R&D organizations developing drugs are therefore looking for new approaches for decision making at all levels of the R&D process, ranging from experimental design and dose selection to rationalization of investment decisions at a project portfolio level. Computational modeling is a candidate technology due to its capability to integrate, process, and represent data and to use computational simulation for prediction purposes.

In biological modeling, model structures directly represent the systems understanding of, e.g., drug ADME properties and drug action, which can be used for hypotheses testing, experimental design, and identification of inconsistencies. In the face of the various types of experimental data generated at the different levels of biological organization along the pharmaceutical R&D process, so far, specific computational models for individual biological systems and levels used in

different project phases have been developed rather independently. This lead to an overemphasized reductionism and prevented a true translational use of models for bridging between project phases and the corresponding model systems used (protein to cell, cell to animal, animal to volunteer, and volunteer to patient). Adequate multiscale approaches, however, are necessary for an in-depth mechanistic description of drug action integrating across biological scales and allowing translation. On the one hand, such approaches clearly need to represent pharmacodynamics at cellular scale to represent basic molecular mechanisms, while on the other hand the mechanisms governing drug ADME need to be taken into account as well. Given the mind-blowing complexity of biology, a pragmatic, purpose-driven representation of physiological function and patho-physiological processes is mandatory nevertheless.

We here present various approaches of computational modeling at the cellular, organ, and whole-body scale. The applicability of the different mathematical formalisms within an integrative multiscale modeling framework is discussed. Compartmental PK/PD models have currently achieved the highest acceptance in pharmaceutical industry and regulatory bodies. The compartmental approach is characterized by a phenomenological representation of biology, however. An integration of models with computational representations of smaller biological scales is therefore usually quite difficult and often impossible. PBPK models offer a much higher level of mechanistic detail and physiological information. As a direct consequence, PBPK models offer a mechanistic framework for the integration of metabolic and signal transduction models, which can often be seen as more elaborate mechanistic PD representations. A small number of dedicated computational models at organ scale are partly very advanced but due to a focus on, e.g., biomechanical properties of the heart in the virtual heart initiative a use of these models for modeling of drug action is not always straight forward.

The relevance of computational models in drug development has already increased for more than a decade and this process is rather accelerating than decelerating. Both pharmaceutical companies and regulatory bodies like the US Food and Drug Administration (FDA) are building up dedicated resources and nowadays computational considerations are already considered as a valuable complement to experimental data. The current investments in dedicated disease models for specific therapeutic areas will further strengthen this position. The rapid development of more and more powerful and user-friendly software platforms for modeling and simulation will also help. Altogether, reality is becoming closer and closer to the vision of an exhaustive representation of knowledge, information, assumption, and data in computational models and the use of simulation for a truly rational decision making. It appears reasonable that modeling and simulation will soon significantly enhance the development of targeted therapeutics with favorable risk-benefit profile thereby markedly providing benefits to industry, patient, and careers.

Acknowledgment This study was supported by the German Ministry for Education and Research (BMBF) through the Systems Biology Networks Virtual Liver and FORSYS partner.

References

1. Anderson AR, WeaverAM, Cummings PT, Quaranta V (2006) Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. Cell 127: 905–915
2. Beal SL, Sheiner V (1994) NONMEM user's guide. NONMEM project group. University of California, San Francisco
3. Becker V, Schilling M, Bachmann J, Baumann U, Raue A, Maiwald T, Timmer J, Klingmuller U (2010) Covering a broad dynamic range: information processing at the erythropoietin receptor. Science 328:1404–1408
4. Briest S, Stearns V (2009) Tamoxifen metabolism and its effect on endocrine treatment of breast cancer. Clin Adv Hematol Oncol 7:185–192
5. Burghaus R, Coboeken K, Gaub T, Kuepfer L, Sensse A, Siegmund HU, Weiss W, Mueck W, Lippert J (2011) Evaluation of the efficacy and safety of rivaroxaban using a computer model for blood coagulation. PLoS One 6:e17626
6. Chassagnole C, Jackson RC, Hussain N, Bashir L, Derow C, Savin J, Fell DA (2006) Using a mammalian cell cycle simulation to interpret differential kinase inhibition in anti-tumour pharmaceutical development. Biosystems 83:91–97
7. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22:151–185
8. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104:1777–1782
9. Edginton AN, Willmann S (2006) Physiology-based versus allometric scaling of clearance in children; an eliminating process based comparison. Paediatr Perinat Drug Ther 7:146–153
10. Edginton AN, Willmann S (2008) Physiology-based simulations of a pathological condition: prediction of pharmacokinetics in patients with liver cirrhosis. Clin Pharmacokinet 47:743–752
11. Eissing T, Conzelmann H, Gilles ED, Allgower F, Bullinger E, Scheurich P (2004) Bistability analyses of a caspase activation model for receptor-induced apoptosis. J Biol Chem 279:36892–36897
12. Eissing T, Kuepfer L, Becker C, Block M, Coboeken K, Gaub T, Goerlitz L, Jaeger J, Loosen R, Ludewig B, Meyer M, Niederalt C, Sevestre M, Siegmund H-U, Solodenko J, Thelen K, Telle U, Weiss W, Wendl T, Willmann S, Lippert J (2011) A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks. Front Physio 2:4
13. Espie P, Tytgat D, Sargentini-Maier ML, Poggesi I, Watelet JB (2009) Physiologically based pharmacokinetics (PBPK). Drug Metab Rev 41:391–407
14. FDA (2004) Innovation or stagnation: challenge and opportunity on the critical path to new medical products. Challenges and Opportunities Report – March 2004
15. Gille C, Bolling C, Hoppe A, Bulik S, Hoffmann S, Hubner K, Karlstadt A, Ganeshan R, Konig M, Rother K, Weidlich M, Behre J, Holzhutter HG (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol 6:411
16. Gstaiger M, Aebersold R (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet 10:617–627
17. Henney AM (2009) Who will take up the gauntlet? Challenges and opportunities for systems biology and drug discovery. EMBO Rep 10(Suppl 1):S9–13
18. Hoehme S, Brulport M, Bauer A, Bedawy E, Schormann W, Hermes M, Puppe V, Gebhardt R, Zellmer S, Schwarz M, Bockamp E, Timmel T, Hengstler JG, Drasdo D (2010) Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. Proc Natl Acad Sci USA 107:10371–10376
19. Hunter PJ, Borg TK (2003) Integration from proteins to organs: the physiome project. Nat Rev Mol Cell Biol 4:237–243

20. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7:198–210
21. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7:165–176
22. Kuepfer L (2010) Towards whole-body systems physiology. Mol Syst Biol 6:409
23. Kuepfer L, Peter M, Sauer U, Stelling J (2007) Ensemble modeling for analysis of cell signaling dynamics. Nat Biotechnol 25:1001–1006
24. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470:187–197
25. Metz JT, Johnson EF, Soni NB, Merta PJ, Kifle L, Hajduk PJ (2011) Navigating the kinome. Nat Chem Biol 7:200–202
26. Nestorov I (2007) Whole-body physiologically based pharmacokinetic models. Exp Opin Drug Metab Toxicol 3:235–249
27. Noble D (2002) Modeling the heart–from genes to cells to the whole organ. Science 295: 1678–1682
28. PricewaterhouseCoopers (2007) Which path will you take? Pharma 2020: The vision
29. Rae JM, Sikora MJ, Henry NL, Li L, Kim S, Oesterreich S, Skaar TC, Nguyen AT, Desta Z, Storniolo AM, Flockhart DA, Hayes DF, Stearns V. (2009) Cytochrome P450 2D6 activity predicts discontinuation of tamoxifen therapy in breast cancer patients. Pharmacogenomics J 9:258–264
30. Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation and tolerance of intracellular noise. Nature 420:231–237
31. Rodriguez B, Burrage K, Gavaghan D, Grau V, Kohl P, Noble D (2010) The systems biology approach to drug development: application to toxicity assessment of cardiac drugs. Clin Pharmacol Ther 88:130–134
32. Rowland M, Tozer TN (2011) Clinical pharmacokinetics and pharmacodynamics – concepts and applications. Lippincott Williams & Williams, Baltimore
33. Savage VM, Deeds EJ, Fontana W (2008) Sizing up allometric scaling theory. PLoS Comput Biol 4:e1000171
34. Schmitt W, Willmann S (2005) Physiology-based pharmacokinetic modeling: ready to be used. Drug Discov Today 2:125–132
35. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol Syst Biol 3:119
36. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26:1003–1010
37. Shlomi T, Cabili MN, Ruppin E (2009) Predicting metabolic biomarkers of human inborn errors of metabolism. Mol Syst Biol 5:263
38. Sung JH, Esch MB, Shuler ML (2010) Integration of in silico and in vitro platforms for pharmacokinetic–pharmacodynamic modeling. Exp Opin Drug Metab Toxicol 6:1063–1081
39. Tawhai MH, Bates JH (2011) Multi-scale lung modeling. J Appl Physiol 110(5):1466–1572
40. Thomas SR (2009) Kidney modeling and systems physiology. Wiley Interdiscip Rev Syst Biol Med 1:172–190
41. Tornoe CW, Agerso H, Jonsson EN, Madsen H, Nielsen HA (2004) Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in NLME using differential equations. Comput Methods Progr Biomed 76:31–40
42. van der Graaf PH, Benson N (2011) Systems pharmacology: bridging systems biology and pharmacokinetics–pharmacodynamics (PKPD) in drug discovery and development. Pharm Res 28(7):1460–1464
43. West GB, Brown JH, Enquist BJ (1997) A general model for the origin of allometric scaling laws in biology. Science 276:122–126
44. Willmann S, Edginton AN, Coboeken K, Ahr G, Lippert J (2009) Risk to the breast-fed neonate from codeine treatment to the mother: a quantitative mechanistic modeling study. Clin Pharmacol Ther 86:634–643

45. Willmann S, Hohn K, Edginton A, Sevestre M, Solodenko J, Weiss W, Lippert J, Schmitt W (2007) Development of a physiology-based whole-body population model for assessing the influence of individual variability on the pharmacokinetics of drugs. J Pharmacokinet Pharmacodyn 34:401–431
46. Willmann S, Lippert J, Schmitt W (2005) From physicochemistry to absorption and distribution: predictive mechanistic modelling and computational tools. Exp Opin Drug Metab Toxicol 1:159–168
47. Willmann S, Lippert J, Sevestre M, Solodenko J, Fois F, Schmitt W (2003) PK-Sim©: a physiologically based pharmacokinetic 'whole-body' model. Biosilico 1:121–124
48. Willmann S, Schmitt W, Keldenich J, Dressman JB (2003) A physiologic model for simulating gastrointestinal flow and drug absorption in rats. Pharm Res 20:1766–1771
49. Willmann S, Schmitt W, Keldenich J, Lippert J, Dressman JB (2004) A physiological model for the estimation of the fraction dose absorbed in humans. J Med Chem 47:4022–4031

Chapter 33
Re-analysis of Bipolar Disorder and Schizophrenia Gene Expression Complements the Kraepelinian Dichotomy

Kui Qian, Antonio Di Lieto, Jukka Corander, Petri Auvinen, and Dario Greco

Abstract The differential diagnosis of schizophrenia (SZ) and bipolar disorder (BD) is based solely on clinical features and upon a subset of overlapping symptoms. Within the last years, an increasing amount of clinical, epidemiological and genetic data suggested inconsistent with the Kraepelinian dichotomy. We performed re-analysis of genome-wide gene expression data obtained from *postmortem* prefrontal cortex (PEC) of both BD and SZ patients with matched controls from four independent microarray experiments. We found 2,577 and 477 genes specifically altered in BD and SZ, respectively. Of these, 164 genes were shared between the syndromes. We identified genes of the transcriptional and post-transcriptional machineries altered in BD and genes of the development changed in SZ. Our results showed that the genomic expression profile of BD and SZ had some similarity but still could be well-distinguished by suitable statistical test.

K. Qian • P. Auvinen
Institute of Biotechnology, University of Helsinki, Helsinki, Finland
e-mail: kui.qian@helsinki.fi; petri.auvinen@helsinki.fi

A. Di Lieto
Neuroscience Centre, University of Helsinki, Helsinki, Finland
e-mail: antonio.dilieto9@gmail.com

J. Corander
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: jukka.corander@helsinki.fi

D. Greco (✉)
Department of Bioscience and Nutrition, Karolinska Institutet, 141 83 Huddinge, Stockholm, Sweden
e-mail: dario.greco@ki.se

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_33,

1 Introduction

Schizophrenia (SZ) and bipolar disorder (BD) are psychiatric syndromes affecting each ∼1% of the population worldwide. Although their incidence is relatively low, these conditions are major contributors to the global burden of diseases. The relatively early onset, and the persistence or fluctuation of symptoms, has devastating consequences on the quality of life of the patients.

SZ is mainly characterized by a combination of hallucinations and delusions and is often associated with specific cognitive deficits. On the other hand, BD is marked by an alternation of elevated and depressed mood with or without psychotic symptoms. Their diagnosis is based solely on clinical features because validating diagnostic tests are, currently, not available. Differential diagnosis is based upon a subset of overlapping symptoms [1]. Despite, their hereditability is estimated to be higher than in other diseases of the central nervous system or in some cancers [2], evidence for genetic candidate factors is still far from being robust or conclusive [3].

Descriptions of the major psychiatric diseases were recorded in the 19th century by Kraepelin, who described two distinct disorders (the Kraepelinian dichotomy): dementia praecox, renamed SZ by Bleuler in 1911, and manic depressive insanity, now called BD. This classification forms the basis of modern diagnostic system as defined in DSM IV and ICD-10. Within the last years, the traditional Kraepelinian dichotomy dividing SZ and BD has been strongly challenged [4]. An emergent amount of clinical, epidemiological and genetic data questioned the model that SZ and BD would be two independent syndromes [4, 5]. Recently, an extensive population-based study on the Swedish registers strongly shows that SZ and BD partially share genetic cause (up to 63%) [6], and another study shows that common polygenic variation that contributes to risk of these syndromes are shared between SZ and BD [7]. Large-scale genome-wide surveys of rare copy number variants (CNVs) [8, 9] and genome-wide association (GWA) studies [10] have recently been undertaken to decipher the number and type of genetic variants involved SZ and BD.

Microarray technologies have been used to address the gene expression in *postmortem* human brain samples. Different groups of genes have been described as differentially expressed in these syndromes: genes encoding for transcripts of synaptic protein [11, 12] cell growth and development of CNS [13], myelination [14], apoptosis [15], mitochondrial dysfunction [16], oligodendrocyte functions [17], receptors channels, transporters and signal transduction [18], oxidative stress [19], neurotransmission [20] and ubiquitination [12]. However, these results are partially confirmed and concordant.

Combining microarray data from independent experiments in re-analysis fashion can generate a more comprehensive understanding of the genome biology of SZ and BD by increasing the statistical power of the analysis. There are two general approaches for re-analyzing microarray experiments: (1) by comparing the results published in several studies; (2) by comprehensively re-analysing the primary data from several experiments [21]. The latter approach has been successfully used in searching for human tissue-selective gene expression patterns [22].

Similarly, here we have carried out re-analysis of Affymetrix GeneChip data from human dorsolateral prefrontal cortex (PFC) of SZ and BD patients as well as healthy subjects publicly available at the Stanley Medical Research Institute (SMRI).

The aims of this study are: (1) to identify families of genes involved in pathophysiology of SZ and BD and (2) to define the gene expression similarities and differences between these syndromes.

2 Materials and Methods

2.1 Microarray Data Collection

Microarray data were collected from the SMRI Online Genomics Database (www.stanleygenomics.org), where 20 independent studies performed on two collections of postmortem human samples (the Stanley Array collection and the Stanley Consortium collection) are stored. Data were selected according to the following criteria: (1) microarrays performed on PEC; (2) the samples extracted from patients with BD or SZ, and normal control individuals; (3) each individual sample profiled only in one microarray (samples used in more than one studies were selected only once); (4) the gene expression would be studied by using Affymetrix U133A GeneChips. For each sample, the microarray raw file (CEL file) and the demographic data were retrieved.

2.2 Microarray Data Quality Control

Extensive QC analysis was carried out on each array data. Particularly, the RNA degradation state, the spatial distribution of foreground and background signals over the slide, as well as the signals from the control probes were inspected. The facilities in the BioConductor packages *affyQCReport* [23] and *affyPLM* [24] were used for this task. Arrays were eliminated from further analysis if matching one of these conditions: (1) artefacts were screened by the affyPLM analysis; (2) the 3′–5′ ratios for the genes GAPDH and β-actin were aberrant and the RNA degradation plot was abnormal.

2.3 Re-annotation of Affymetrix Probes

To account for recent advances in genomics, the probes on each chipset were re-annotated and re-arranged in newly defined probe sets according to the Entrez Gene

database. To achieve this, custom CDF v.12 files were retrieved from the BrainArray Laboratory website (http://brainarray.mbni.med.umich.edu). After re-assignment, each new probe set contained only probes perfectly matching with an individual and unique gene sequence from the Entrez Gene database.

2.4 Microarray Data Pre-processing

Gene expression values for each re-annotated probe set were calculated by RMA algorithm [25] implemented in the BioConductor *affy* package [26]. Briefly, in the RMA algorithm, the signals of the probes of each probe sets are background corrected, normalized using the quantile method and summarized by the median polish technique [24].

2.5 Differential Expression Analysis

To testing which of the 17 demographic variables are associated with the genes expression value, the global effect of each demographic variable on the totality of the genes analysed was evaluated using the BioConductor package *globaltest* [27]. The variables containing missing data were omitted, as they would have effects on the accuracy of the testing. In this analysis, three different models were used according to the nature of each variable: (1) logistic regression model for the discrete variables with two categories; (2) multinomial regression model for discrete variables with more than two categories; (3) linear regression model for continuous numerical variables. Variables with significant effects (p-value < 0.001) were selected and used in the statistical test models.

A linear model including strong associated demographic variables was used to explain the expression level of each gene across all the arrays of the dataset. Contrasts were defined to evaluate the expression differences between BD *versus* control and SZ *versus* control, and tested by moderated t-test. The Bioconductor package *limma* [28] was used for this task. Genes presenting p-value <0.05 were considered to be significantly expressed and used for further functional analysis.

2.6 Functional Analysis

The DAVID functional annotation tool was used to search for over-represented Gene ontology families by the Fisher's exact test [29, 30]. Gene ontology families with nominal p-value <0.05 were considered to be significant. This analysis was done using all the differentially expressed genes of each contrast analysed and the subgroups of upregulated and downregulated genes.

WebGestalt2 (http://bioinfo.vanderbilt.edu/webgestalt/) was used to study the gene set enrichment in KEGG pathways, targets of transcription factors, targets of microRNAs and the cytogenetic bands [31]. For all these test, the group of all the genes in the human genome was used as a reference set and three gene lists were utilized as input for the test: (a) all differently expressed genes in BD, (b) all differently expressed genes in SZ and (c) common genes in gene list a and b. Results with multiple test adjustment of false discovery rate (BH) <0.05 were considered as significant.

2.7 Diagnostic Cross-Validation

To accurate classification of our samples, we performed cross-validation on 134 arrays. Straight leave-one-out cross-validation (LOOCV) was carried out by *MLInterfaces* package, using the diagnosis as the classes. In addition, Bayesian model averaging (BMA) method for gene selection and classification was utilized to perform LOOCV through *iterativeBMA* package [32].

3 Results

A total of 400 CEL files from four independent experiments were collected and checked about their RNA and image quality. Five of them presented RNA degradation and 261 arrays showed varying degrees of artefacts or nonuniform spatial distribution of probe signals, leading to a total number of 266 arrays ($\sim$66%) that were eliminated from our dataset. Details of the dataset information were shown in (Table 33.1). All of the 134 arrays that successfully passed the quality control were re-annotated according to the probe mapping to the Entrez Gene database [33]. In the original annotation released from the manufacturer Affymetrix, 21,722 probe sets were, respectively, present on the U133A chipset. After re-annotation, as many as 11,911 probe sets were newly defined.

3.1 Gene Expression Modelling and Statistical Test

A linear model was fitted taking into account the demographic variables with a possible strong effect on the gene expression: study ID, collection type and brain pH ($p < 0.001$, Table 33.2). Three studies were in collection A and one in collection C, suggesting the study ID would be more suitable than collection type. Finally, study ID and brain pH were included in the statistical test for gene expression.

The p-value distribution patterns of BD and SZ test results were investigated (Fig. 33.1). For different p-values, there were always an obvious high number

Table 33.1 Summary of the selected microarray experiments in SMRI

Study ID	Investigator	Samples	RNA degradation	Artefact	Eliminated	Array used	Control	Bipolar disorder	Schizophrenia	Collection*	Region	Array type
1	AltarA	98	3	63	66	32	17	3	12	A	FrontalBA46	hgU133A
2	AltarC	72	2	40	42	30	10	9	11	C	FrontalBA46/10	hgU133A
3	Bahn	101	0	57	57	44	9	21	14	A	FrontalBA46	hgU133A
7	Kato	102	0	74	74	28	9	10	9	A	FrontalBA46	hgU133A

The study IDs and investigators' information were named from the Stanley Medical Research Institute online genomics database. Artefact presents the number of arrays that have the artefacts or probe signal spatial bias in their pseudo image by *affyPLM* analysis. Collection is the origin of the samples in SMRI database.
*A = array collection, C = neuropathology consortium

Table 33.2 The effect of each demographic variable on the gene expression profiles. PMI: Postmortem interval

	Bipolar disorder	Schizophrenia
Study ID**	7.57×10^{-8}	1.22×10^{-12}
Collection type**	0.00039	3.00×10^{-5}
Age	0.585	0.143
Sex	0.0569	0.0372
Race	0.0761	0.517
Axis I primary diagnosis	0.00269	0.492
PMI	0.126	0.178
Brain PH**	2.91×10^{-6}	0.000633
Left brain	0.383	0.211
Suicide status	0.105	0.893
Psychotic feature	0.0129	0.278
Rate of death	0.0131	0.0163
Exacerbation	0.00244	0.268
Smoking at time of death	0.719	0.854
Lifetime alcohol	0.834	0.9
Lifetime drugs	0.182	0.0419
Lifetime anti-psychotics	0.214	0.428

** p-Value < 0.001

of genes in range of smallest p-value in BD, while for SZ, there were no such pattern observed (p-value < 0.05 and p-value < 0.001, the 1st bar on left for each histogram in Fig. 33.1). With the threshold of p-value < 0.05, a total of 2,577 genes were found differentially expressed in BD *versus* control (Fig. 33.1a). Out of these, 1,319 were upregulated and 1,259 were downregulated. As many as 477 genes were differentially expressed in SZ against control samples (Fig. 33.1b), of which 197 were induced and 280 were repressed. A total of 164 genes were both dysregulated in BD and SZ, of which some had been reported before, such as AIF1, BID, CACNA1A, CYP2C19, JMJD7-PLA2G4B, NDE1, NOS1AP, NPY, PENK, PER1, PTGS1, SST and TAC1.

Further, the common genes differentially expressed in both disorders showed similar expression pattern (Fig. 33.2). We checked this phenomenon by utilizing four different linear models. In the univariate statistical model Y $\sim$ diagnosis (Fig. 33.2d) as well as in a multivariate model Y $\sim$ diagnosis + pH (Fig. 33.2c), all the common differentially expressed genes show the same altered direction. In the model Y $\sim$ diagnosis + study, DDX39B (Entrez Gene ID 7919) and CUTC (Entrez Gene ID 51076) were suppressed in BD but induced in SZ. Meanwhile, CXorf36 (Entrez Gene ID 79742) were induced in BD but suppressed in SZ (Fig. 33.2b). In the model Y $\sim$ diagnosis + study + pH, only DDX39B (Entrez Gene ID 7919) was upregulated in SZ and downregulated in BD (Fig. 33.2a).

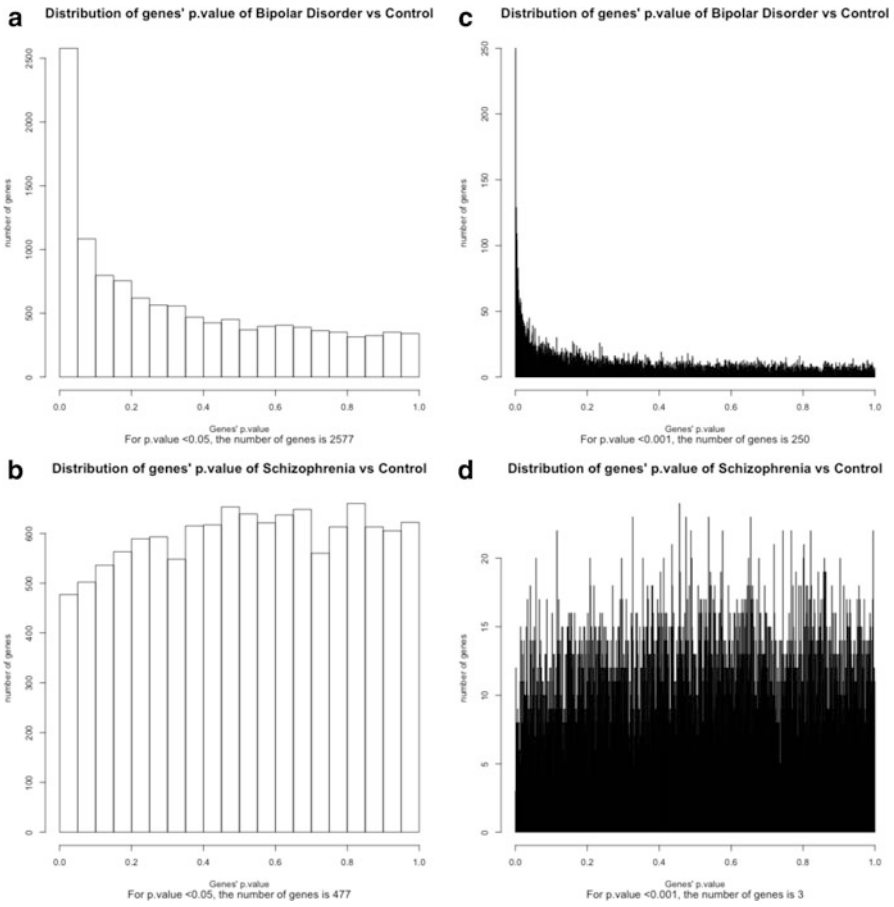


Fig. 33.1 Distribution of genes' p-values of schizophrenia (**b**, **d**) and bipolar disorder (**a**, **c**). Number of genes (y-axis) represents the count within the range of corresponding p-value (x-axis). Two different width of column was used: 0.05 (**a**, **b**) or 0.001 (**c**, **d**) for each column

3.2 Functional Analysis

For more functional annotation, we looked for a number of features in the dataset. The most significant enriched KEGG pathways were "metabolic pathway" for BD ($p = 1.67 \times 10^{-44}$), "MAPK signalling pathway" for SZ ($p = 2.04 \times 10^{-5}$) and "arachidonic acid metabolism" for the common gene list ($P = 0.0315$) (Table 33.3). We were also interested about the possible transcription factors that could be involved and as a result of this analysis we saw the most significantly enriched transcription factor consensus sequence was GGGCGGR for SP1 in BD ($p = 5.17 \times 10^{-95}$), TTGTTT for FOXO4 in SZ ($p = 4.51 \times 10^{-17}$) and

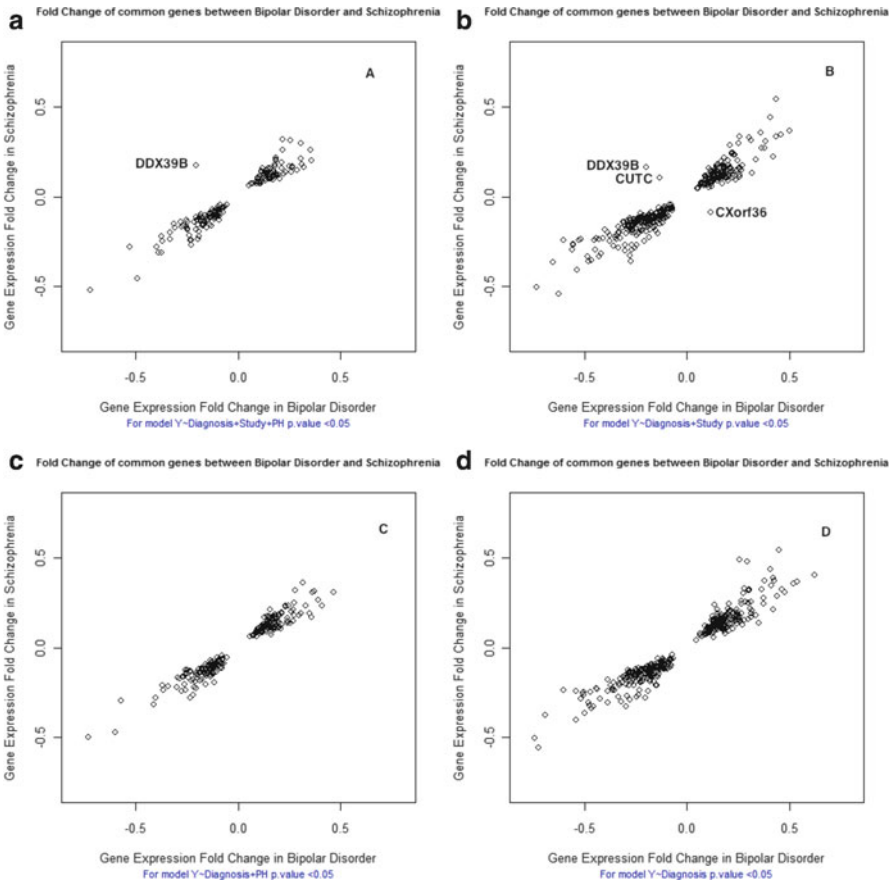


Fig. 33.2 Fold change of common genes between bipolar disorder (BD) and schizophrenia (SZ). The fold change of each gene in BD was plotted in x-axis and SZ in y-axis. Four different statistical models were checked and suggested that this phenomenon was not specific to models

GTGACGY for E4F1 in common gene list ($p = 0.0003$). In addition to the transcription factor, one obvious target would be the possible involvement of miRNA regulations. The most significant enriched miRNA consensus sequence was TGCTGCT (MIR-15A, MIR-16, MIR-15B, MIR-195, MIR-424 and MIR-497) in BD ($p = 3.19 \times 10^{-33}$), TATTATA (MIR-374) in SZ ($p = 5.92 \times 10^{-5}$) and CACTGCC (MIR-34A, MIR-34C and MIR-449) for the common gene list ($p = 4.21 \times 10^{-5}$). We did also analyse the higher-order organization involved and observed some cytobands that were enriched in this data. The most significant enriched the cytogenetic bands were chr16q22 for BD ($p = 2.87 \times 10^{-6}$), chr19p for SZ ($p = 0.0068$) and chr2p14 for common gene list ($p = 0.0093$).

Table 33.3 Enriched top KEGG pathways with adjusted p-value

KEGG pathways	Adjusted p-value
BD	
Metabolic pathways	1.67×10^{-44}
Pathways in cancer	1.95×10^{-11}
Endocytosis	1.56×10^{-9}
Wnt signalling pathway	3.93×10^{-9}
Purine metabolism	6.15×10^{-8}
Melanogenesis	6.24×10^{-8}
Chemokine signalling pathway	6.29×10^{-8}
Spliceosome	6.29×10^{-8}
Cytokine–cytokine receptor interaction	1.41×10^{-7}
MAPK signalling pathway	4.83×10^{-7}
SZ	
MAPK signalling pathway	2.04×10^{-5}
Arachidonic acid metabolism	8.77×10^{-5}
Pathways in cancer	8.77×10^{-5}
Focal adhesion	0.0002
Long-term depression	0.0002
Chemokine signalling pathway	0.0006
Neurotrophin signalling pathway	0.0008
Regulation of actin cytoskeleton	0.001
Metabolic pathways	0.001
Fc epsilon RI signalling pathway	0.0019
Common gene	
Arachidonic acid metabolism	0.0315
Glycerophospholipid metabolism	0.0315
Homologous recombination	0.0375
Linoleic acid metabolism	0.0375
Nucleotide excision repair	0.0489
Spliceosome	0.0489
Vascular smooth muscle contraction	0.0489

3.3 Class Prediction and Cross-Validation

Since there was no test set in this dataset, the LOOCV procedure was applied. We used two different LOOCV algorithms: k-nearest neighbour classification and Bayesian model averaging classification. The k-nearest neighbour classification showed about 27% error rate (12/43) to classify between BD and control samples, and much higher error rate $\sim$59% (27/46) for SZ. On the other hand, the iterative BMA algorithm showed excellent prediction capability (Table 33.4), as all the 43 BD samples was predicted into the BD class, and so were the 46 SZ and the 45 controls samples.

Table 33.4 Results of leave one out cross-validation

		Predicted			Predicted			Predicted	
		BD	Control		SZ	Control		BD	SZ
k-Nearest neighbour classification	BD	31	12	SZ	19	27	BD	29	14
	Control	13	32	Control	17	28	SZ	22	24
Bayesian model averaging classification	BD	43	0	SZ	46	0	BD	43	0
	Control	0	45	Control	0	45	SZ	0	46

4 Discussion

We have integrated and re-analysed publicly available microarray gene expression data of *postmortem* PEC of BD and SZ patients, and control subjects. The integrated large-scale dataset had been carefully re-analysed at every step, extensive statistical modelling had been utilized and several aspects of the biological annotations were present in the results.

The independence of the samples is an issue when working with data based on a common brain collection, as some subjects can be utilized for more than one assay. This is the case with the brain collections available at the Stanley Foundation. In the year 2008, Choi and collaborators used a post hoc correction that took into account the over-estimation of the degrees of freedom [34]. Here, we strictly selected the data based on the quality of arrays, so that microarrays hybridized to the same subject were included only once in the final analysis. We are convinced that this quality control step increases the accuracy of the final results by improving the accuracy of the quality of the utilized data.

It is always a difficult task to deal with multiple factor experiments, especially for clinical samples where many biological and demographic variables can affect the expression of many genes. Several variables of this dataset had been reported to be associated with the disorders, such as age [35–38], gender [39, 40], PMI [41, 42] and lifetime drug usage [43, 44]. However, we showed that even though many of them had modest effects on BD or SZ (p-value < 0.05 but > 0.001), the most important variable was pH (p-value $= 2.91 \times 10^{-6}$ for BD, p-value $=$ 0.000633 for SZ), which was already reported in many studies [42, 44–46].

Under the null hypothesis of no differential expression in statistical test, the p-values would be uniformly distributed in the range [0, 1] [47]. The histograms of p-value of BD *versus* control were densely distributed near zero and become less as the p-values increased and this pattern was shown to be the same on two different scales (Fig. 33.1a and c). These distributions indicated that the test method was well-fit in our linear model and suggests that genes with low p-value were reliably differentially expressed. The flat histograms of SZ *versus* control did not indicate

any violation of the assumptions of statistical test but suggested that the genes with low p-value might not be statistically significant after adjusting for multiple testing (Fig. 33.1b and d).

Even though the p-value distributions were quite different between BD and SZ, the genes both altered in SZ and BD showed similar expression patterns (Fig. 33.2). This would indicate that there are some common molecular mechanisms shared by BD and SZ.

We observed several biologically relevant aspects of BD and SZ physiopathology. The post-transcriptional regulation by miRNAs has been heavily investigated also in the field of the neurosciences [48]. For the genes differentially expressed both in SZ and BD, the predicted miRNA binding sites were also consistent with recent reports on hsa-mir-34a [49] and has-mir-195 [50]. Many important transcription factors involved in BD and SZ pathogenesis were also retrieved in our predicted list: CREB, OCT, P53, SP1, TATA, ATF4 and MYC [51–56]. Several KEGG pathway in our gene set enrichment results were already studied, such as arachidonic acid metabolism, linoleic acid metabolism and nucleotide excision repair [57–59].

Classification of gene expression data could be utilized to predict the diagnostic category of new samples [60–63]. Our results also suggested that the profiles of SZ are less different from controls than BD and hence more difficult to classify (Table 33.3).

Our results suggest that the Bayesian approach is more efficient and reliable as compared to the more traditionally utilized k-nearest neighbour method. Though this may be depended on the selection of samples and many critical pre-steps we used, this method would be a great tool to classify new microarray data of SZ and BD as well as to support their differential diagnosis.

Taken altogether, our results suggest that, even though BD and SZ share some features of their expression profiles, they are still characterized by distinct gene expression patterns, supporting the Kraepelinian dichotomy. Supporting the dichotomy at gene expression level might have a major impact not only at diagnostic level but also in order to identify novel target genes for development of specific drugs, consequently improving the care of patients.

Alternative approaches for better studying gene expression in complex tissues could involve the novel sequencing technologies that allow better resolution and precision than microarray measurements [64]. Thus, we suggest that future investigation by more accurate methods like deep sequencing can be useful to highlight smaller expression differences as well as to detect qualitative differences of the transcripts, such as alternative splicing, differential usage of promoters usage of alternative polyadenylation sites.

Acknowledgements This study has been funded by the Institute of Biotechnology, University of Helsinki (Finland), and by the Academy of Finland (PA). DG was funded by the Paulon Säätiö (Finland). The work of JC was supported by the grant no. 121301 from Academy of Finland. The authors are deeply grateful to Prof. Eero Castrén and Dr Iiris Hovatta for critical comments on the manuscript.

References

1. American Psychiatric Association Task Force on DSM-IV (2000) Diagnostic and statistical manual of mental disorders: DSM-IV-TR, 4th edn. American Psychiatric Association, Washington, pp xxxvii, 943
2. Consortium WTCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678
3. Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60(12):1187–1192
4. Craddock N, O'Donovan MC, Owen MJ (2006) Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. Schizophr Bull 32(1):9–16
5. Consortium IS (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237–241
6. Lichtenstein P et al (2009) Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373(9659):234–239
7. Purcell SM et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748–752
8. Stefansson H et al (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455(7210):232–236
9. Hunsberger JG et al (2009) micrornas in mental health: from biological underpinnings to potential therapies. Neuromol Med 11(3):173–182
10. Burmeister M, McInnis MG, Zollner S (2008) Psychiatric genetics: progress amid controversy. Nat Rev Genet 9(7):527–540
11. Mirnics K et al (2000) Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. Neuron 28(1):53–67
12. Ryan MM et al (2006) Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatry 11(10):965–978
13. Nakatani N et al (2006) Genome-wide expression analysis detects eight genes with robust alterations specific to bipolar I disorder: relevance to neuronal network perturbation. Hum Mol Genet 15(12):1949–1962
14. Tkachev D et al (2007) Further evidence for altered myelin biosynthesis and glutamatergic dysfunction in schizophrenia. Int J Neuropsychopharmacol 10(4):557–563
15. Benes FM et al (2005) The expression of proapoptosis genes is increased in bipolar disorder, but not in schizophrenia. Mol Psychiatry 11(3):241–251
16. Iwamoto K, Bundo M, Kato T (2005) Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. Hum Mol Genet 14(2):241–253
17. Tkachev D et al (2003) Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. Lancet 362(9386):798–805
18. Iwamoto K et al (2004) Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. Mol Psychiatry 9(4): 406–416
19. Prabakaran S et al (2004) Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. Mol Psychiatry 9(7):684–697, 643
20. Hashimoto T et al (2008) Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. Mol Psychiatry 13(2):147–161
21. Larsson O, Wennmalm K, Sandberg R (2006) Comparative microarray analysis. OMICS 10(3):381–397
22. Greco D et al (2008) Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. PLoS One 3(4):e1880
23. Parman C, Halling C (2009) affyQCReport: a package to generate QC reports for affymetrix array data

24. Bolstad BM et al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185–193
25. Irizarry RA et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264
26. Gautier L et al (2004) affy – analysis of Affymetrix genechip data at the probe level. Bioinformatics 20(3):307–315
27. Goeman JJ et al (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20(1):93–99
28. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (eds) Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer, pp 397–420
29. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13
30. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
31. Zhang B, Kirov S, Snoddy J (2005) webgestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33(Web Server issue):W741–748
32. Yeung KY, Bumgarner RE, Raftery AE (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21(10):2394–2402
33. Maglott D et al (2007) Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 35(Database issue):D26–31
34. Choi KH et al (2008) Putative psychosis genes in the prefrontal cortex: combined analysis of gene expression microarrays. BMC Psychiatry 8:87
35. Pedersen CB, Mortensen PB, Cantor-Graae E (2011) Do risk factors for schizophrenia predispose to emigration? Schizophr Res 127(1–3):229–234
36. Fatjo-Vilas M et al (2011) Dysbindin-1 gene contributes differentially to early- and adult-onset forms of functional psychosis. Am J Med Genet B Neuropsychiatr Genet 156(3):322–333
37. Gruber O et al (2011) A systematic experimental neuropsychological investigation of the functional integrity of working memory circuits in major depression. Eur Arch Psychiatry Clin Neurosci 261:179–184
38. McIntosh BJ et al (2011) Performance-based assessment of functional skills in severe mental illness: results of a large-scale study in China. J Psychiatr Res:45(8):1089–1094
39. Mueser KT et al (2010) Neurocognition and social skill in older persons with schizophrenia and major mood disorders: An analysis of gender and diagnosis effects. J Neurolinguistics 23(3):297–317
40. Pedersen CB, Mortensen PB, Cantor-Graae E Do risk factors for schizophrenia predispose to emigration? Schizophr Res 127(1–3):229–234
41. Hashimoto K, Sawa A, Iyo M (2007) Increased levels of glutamate in brains from patients with mood disorders. Biol Psychiatry 62(11):1310–1316
42. Rollins B et al (2009) Mitochondrial variants in schizophrenia, bipolar disorder, and major depressive disorder. PLoS One 4(3):e4913
43. Nasrallah HA, Brecher M, Paulsson B (2006) Placebo-level incidence of extrapyramidal symptoms (EPS) with quetiapine in controlled studies of patients with bipolar mania. Bipolar Disord 8(5 Pt 1):467–474
44. Lewis R, Bagnall AM, Leitner M (2005) Sertindole for schizophrenia. Cochrane Database Syst Rev (3):CD001715. Review. PMID: 16034864 [PubMed - indexed for MEDLINE]
45. Thompson Ray M et al (2011) Decreased BDNF, trkb-TK+ and GAD(67) mrna expression in the hippocampus of individuals with schizophrenia and mood disorders. J Psychiatry Neurosci 36(1):100048
46. Wang JF et al (2009) Increased oxidative stress in the anterior cingulate cortex of subjects with bipolar disorder and schizophrenia. Bipolar Disord 11(5):523–529
47. Pounds SB (2006) Estimation and control of multiple testing error rates for microarray studies. Brief Bioinform 7(1):25–36

48. Moreau MP et al (2011) Altered microrna expression profiles in postmortem brain samples from individuals with schizophrenia and bipolar disorder. Biol Psychiatry 69(2):188–193
49. Kim AH et al (2010) microrna expression profiling in the prefrontal cortex of individuals affected with schizophrenia and bipolar disorders. Schizophr Res 124(1–3):183–191
50. Dinan TG (2010) micrornas as a target for novel antipsychotics: a systematic review of an emerging field. Int J Neuropsychopharmacol 13(3):395–404
51. Gershon ES, Alliey-Rodriguez N, Liu C (2011) After GWAS: searching for genetic risk for Schizophrenia and bipolar disorder. Am J Psychiatry 168(3):253–256
52. Souza BR et al (2011) Downregulation of the camp/PKA pathway in PC12 cells overexpressing NCS-1. Cell Mol Neurobiol 31(1):135–143
53. Yuan P et al (2010) Altered levels of extracellular signal-regulated kinase signaling proteins in postmortem frontal cortex of individuals with mood disorders and schizophrenia. J Affect Disord 124(1–2):164–169
54. Yuan P et al Altered levels of extracellular signal-regulated kinase signaling proteins in postmortem frontal cortex of individuals with mood disorders and schizophrenia. J Affect Disord 124(1–2):164–169
55. Kakiuchi C et al (2007) Association analysis of ATF4 and ATF5, genes for interacting-proteins of DISC1, in bipolar disorder. Neurosci Lett 417(3):316–321
56. Ubhi K, Price J (2005) Expression of POU-domain transcription factor, Oct-6, in schizophrenia, bipolar disorder and major depression. BMC Psychiatry 5:38
57. Iwayama Y et al (2010) Association analyses between brain-expressed fatty-acid binding protein (FABP) genes and schizophrenia and bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 153B(2):484–493
58. Quinones MP, Kaddurah-Daouk R (2009) Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. Neurobiol Dis 35(2):165–176
59. Adibhatla RM, Hatcher JF (2010) Lipid oxidation and peroxidation in CNS health and disease: from molecular mechanisms to therapeutic opportunities. Antioxid Redox Signal 12(1): 125–169
60. Shedden K et al (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 14(8):822–827
61. van't Veer LJ, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452(7187):564–570
62. Lee SC et al (2009) Post-treatment tumor gene expression signatures are more predictive of treatment outcomes than baseline signatures in breast cancer. Pharmacogenet Genomics 19(11):833–842
63. Zaas AK et al (2010) Blood gene expression signatures predict invasive candidiasis. Sci Transl Med 2(21):21ra17
64. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135–151

Chapter 34
Bringing Together Models from Bottom-Up and Top-Down Approaches: An Application for Growth of *Escherichia coli* on Different Carbohydrates

Andeas Kremling

Abstract Modeling in systems biology follows two lines: a data driven top-down approach that integrates experimental data from various "omics" technologies and a model based bottom-up approach where the model structure is given and kinetic parameters are chosen in such a way that an experimental observation can be reproduced quantitatively or qualitatively. Mathematical models are frequently used to elucidate cellular design principles in order to understand complex biochemical networks better. To show that both approaches lead to a consistent description of cellular dynamics, mathematical models from both approaches are explored. On the level of transcription factor activities a sufficient qualitative agreement is observed. Experimental data for the classical growth experiment of *Escherichia coli* on two carbon sources, glucose and lactose is available to set up the data driven model and to support the theoretical findings from the bottom-up approach.

1 Introduction

A quantitative description of cellular processes offers new possibilities in medical applications or biotechnology. This requires the availability of time course data of the interesting state variables that nowadays can be found in a number of data bases. In systems biology two approaches are established to derive a quantitative description. In the bottom-up approach, starting from smaller networks, the structure of the mathematical model is based on balance equations that are represented by differential equations. These equations are characterized by a number of (in most cases unknown) kinetic parameters. To determine the parameters, often the models have to be calibrated based on time course data. In general, time course data are

A. Kremling (✉)
Department of Systems Biotechnology, Technische Universität München,
Boltzmannstr. 15, 85748 Garching, Germany
e-mail: a.kremling@lrz.tum.de

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_34,

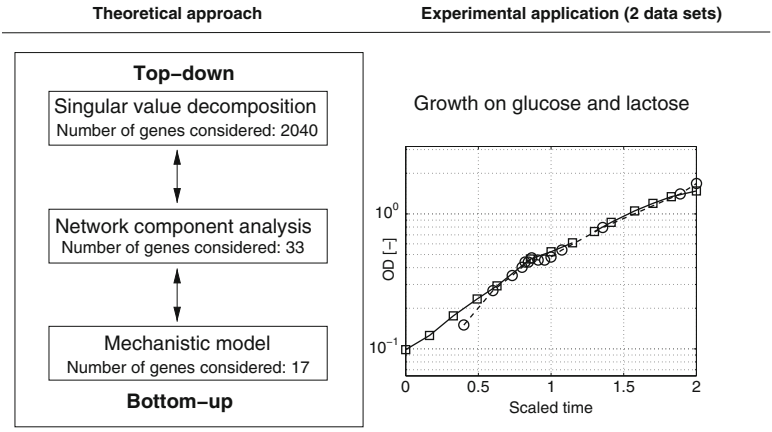


Fig. 34.1 Outline of the approach and experimental data used for application. The optical densities in both data sets are very similar (*circles* [2], *squares* [1]). Time scaling was performed: phase 1 characterizes growth on glucose while phase 2 characterizes growth on lactose

not available for all state variables in the model. Therefore, simulation studies are performed to elucidate the dynamics of those state variables and to determine for example parametric sensitivities. In contrast, to perform a top-down approach, data from different cellular levels (named "omics") are integrated and analyzed. These data characterize the overall status of the cell, since all transcripts or all fluxes are determined and are often used together with techniques from multivariate data analysis to infer properties of the cellular network.

In the present contribution different methods are applied to show the consistency between the two approaches. This is achieved on the level of transcription factor activities. Transcription factors are responsible for coordinated expression of genes during growth, but also in the case of changes of environmental situations. A classical example is growth of *Escherichia coli* on two substrates. It is well known, that during growth on glucose and lactose, the first one is consumed while the second one is not until the first one is run out. In the situation when glucose is running out, several cellular programs are started to cope with the stress and to adapt to the new situation. Here, two data sets are used and three models are introduced – one based on a bottom-up approach, and two based on a top-down approach – to simulate the activities of several transcription factors and to show a good qualitative agreement between bottom-up and top-down approaches. Figure 34.1 summarizes the approaches giving also the number of genes represented in the respective models.

A prerequisite for the application of the proposed strategy is a consistent data set. Two data sets are used [1, 2] that show a comparable optical density as shown for the two experiments in Fig. 34.1. Since different initial conditions are used for the substrates, time scaling was performed. In next sections models are introduced

and in Sect. 5 the simulation results are compared. Section 6 focuses on a further regulatory circuit observed during the experiment, namely, the stress response. Finally by-product secretion is analyzed in Sect. 7.

2 Bottom-Up Model

To describe carbohydrate uptake in *E. coli* a very detailed model was introduced [1] that comprises uptake and metabolism of six carbohydrates, including the description of global signaling, signal processing, and gene expression. The model was validated by fitting kinetic parameters against a very comprehensive experimental data base (18 experiments with 5 different strains) and describes expression of 17 key enzymes, 38 enzymatic reactions, and the dynamic behavior of more than 50 metabolites. In contrast to a model that was published very recently on the same topic [11], the model shows predictive character and is able to forecast experiments with good accuracy [9].

In the central pathways two metabolites have a distinguished role in the model. On the one hand, the phosphoenolpyruvate (PEP) to pyruvate ratio determines the degree of the phosphorylation of the proteins of the phosphotransferase system (PTS). This is described in detail in [7, 8]. Here, protein EIIA of the PTS in its phosphorylated form is an activator for the adenylate cyclase which produces cAMP. cAMP is a co-factor for transcription factor Crp that is involved in gene regulation of a number of genes related to central metabolism as well as others (for the current number see the Ecocyc database [6]). This straight forward activation is shown in Fig. 34.2. On the other hand, fructose-1,6-bisphosphate is a co-factor for transcription factor FruR (also named Cra). FruR is mainly involved in gene expression of genes of glycolysis and gluconeogenesis. FruR interacts with fructose-1,6-bisphosphate and activates or inhibits the respective genes. However, depending on the mode of regulation, the relationship between co-factor and transcription factor is more complex, as summarized in Fig. 34.2. As can be seen in the inserted

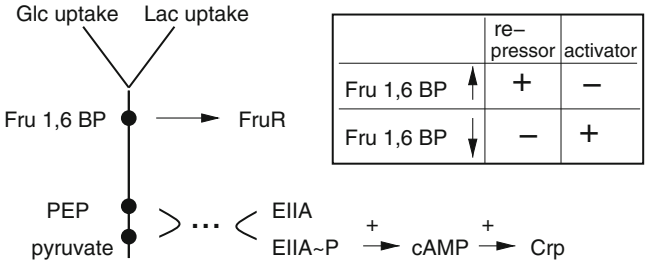


Fig. 34.2 Scheme that describes the two signaling pathways starting from central pathway glycolysis and ending at the two transcription factors FruR and Crp, respectively. Details are given in the text

table, high values of fructose-1,6-bisphosphate change the sign of the regulatory action. If FruR acts as a repressor, high values of fructose-1,6-bisphosphate lead to a deactivation of FruR and therefore the respective genes are transcribed (plus sign). Contrary, if FruR is an activator the deactivation leads to a decrease of the regulatory activity (minus sign). If the level of fructose-1,6-bisphosphate is low, the opposite values are true. For the experimental situation – growth on glucose and lactose – fructose-1,6-bisphosphate is rather high and therefore the sign of the regulatory action has to be inverted.

3 Top-Down Models

3.1 Singular Value Decomposition

Experimental data obtained from an "omics" approach can be analyzed in different ways. A straight forward approach is the elucidation of the characteristic (or dominant) modes of the data with the help of a singular value decomposition (SVD) approach [5]. Formally, a given data set, the mRNA data is given as a matrix with every row representing the time course of one specific mRNA. SVD decomposes the matrix **mRNA** in three matrices $\mathbf{U}$, Σ, and $\mathbf{V}^T$ ($\mathbf{U}$ and $\mathbf{V}$ are orthonormal systems, i.e., $\mathbf{U}\,\mathbf{U}^T = \mathbf{I}$, $\mathbf{V}\,\mathbf{V}^T = \mathbf{I}$) in such a way that

$$\mathbf{mRNA} = \mathbf{U} \cdot \Sigma \cdot \mathbf{V}^T. \tag{34.1}$$

Matrix **mRNA** has n rows (one for each gene) and t_k columns. In economy size, the dimension of the respective other matrices are: $\mathbf{U}$ $n \times t_k$, Σ $t_k \times t_k$, $\mathbf{V}$ $t_k \times t_k$. The matrix Σ contains the singular values in decreasing order. It is observed that most of the values are small compared to the first ones, so one can reduce the matrix Σ in such a way that only the main modes of the system are calculated to decompose matrix **mRNA**. The number of singular values considered here is s. So, the matrices have the following dimensions: $\mathbf{U}'$ $n \times s$, Σ' $s \times s$, $\mathbf{V}'$ $s \times t_k$. To be conform with the approach introduced in the next section, two matrices are combined, so the overall mRNA dynamics is given by the following equation:

$$\mathbf{mRNA} = \mathbf{U}' \cdot \Sigma' \cdot \mathbf{V}'^T = \mathbf{US} \cdot \mathbf{V}'^T, \tag{34.2}$$

where $\mathbf{US}$ ($n \times s$) represents the strength of coupling of each mode $\mathbf{V}'^T$ ($s \times t_k$). The rows of $\mathbf{V}'^T$ can be interpreted as the time course data of key players of the system. However, in general, it represents a linear combination of transcription factor concentrations and activities.

Based on the available data set with transcriptomic data [2] $n = 2040$ genes are considered in the model with $t_k = 18$ time points.

3.2 Network Component Analysis

A cognate approach incorporating detailed biological knowledge was applied and a further model was developed. The model is based on the concept of Network Component Analysis (NCA) [10] that allows a semi-quantitative description of gene expression based on measured transcriptomic data. In brief, the approach is as follows. The number of selected genes is N and the number of selected transcription factors is m. The dynamics of a single gene (i) is described with an ordinary differential equation:

$$\dot{mRNA}_i \;=\; k_i \;\; TF_1^{k_{1i}} \cdot TF_2^{k_{2i}} \cdots \times TF_m^{k_{mi}} \;-\; k_z\, mRNA_i \tag{34.3}$$

with the last term describing the degradation of the mRNA. Parameters k_{ji} are related to the strength of each transcription factor TF_j binding to the respective control sequence: if $k_{ji} > 0$, then the transcription factor is an activator, while $k_{ji} < 0$ points to an inhibition. Assuming that the dynamics of mRNA is faster than protein synthesis, a steady-state assumption holds true and the following equation results after fixing a set point (subscript 0):

$$\frac{mRNA_i}{mRNA_{i0}} \;=\; \left(\frac{TF_1}{TF_{10}}\right)^{k_{1i}} \left(\frac{TF_2}{TF_{20}}\right)^{k_{2i}} \cdots \left(\frac{TF_m}{TF_{m0}}\right)^{k_{mi}}. \tag{34.4}$$

Taking logarithm ($\log_2$) leads to:

$$\log \frac{mRNA_i}{mRNA_{i0}} = k_{1i} \log\left(\frac{TF_1}{TF_{10}}\right) + k_{2i} \log\left(\frac{TF_2}{TF_{20}}\right) + \cdots + k_{mi} \log\left(\frac{TF_m}{TF_{m0}}\right) \tag{34.5}$$

which can be written in matrix form:

$$\mathbf{mRNA} \;=\; \mathbf{K} \cdot \mathbf{TF}\,, \tag{34.6}$$

with $\mathbf{K}$ is $N \times m$ coupling matrix representing the effect of each transcription factor on the respective gene and $\mathbf{TF}$ is a $m \times t_k$ matrix of transcription factor activities (t_k is again the number of available data points). The aim is now to decompose matrix $\mathbf{mRNA}$ to get both K as well as $\mathbf{TF}$. Note that the entries of $\mathbf{K}$ have to be specified before (value 0 if a transcription factor is not involved in the regulation of the gene and 1 as starting value for the algorithm, if a transcription factor is involved) the algorithm starts, that is, the structure of the model has to be given and NCA determines the coupling strength and the time course of transcription factor activities. To solve the problem, the following objective function is minimized:

$$\min \; ||\mathbf{mRNA} \;-\; \mathbf{K} \cdot \mathbf{TF}||^2 \tag{34.7}$$

considering the difference between measured data and model simulation. Further details and the algorithm as MATLAB file can be found in the original paper [10].

The model is based on available transcriptomic data [2], but in contrast to the SVD approach, is reduced to focus on central metabolism. It comprises 50 transcriptional units (75 genes) and $m = 4$ transcription factors (Crp, ArcA, FruR, and GalS). After filtering out genes with no entry in the database (no experimental evidence that the gene is under control of one of the transcription factors) the final model contains $N = 33$ genes, representing the central metabolism. The choice is based on prerequisites of the algorithm and the experimental conditions chosen. So, transcription factor Fnr, related to genes that are involved in oxygen consumption is not considered. Also, several other transcription factors cannot be integrated or are not significant, e.g., considering transcription factor Fis showed that this transcription factor has only marginal influence on the calculations.

4 mRNA–Protein Relationship

An interesting observation is that in many cases the dynamical behavior of mRNA and protein seem not to be correlated. However, for steady-state data, a good agreement was shown [12, 14]. Experimental data, protein, and mRNA time course data are available for example for the LacZ protein from the two data sources. Due to the fact that a dynamical system is analyzed, a modeling approach was also chosen here to correlate mRNA and protein dynamics. As a starting point, a differential equation for the protein is set-up:

$$\dot{P} = k_{\text{syn}}\ \text{TA}\ mRNA\ -\ (\mu\ +\ k_{\text{d}})\ P. \tag{34.8}$$

The equation takes into account, that the rate of protein synthesis is proportional to the amount of mRNA (*mRNA*) available and to the amount of proteins (TA) representing the translation apparatus (parameter k_{syn}) and a degradation term (parameter k_{d}). In translation, several proteins are involved that are also subject to dynamical changes as can be seen from the experimental data (see below). The rate of degradation is proportional to the protein concentration; furthermore dilution by growth (specific growth rate is μ) is considered. The mRNA data is scaled with respect to a chosen value at the beginning of the experiment. This can be done with the protein concentration in the same way. Having $p = P/P_0$, with P_0 is being the set point for the protein, $\text{ta} = \text{TA}/\text{TA}_0$ with TA_0 is the corresponding operation point for the translation apparatus, and the fact that the mRNA concentration can be reconstructed from the measured data $mRNA_M$, the equation from above reads:

$$(\dot{p P_0})\ =\ k_{\text{syn}}\ \text{ta}\ \ \text{TA}_0\ \ mRNA_0\ \ 2^{mRNA_M}\ -\ (\mu + k_{\text{d}})\ p\ P_0. \tag{34.9}$$

For the experimental growth situation considered, growth on glucose takes place in the first growth phase. Here we assume that in the exponential phase the rate of synthesis and degradation for most of the proteins is well balanced and hence, for

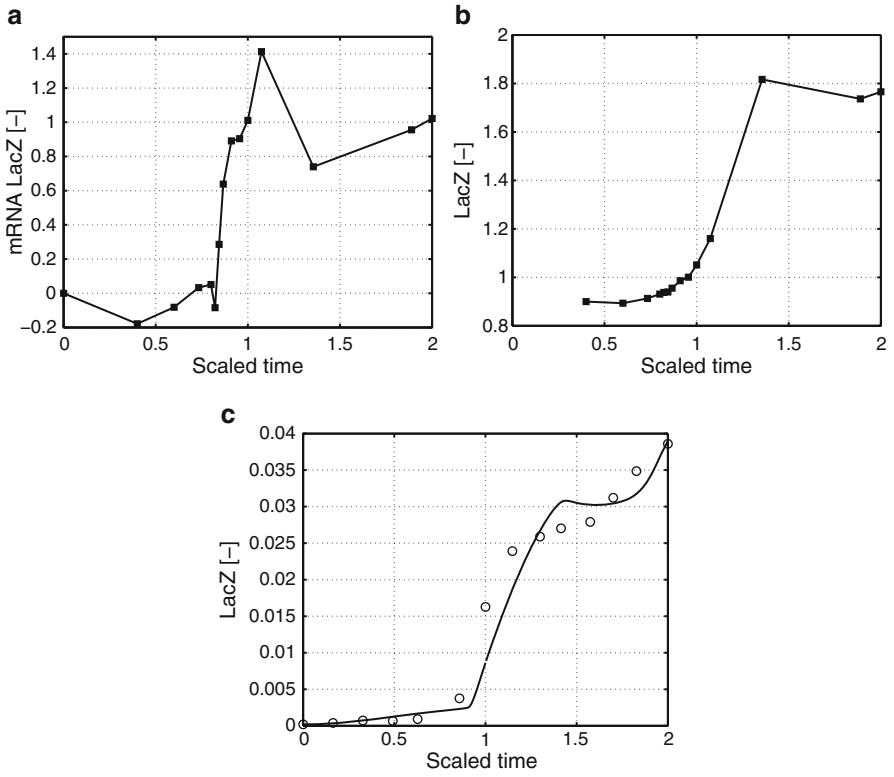


Fig. 34.3 *Left*: mRNA data for LacZ. *Middle*: Corresponding protein time course based on (34.11). *Right*: Measured time course for LacZ (symbols) and simulation data from the detailed dynamic model [1]

the operation point considered (i.e., ta = 1, $p = 1$, and $mRNA_M = 0$), the following equation will hold true:

$$k_{\text{syn}}\ \text{TA}_0\ mRNA_0\ =\ (\mu\ +\ k_{\text{d}})\ P_0. \tag{34.10}$$

Plugging in this equation, the equation for the scaled protein results in:

$$\dot{p}\ =\ \left(\text{ta}\ 2^{mRNA_M}\ -\ p\right)\ (\mu + k_{\text{d}}). \tag{34.11}$$

The equation is based only on a single unknown parameter, k_{d} that allows to adjust the model in such a way that the experimental data are matched. To validate the approach, a comparison with experimental data for LacZ based on the two data sets is shown in Fig. 34.3. Note that a first calculation was performed to determine "ta" (transcription/translation apparatus) and that this results were used for the example here. Comparing the results from the proposed approach (middle plot) with

the experimental and simulation data [1] (right plot) reveals that both time course data for the LacZ protein show qualitatively a sufficient agreement. Note that the simulation in the middle plot is based only on the few data points in the left plot, while the dynamic simulation in the right plot is continuous.

5 Transcription Factors Activities During Growth on Glucose and Lactose

The modeling approaches introduced so far are applied to growth of *E. coli* on glucose and lactose. This is a very classical experiment to demonstrate genetic control and is discussed very frequently in the literature. To compare transcription factor activities, two data sets were used. In [1] 18 experiments are used to calibrate the comprehensive model introduced in Sect. 2. In [2] an experiment with the focus on transcriptomic data during the same environmental condition was introduced. To compare the data sets, a time scale normalization was performed. This was necessary since different initial values for glucose and lactose were used. This leads to the observation that glucose runs out to different time points. So, the experimental data was subdivided into two phases: Phase 1 corresponds to growth on glucose while phase 2 growth on lactose. As was shown already above, the two time courses for biomass show a good agreement.

5.1 Results for the Singular Value Decomposition

For the complete data set in [2] a SVD composition was performed. The data set contains 2040 genes and 18 time points. Figure 34.4 shows the dynamics of the first two modes that are the first two columns of matrix **V** (see above). The SVD reveals a characteristic mode on a short time scale (1–1.5 h) and a characteristic mode on a longer time scale (4 h). Sequential growth of *E. coli* on the two substrates, as can be seen in the figure, leads to a short time period where the organism must adapt to the altered situation. As discussed in [13], the stringent response, mediated by the alarmone guanosine 3',5'-bispyrophosphate (ppGpp) coordinates gene expression in the transition time from growth on glucose to growth on lactose. Mode 1 from the SVD approach corresponds very well to this time frame (compare Fig. 1 in [13]; it shows expression data of genes that were under control of ppGpp. Direct measurement for ppGpp for this type of experiment is not available, however, for an experiment with glucose and succinate as substrates, ppGpp was measured [4]. The time course of ppGpp during this experiment reveals a fast time constant and the system reaches a new steady state after 30 min). It is the dominant mode, that is, the dynamics of all 2040 genes can be explained in large part by this mode. Mode 2 shows a short decrease when glucose is running out, then quickly increases

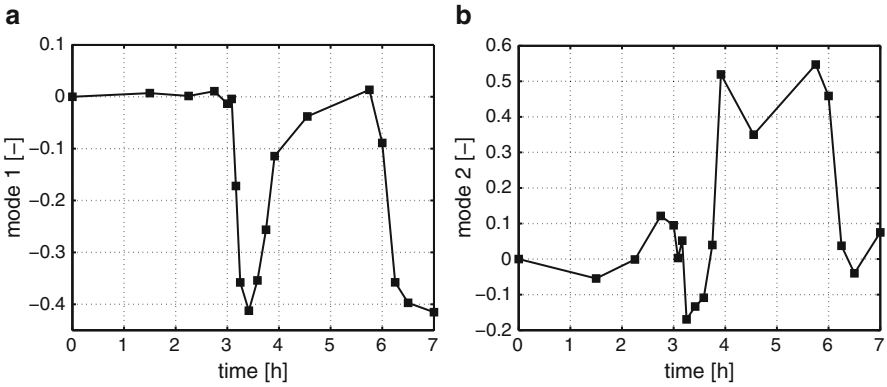


Fig. 34.4 The first two characteristic modes from SVD. *Left*: Mode 1 with a short time scale; *right*: Mode 2 on a longer time scale. Time is not scaled here

and is present during the overall time period until lactose is running out. This mode reflects growth on lactose where many genes are adjusted to meet the cellular requirements.

5.2 Transcription Factors Connectivities and Activities

NCA was applied several times for different systems. However, a problem that is seldomly addressed in the literature is the correct sign of the entries in matrix K (coupling strength). The signs should be in agreement with data base entries, e.g., Ecocyc.

In the model, four transcriptions factors were considered: Crp, ArcA, FruR, and GalS. Figure 34.5 shows the elements of matrix K for every transcription unit in the model. In the same plots, the entry of the correct sign according to Ecocyc is also given as gray bar. Note, that according to Fig. 34.2 the sign for repressor FruR has to be inverted. Comparing the database entries for positive or negative regulation, the following results are obtained. The entries are correct for the cAMP·Crp regulated genes in 82.35%; for ArcA in 100%, for FruR for 70%, and for GalS in 80% (note that if a calculated value is zero or below 0.05, the gene is not counted in the statistics, e.g., *pfl*, *cydAB*, and *rpoS* that are under control of ArcA).

Figure 34.6 left shows the simulation results for the complex cAMP·Crp for the bottom-up model [1]. On the right side the corresponding result for NCA is shown. The second line of plots shows the dynamics of fructose-1,6-bisphosphate·FruR in both approaches. As can be seen, transcription factor activities from the bottom-up approach and from the top-down approach agree qualitatively well.

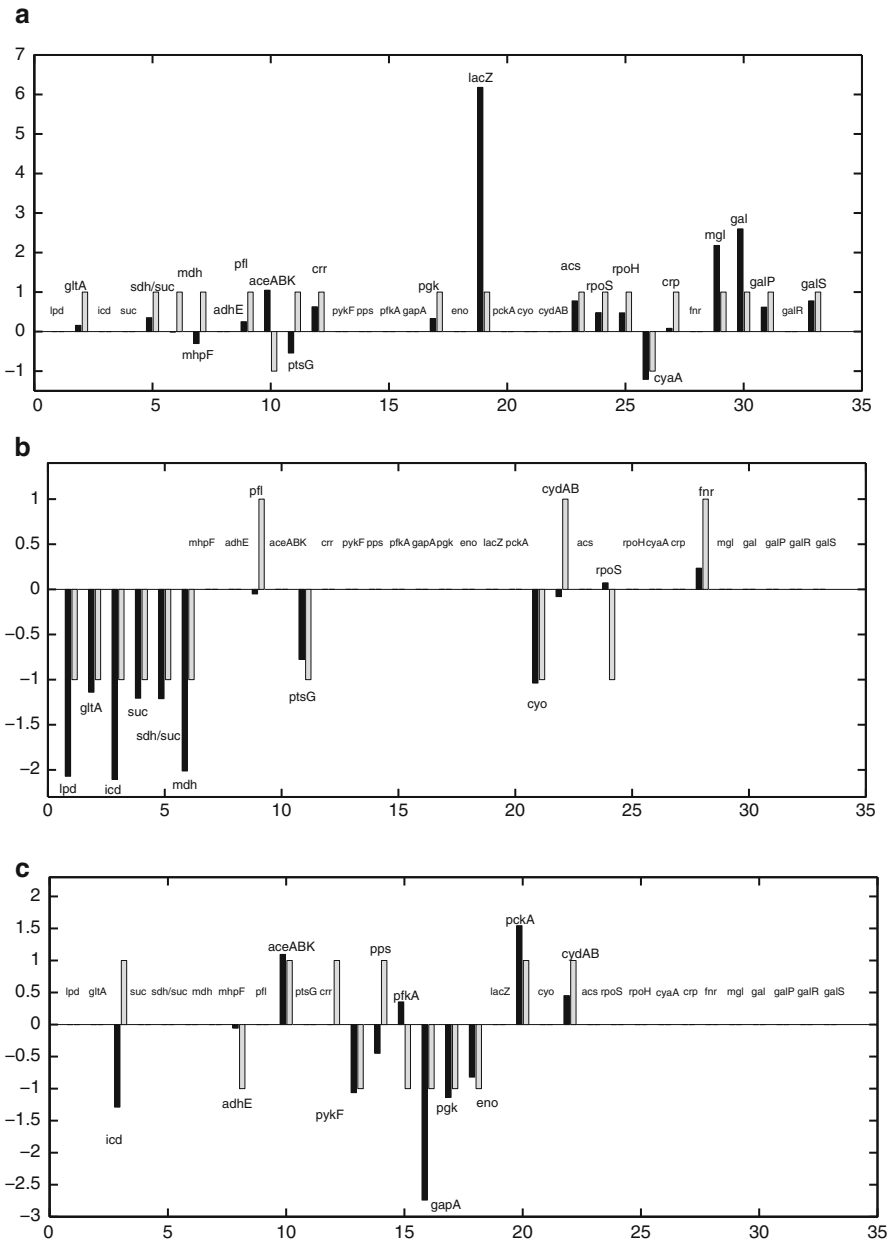


Fig. 34.5 Transcription factor connectivity matrix. From top: cAMP·Crp, ArcA, FruR, and GalS. *Black bars* indicate the value of the connectivity while *gray bars* indicate the correct sign based on data base research

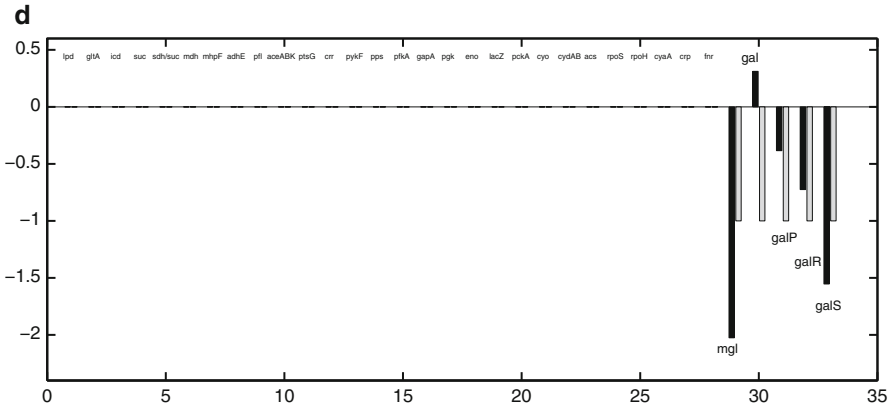


Fig. 34.5 (continued)

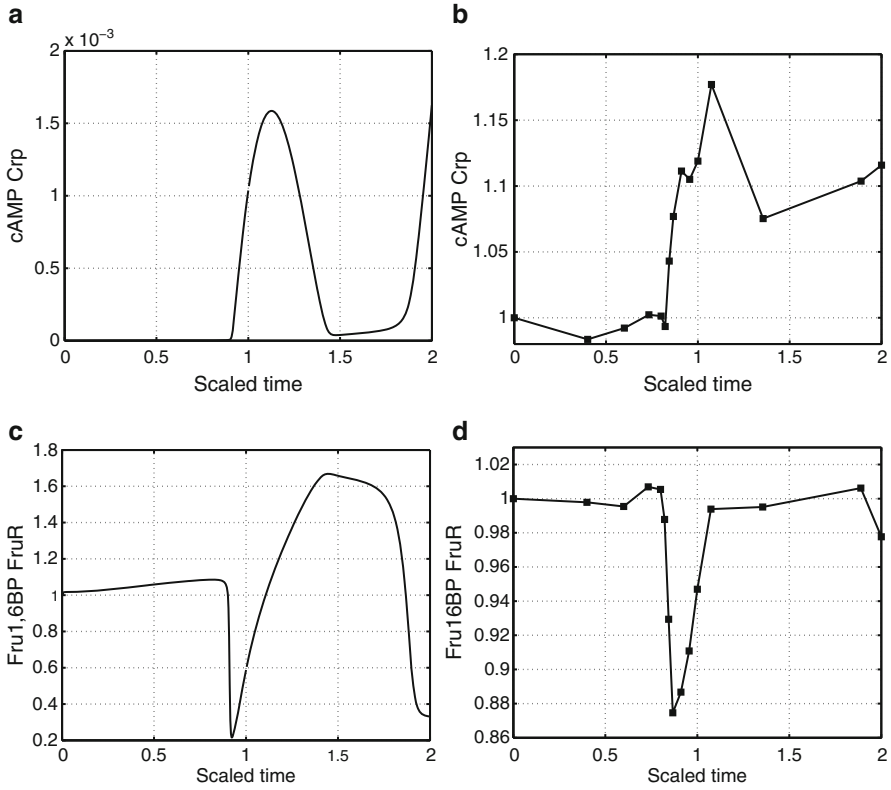


Fig. 34.6 Comparison of transcription factor activities. *Left*: Simulation results with the bottom-up model. *Right*: Results with the NCA model. First line: transcription factor cAMP·Crp; second line transcription factor Fru 1,6-Bisphosphate·FruR. Note that on the left side the unit for the intracellular component is μ mol/g DW while on the right side the activities are dimensionless

6 Stress Response

Running out of the major carbon source glucose leads to a sudden shut-down of the genes for the translation apparatus and an immediate recovery at the beginning of growth on lactose. This cellular response is called stringent response and is a part of the stress response of the cell. The stringent response is monitored with the time course of nine genes involved in transcription and translation. The choice of genes is based on the selection in [2, Fig. 4]; note, that the choice mainly focuses on genes of the 50S ribosomal subunit and the 30S ribosomal subunit. To determine the effects on the protein level, the approach shown above was applied. Using (34.11) leads to the dynamics shown in Fig. 34.7. Interestingly, the fast dynamics that can be seen on the mRNA level is not reflected on the calculated protein level. Based on the fact that both, synthesis and degradation (dilution by growth) decrease after the run off of glucose, the protein level remains nearly unchanged during the whole experiment.

A further stress response is the activation of the σ^S-factor. This sigma factor is under control of cAMP·Crp and it is well known that also translation and protein degradation are regulated (not included in the model). Figure 34.8 shows the fast synthesis of the mRNA of σ^S when glucose is running out. However, since the specific growth rate at the end of the glucose phase (and before the run out) is still high, the protein level decreases – due to dilution – indicating a non-stress situation. Quickly after the run out of glucose, the level of σ^S increases and finally decreases in the middle of the second growth phase.

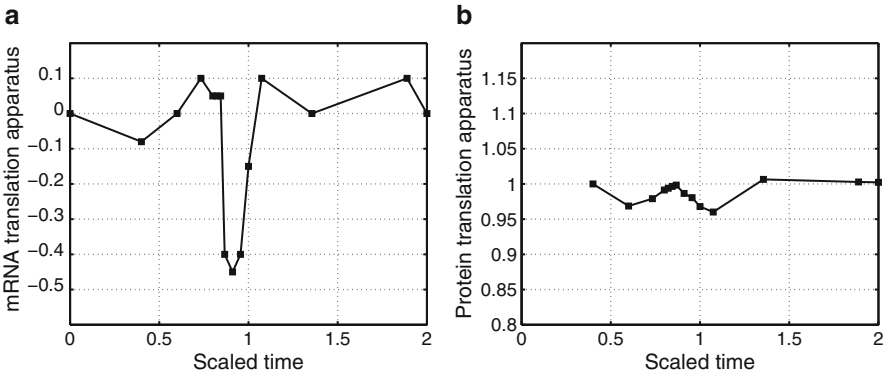


Fig. 34.7 Dynamics of the translation apparatus (*left*: mean value of nine genes (*rpsP*, *rpsB*, *rplM*, *rpsM*, *rplN*, *rpsJ*, *rpsL*, *rplU*, *rpsF*) from [2, Fig. 4] and therein; *right*: dynamics of the protein)

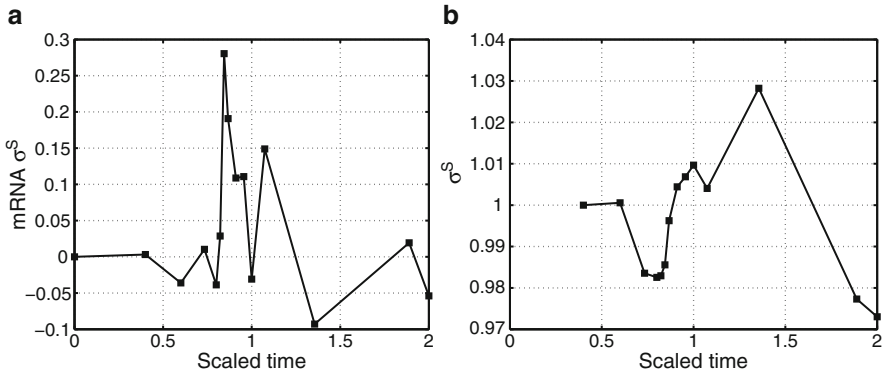


Fig. 34.8 Dynamics of the σ^S-factor. *Left*: mRNA from [2] and *right*: protein

7 Product Secretion During Growth on Lactose

Growth on lactose leads to a lower specific growth rate than growth on glucose. In [3] it is discussed that maybe ethanol is produced in large amounts during growth on lactose (Fig. 4 therein) while in [1] it is observed that galactose is produced. In the following, (34.11) is used to monitor the dynamics of the galactose transporters, the galactose operon, the galactose regulators, and the enzymes responsible for by-product synthesis from glycolysis (lactate, ethanol, and acetate).

Figure 34.9 shows the calculated time course data for the two galactose regulators GalS and GalR, the proteins of the galactose operon *galETKM*, and the galactose transport systems Mgl (three genes) and GalP.

From database entries it is known that GalS is under control of cAMP·Crp but not GalR. This can be seen in the time course on the left side: GalS increases in a small amount while GalR decreases based also on the negative control from GalS.

Figure 34.10 finally shows the calculated time course data for the proteins involved in by-products synthesis. The enzyme involved in ethanol synthesis shows an increase while the enzyme involved in lactate synthesis decreases. In comparison with the galactose pathway, the induction of the ethanol pathway as predicted in [3] is rather marginal. The observation that acetate is produced during a high specific growth rate of *E. coli* on glucose and other carbon sources is documented in several publications (e.g., [14] and references therein). Very recently it was shown that both pathways involved in acetate synthesis and degradation (Pta, Ack, and Acs, respectively) are repressed during growth on glucose, however, the strength of repression is different leading to the excretion of acetate for high growth rates [14]: for high specific growth rates the *acs* operon is severely repressed while *pta* and *ackA* are repressed moderately. Here, we confirm the results while monitoring the protein levels during higher values of cAMP·Crp. As can be seen in the figure, the Acs protein increases in the second growth phase reflecting a strong repression

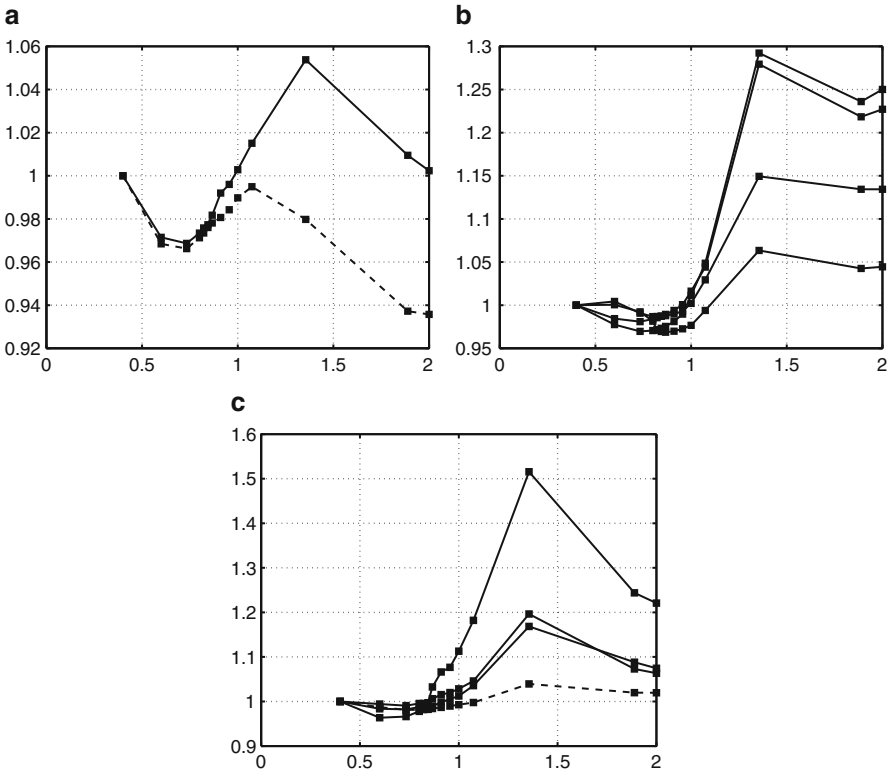


Fig. 34.9 Time course for enzymes involved in galactose transport and degradation. *Left*: GalS and GalR (dashed). *Middle*: GalETKM. *Right*: MglABC and GalP (dashed). Protein concentrations are calculated as described above and are based on the mRNA data from [2]

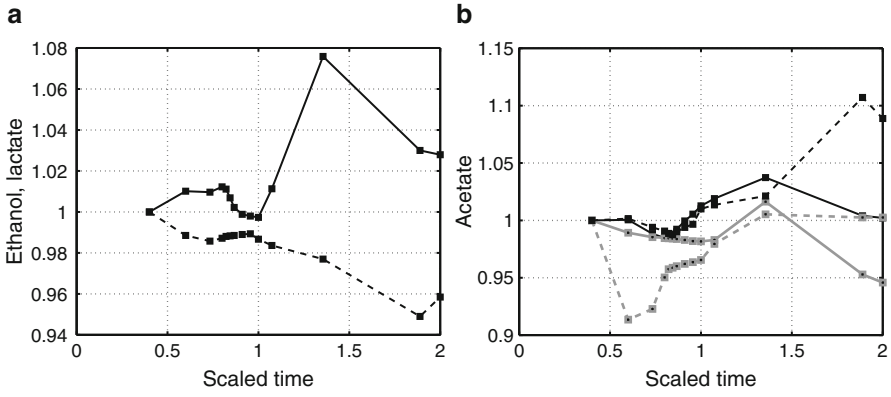


Fig. 34.10 Time course for enzymes involved in product formation. *Left*: Ethanol (AdhE) and lactate (Ldh, dashed). *Right*: Acetate (Acs (black) and PoxB (black dashed); Pta (gray) and AckA (gray, dashed)). Protein concentrations are calculated as described above and are based on the mRNA data from [2]

in first growth phase. In contrast, Pta and AckA show a very different behavior. Both protein levels drop during the first hour and resume afterwards. Surprisingly, gene *poxB* that is known to be not repressed by cAMP·Crp is also induced during growth on lactose. PoxB also synthesizes acetate from acetyl-CoA.

8 Discussion

Mathematical modeling is a powerful tool in systems biology to reproduce experimental data based on the knowledge of the underlying biochemical network and to formulate and test hypotheses. Here, mathematical modeling is applied to show that models based on the bottom-up and the top-down approach lead to a consistent behavior with respect to the dynamics of state variables of the system. In bottom-up models, state variables are chosen in such a way that the network of interest is adequately described. In general, metabolites, mRNA, or proteins are chosen as state variables and balance equations are set up to describe in which way the state variables change over time. Thereby, reaction rates are formulated that describe either a metabolic flux or activation/inhibition in signaling networks. A main drawback of this approach is the high number of unknown kinetic parameters needed for the reaction kinetics. Models based on a top-down approach integrate experimental data based on "omics" technologies (transcriptome, proteome, etc.) together with biological knowledge. Several approaches are described – either to find all significant interconnections between two nodes in the network or to explain the behavior of nodes based on additional variables. The latter approach is used in SVD of time course data and also in NCA. In NCA additional biological knowledge is incorporated in form of basic information on the network structure (gene is controlled by a transcription factor or not). All three approaches are related by the observation that they produce the dynamics of transcription factor activities: the bottom-up network includes metabolites in central metabolism that are starting points of signaling pathways that end with the activation/deactivation of regulators; in the SVD approach, characteristic modes are calculated that represent the overall dynamics of the system and in NCA selected transcription factors are used to describe the dynamics of the respective genes. For a classical example from microbiology, diauxic growth on glucose and lactose, all approaches are used to calculate the respective time course data. Comprehensive experimental time course data from two sources is used to calibrate the dynamical bottom-up model and to calculate the coupling strength of the transcription factors for the respective genes.

Using SVD the two first singular values and the corresponding entries in the U and the V matrices are analyzed. The first singular value (largest absolute value) corresponds to a initial fast response of the cell named stringent response: genes that are involved in amino acid synthesis or the transcription and translation apparatus are shut down and resume after approximately one hour. A very similar behavior is observed when looking on the time course of FruR with the NCA approach. Effector

Fructose-1,6-bisphosphate of FruR in this way is a candidate to transduce the signal (running out of glucose) to the components of the stringent control circuit. The second singular value reflects the dynamics on a longer time scale and describes the adaption of the organism to the new condition. A comparable dynamics is seen in the time course of cAMP·Crp in the NCA approach. Starting point for this signaling pathway is the ratio of two metabolites, PEP, and pyruvate. However, before activating transcription factor Crp a further element with own dynamics, namely, the synthesis of cAMP comes into play. The synthesis and excretion of cAMP shows dynamics on a longer time scale and therefore characterises the transient behavior of the cells from growth on glucose to growth on lactose.

NCA was applied in several cases, however, the entries of matrix K are not discussed at all. Here, it is shown that the sign of the entries in matrix K show a good agreement with the entries in data bases. For transcription factor FruR the error of 30% might be due to the fact that other regulators contribute to transcription of the respective genes.

The stringent response leads to an immediate shut down of transcription of a number genes involved in biosynthesis. When looking at the transcription and translation apparatus, the shut down of transcription is not seen on the protein level. This is due to the fact that both the rate of synthesis and the specific growth rate change with time. This leads to a very well balanced behavior of the proteins.

During growth on lactose, galactose is produced in large amounts as observed in [1]. This is also reflected in the microarray data in [2] where the genes for galactose uptake and metabolism are induced. Acetate secretion is observed during growth on glucose and lactose. Monitoring the proteins confirms a recent result described in [14] that the *acs* gene is repressed much more than *ackA* and *pta*. However, a further pathway via *poxB* shows a higher value in the second growth phase on lactose. This corresponds well with an increase of the specific acetate production on lactose as observed in [1]: in the glucose phase the yield coefficient is 0.35 g/g while in lactose growth phase it increases to 0.45 g/g (data calculated from material provided in the supplement).

In the recent literature, bottom-up and top-down modeling approaches are used "stand alone" in many applications. Here, evidence is provided that for systems very well understood from a biological point of view, both approaches lead to comparable simulation results of unmeasured state variables, here transcription factor activities. This gives hope that in the future, both approaches can benefit from each other in a more intensive way than nowadays. A possible way to do this is to quantitatively map the transcription factor activities from the bottom-up approach to the top-down approach (e.g., using empirical functions). So, the bottom-up model can be used to simulate the expression of the genes from the top-down model under different conditions than presented here.

Acknowledgments Funding in part by the German BMBF during the FORSYS initiative is acknowledged.

References

1. Bettenbrock K, Fischer A, Kremling K, Sauter JT, Gilles ED (2006) A quantitative approach to catabolite repression in *Escherichia coli*. J Biol Chem, 281:2578–2584
2. Chang DE, Smalley DJ, Conway T (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. Mol Microbiol 45(2):289–306
3. Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics 24(18):2044 –2050
4. Harshman RB, Yamazaki H (1971) Formation of ppGpp in a relaxed and stringent strain of *Escherichia coli* during diauxie lag. Biochemistry 10(21):3980–3982
5. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 97(15):8409 –8414
6. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucl Acids Res 37(Database):D464–D470
7. Kremling A, Bettenbrock K, Gilles ED (2007) Analysis of global control of *Escherichia coli* carbohydrate uptake. BMC Systems Biology 1:42
8. Kremling A, Bettenbrock K, Gilles ED (2008) A feed-forward loop guarantees robust behavior in *Escherichia coli* carbohydrate uptake. Bioinformatics 24:704–710
9. Kremling A, Kremling S, Bettenbrock (2009) A comparison of modeling approaches. *FEBS Journal* 276:594–602
10. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP (2003) Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci USA 100(26):15522 –15527
11. Nishio Y, Usada Y, Matsui K, Kurata H (2008) Computer-aided rational design of the phosphotransferase system for enhanced glucose uptake in *Escherichia coli*. Mol Sys Biol 4:160
12. Shimizu K (2004) Metabolic flux analysis based on 13C-labeling experiments and integration of the information with gene and protein expression patterns. Adv Biochem Eng/Biotechnol 91:1–49 PMID: 15453191
13. Traxler MF, Chang DE, Conway T (2006) Guanosine 3',5'-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. Proc Natl Acad Sci USA 103(7):2374 –2379
14. Valgepea K, Adamberg K, Nahku R, Lahtvee PJ, Arike L, Vilu R (2010) Systems biology approach reveals that overflow metabolism of acetate in *Escherichia coli* is triggered by carbon catabolite repression of acetyl-CoA synthetase. BMC Syst Biol 4(1):166

Chapter 35
A Differential Equation Model to Investigate the Dynamics of the Bovine Estrous Cycle

H.M.T. Boer, C. Stötzel, S. Röblitz, and H. Woelders

Abstract To investigate physiological factors affecting fertility of dairy cows, we developed a mechanistic mathematical model of the dynamics of the bovine estrous cycle. The model consists of 12 (delay) differential equations and 54 parameters. It simulates follicle and corpus luteum development and the periodic changes in hormones levels that regulate these processes. The model can be used to determine the level of control exerted by various system components on the functioning of the system. As an example, it was investigated which mechanisms could be candidates for regulation of the number of waves of follicle development per cycle. Important issues in model building and validation of our model were parameter identification, sensitivity analysis, stability, and prediction of model behavior in different scenarios.

H.M.T. Boer (✉)
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands

Adaptation Physiology Group, Department of Animal Sciences, Wageningen University, Wageningen, The Netherlands
e-mail: Marike.Boer@wur.nl

C. Stötzel • S. Röblitz
Department of Numerical Analysis and Modeling, Computational Systems Biology Group, Zuse Institute Berlin (ZIB), Berlin, Germany
e-mail: stoetzel@zib.de; susanna.roeblitz@zib.de

H. Woelders
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands
e-mail: Henri.Woelders@wur.nl

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_35,

1 Fertility in Dairy Cows

Bovine fertility is the subject of extensive research in animal sciences, especially because fertility of dairy cows has declined during the last decades. Subfertility has negative implications for dairy farm profitability, sustainability of animal production, and animal welfare, as it takes more time and effort to get cows to be pregnant. The decline in fertility has coincided with selection for a higher milk yield, and is manifested in alterations in hormone patterns, reduced expression of estrous behavior, and lower conception rates. However, it is unknown if and how high milk yield and subfertility are causally related. Systems biology approaches, including the use of mathematical models, can help to increase our understanding of the complex interplay of factors involved in the reproductive cycle. Such models can be very valuable in studying effects of, e.g., stress or disease on reproduction [1].

The bovine estrous cycle is the hormonally controlled recurrent period when the cow is preparing for reproduction by producing a fertilizable oocyte. The main tissues and organs involved in the regulation of the estrous cycle are the ovaries, the uterus, the hypothalamus, and the anterior pituitary. These organs interact via hormones in the blood. A normal cycle includes two or three wave-like patterns of follicle development, in which a cohort of follicles starts to grow. The length of the estrous cycle is often taken to be approximately 21 days, but the cycle length may be shorter in two-wave cycles than in three-wave cycles. The first one or two waves produce a dominant follicle that does not ovulate, but undergoes regression under influence of P4 (see abbreviation key in the caption of Fig. 35.1). The dominant follicle in the last wave produces increasing amounts of E2, triggering the surge of LH, which induces ovulation. Once an oocyte is successfully ovulated, the remains of the follicle form a new P4-producing CL.

In this chapter, we briefly describe the development of a mathematical model of the bovine estrous cycle, we discuss how such a model could be validated, and we show an example of how the model can be used to investigate patterns of follicle development. The model summarizes physiological knowledge and empirical data, and thereby provides insight in the regulatory structure of the system.

2 Modeling the Bovine Estrous Cycle

The endocrine and physiologic regulation of the bovine estrous cycle has been studied extensively. For some specific mechanisms or parts of the system mathematical models have been developed (reviewed in [2]), but mostly these models were of limited scope and do not contain all the major tissues and hormones necessary for simulation of the dynamics of follicle development over consecutive cycles. From the mathematical point of view, many biological processes, such as hormonal interactions, can be modeled with the help of differential equations, which describe the rates of change of the involved substances over time. We developed a

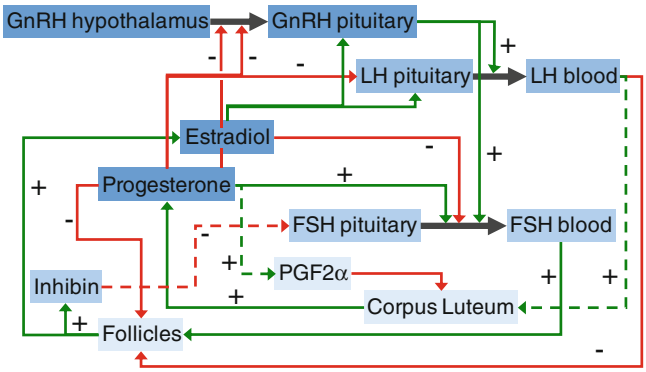


Fig. 35.1 Key components of the biological system and their interactions. "+" and "−": inhibiting and stimulating effects respectively. *Dashed lines*: time delay. *Foll* follicular function (in the model representing the combined capacity of all follicles present at any time to produce E2 and Inh), *CL* corpus luteum (in the model representing the capacity of the CL to produce P4, rather than the physical size of the CL), *P4* progesterone, *E2* estradiol, *Inh* inhibin, *GnRH* gonadotropin releasing hormone, *FSH* follicle stimulating hormone, *LH* luteinizing hormone, *PGF2α* prostaglandin F2α

mathematical model of the dynamics of the bovine estrous cycle on individual cow level that is able to simulate follicle and CL development and the periodic changes in hormones levels that control these processes by a set of linked differential equations. We performed an extensive literature research on how the individual components of the cycle function together, obtained abstraction levels that display the most important mechanisms, and constructed a flow chart of their interactions. The key components of the biological system and their interactions incorporated in the model are shown in Fig. 35.1.

We derived a differential equation for each of the components mentioned in Fig. 35.1. This initial model contains 12 ordinary and delay differential equations and 54 parameters [3] and is partly based on previous work by Selgrade and colleagues [4] and Reinecke and Deuflhard [5] on modeling the human menstrual cycle. Hill functions are used to model the non-linear stimulating and inhibiting effects of hormones. In the model, the amount of GnRH in the hypothalamus is a result of synthesis in the hypothalamus and release into the pituitary and is affected by P4 and E2. FSH is synthesized in the pituitary when the level of Inh is low. FSH release is stimulated by GnRH and inhibited by E2. LH synthesis in the pituitary is stimulated by E2 and inhibited by P4, and LH release is stimulated by GnRH. Follicle development is stimulated by FSH and inhibited by P4 and the LH surge. The production of P4 is proportional to CL function. PGF2α induces CL regression and is stimulated by P4 with a time delay. The production of E2 and Inh is proportional to follicular function. Simulation results (Fig. 35.2) show that a set of equations and parameters was obtained that describes the system consistent with empirical knowledge. Even though the majority of the mechanisms included in the model are based on relations that in literature have only been described qualitatively, the model output is surprisingly well in line with empirical data.

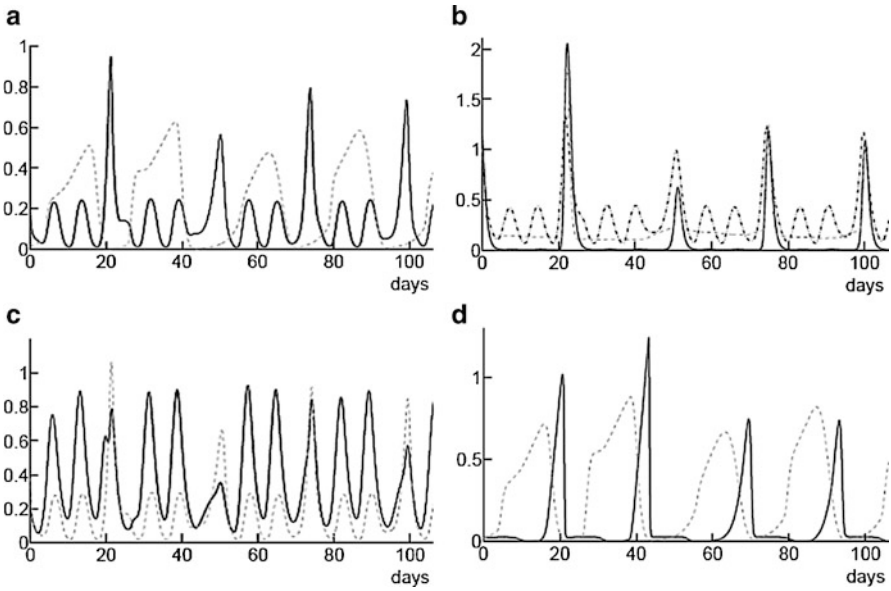


Fig. 35.2 Model parameterization generating estrous cycles of approximately 21 days, with three peaks of FSH and three corresponding waves of follicular growth. The third wave of follicular growth takes place when P4 levels are low, which results in increasing levels of E2. This causes an LH surge, which then triggers ovulation. (**a**) Foll (*solid line*) and CL (*dashed line*). (**b**) GnRH (*solid line*), LH (*dashed line*), and E2 (*dashed–dotted line*). (**c**) FSH (*solid line*) and Inh (*dashed line*). (**d**) PGF2α (*solid line*) and P4 (*dashed line*). The equations are expressed on a relative scale in order to simplify parameter estimation, and therefore the y-axis of the figures is dimensionless

3 Model Validation

There is no general procedure for model validation. The most important aspect is whether certain model simulation outcomes match with some given experimental data. Model validation therein aims to assess the predictive accuracy of the numerical model, and thereby to build confidence in the model. A model has an added value when it not only matches given data but also gives insight into certain processes that cannot be observed by measurements, and thus hints to explanations for certain phenomena. Here we discuss four steps that we consider to be important for the model building and validation of our specific model of the bovine estrous cycle: parameter identification, sensitivity analysis, stability, and prediction of model behavior in different scenarios.

Our model describes the interactions between key components of the bovine estrous cycle. For solving the system of differential equations, the solver RADAR5 [6] developed for the solution of stiff delay differential equations was used. The main difficulty lies in the identification of the involved parameters. Most parameter values in the model are neither measurable nor available in literature, and sometimes

even the range of values is completely unknown. For a model of a complex system with various components functioning together, this leads to a large number of differential equations and unknown parameters. Under these circumstances, estimating all parameters simultaneously is impossible. For our model, we used a model decomposition approach to obtain a good initial guess of the parameter values for the optimization procedure. The model was decomposed into disjoint model parts, and parts of the model were temporarily replaced by input curves based on published data of hormone profiles of cows with a normal estrous cycle. A first subset of parameters was then estimated, and step by step the output functions for the other model parts were fitted, until finally a closed network was obtained [7]. Parameters were estimated with software developed at the Zuse Institute (NLSCON). This software uses subtle mathematical techniques such as affine covariant Gauss–Newton methods that take into account sensitivities and linear dependencies of the parameters [8].

A sensitivity analysis for the complete set of model parameters has been performed with techniques described in [8]. A higher sensitivity means that a change in the value of the parameter has a larger effect on the model solution. Sensitivity analysis can, therefore, identify the parts that need a more precise parameter estimation. It is an important step not only in the parameter estimation algorithm but also in model validation, since it quantifies the relative importance of parameters. Thereby, it shows if the model does not depend unexpectedly strong on biologically less relevant parameters. The sensitivity analysis of our model confirmed that parameters that are very important for follicle development and cycle length had a high impact on the model solution.

Model validation also deals with the question of stability. Stability investigates how changes in the model input affect model output. In a stable model, small perturbations should not disturb the qualitative behavior of the system. As can also be observed in Figs. 35.2 and 35.3, some parameterizations of our model produce a stable limit cycle (periodic behavior), while others generate consecutive estrous cycles that are not entirely identical (quasi-periodic behavior). The variations between simulated cycles are thus not an intrinsic characteristic of the model, but depend on the parameterization. In the bovine, a new population of follicles is recruited in each cycle, with a different number and size, leading to differences in the hormonal profiles that are the result. We, therefore, think the variation between estrous cycles is not only due to changes in external factors for that cow but also arises from the fact that each cycle presents slightly new and somewhat different "starting values" for the next cycle, which we think that our model can mimic. Stability of the model is also an essential requirement to handle variation between individuals. With one single model, we aim at finding parameterizations for individual measurement data. This could be done by defining input functions of individual time series, but also by simulating external influences like effects of nutrition or stress. However, experimental data available in the literature often do not meet the requirements for these individual parameterizations, because either the time scale of investigation is too short or the data lack information of certain experimental parameters.

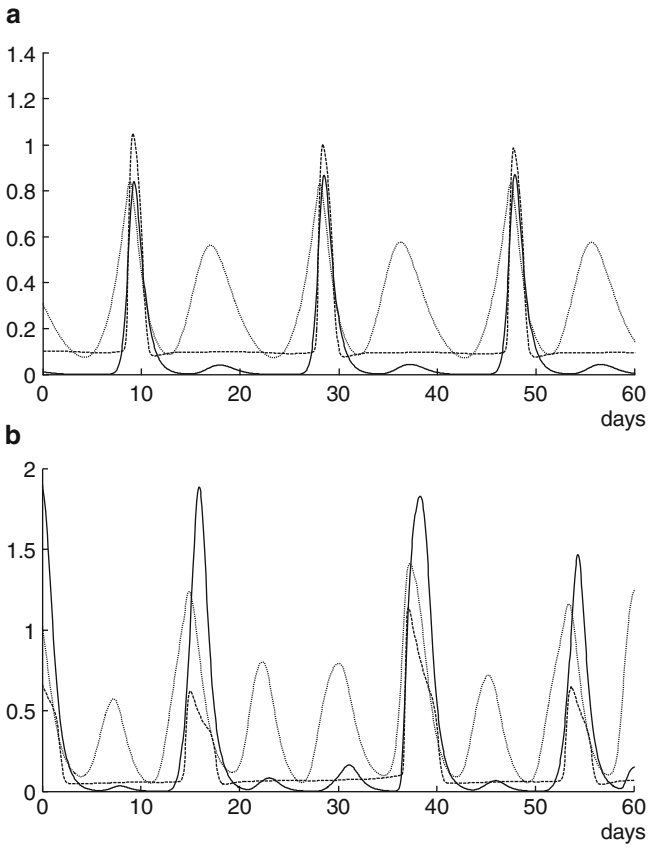


Fig. 35.3 A change in specific parameter values can result in a series of two-wave cycles (**a**) or alternating three- and two-wave cycles (**b**). E2 (*dotted line*), GnRH (*solid line*), and LH (*dashed line*). This figure was obtained by decreasing the parameter that represents the maximum inhibiting effect of P4 on follicular function

Apart from fitting to individual data, the model could be used to determine the level of control exerted by various system components on the functioning of the system. This could be done by changing the value of specific parameters, aiming to obtain a certain model output, or by mimicking, e.g., external hormone administration. Experimental data to verify the predicted causes of certain phenomena are not always available, but the simulation could provide some likely candidates involved in the regulation of certain mechanisms that could be tested in further experiments. Further, the model can serve as a basis for more elaborate models and simulations, with the ability to study effects of external manipulations and genetic differences. Summarizing, there are many possible model applications, and therefore we should think carefully about what we want to investigate and which parts of the model need, therefore, to be validated.

4 Using the Model to Investigate Patterns of Follicle Development

The model was initially parameterized to generate three waves of follicle development per cycle. One model application that has already been performed was to investigate which mechanisms could be likely candidates for regulation of the number of waves in the bovine estrous cycle. This specific research question allowed us to predict the temporal behavior of the system, to optimize parameters, and to study the sensitivity of dynamical processes with respect to its initial parameter values. A normal bovine estrous cycle contains two or three waves in which a cohort of follicles starts to grow. However, the reason for cycles being of the two or three wave types is unclear. Some studies report better fertility in three-wave cycles compared to two-wave cycles [9], and it has been suggested that the older and larger ovulatory follicles in cycles with two waves contain oocytes of less quality than cycles with three waves [10]. However, other studies showed no difference [11]. A better understanding of endocrine mechanisms regulating follicle development is important to obtain more precise control of the estrous cycle, which can help improvement of pregnancy rates. In the bovine, the follicle that is dominant at the moment of CL regression develops to become the ovulatory follicle. We assumed that there may be two mechanisms by which the follicle-wave pattern can be influenced. One is the rate of follicle growth and the other is the time point of CL regression. In our model, follicle growth is stimulated by FSH and inhibited by P4. Therefore, the first mechanism might be induced by changing the effect of FSH or P4 on follicle growth, or by changing FSH or P4 synthesis. The second mechanism, i.e., the time point of CL regression, is expected to have an effect on the follicular-wave pattern because two-wave cycles can occur when the CL starts to regress at an earlier time point, e.g., because of an earlier increase of PGF2α. We have selected ten parameters in our model that relate to these two overall mechanisms, and we have tested whether changing the value of these parameters affects the number of waves per cycle in the model simulations. For this purpose, the model was extended with an extra equation, which is described in detail in [12]. In brief, the fixed time delays for the effect of the increase in P4 levels on PGF2α release (which limited the predictive ability for this part of the model) were replaced by a mechanism in which the ability to synthesize PGF2α develops over time under the influence of P4. PGF2α levels now rise because P4 stimulates the production of enzymes and receptors required for PGF2α production, which was previously included as a "black box" by using large delays.

Simulation results showed that a change in the value of specific parameters involved in the regulation of follicle growth rate or the time point of CL regression can change the number of waves in a cycle (Fig. 35.3). Of the ten parameters tested, six affected the number of waves per cycle. Like in real cows, the period of oscillations (cycle length) appeared to be variable. Cycles with two waves had a shorter cycle length. In non-ovulatory waves of two-wave cycles, FSH levels were higher, Foll (follicular capacity to produce E2 and Inh) was larger, and therefore

also E2 and Inh levels were higher compared to non-ovulatory waves of three-wave cycles. The two-wave cycles obtained by a change in follicle growth rate were due to a later emergence of the second wave, while the two-wave cycles obtained by a change in time point of CL regression were caused by a shorter CL life span.

The simulation results thus showed that several components of our model of the bovine estrous cycle can affect the pattern of follicle growth, and some of them are plausible biological mechanisms that could explain these patterns. The model appeared to be sufficiently stable when simulation of two-wave cycles was performed. A reason of poor reproductive performance could be suboptimal matching of follicle growth rate and the time point of CL regression. An earlier time point of CL regression (and therefore a shorter cycle) induces a switch from three to two waves, because when P4 levels are sufficiently decreased at the second wave, this will become the ovulatory wave. Although in the bovine, two-wave cycles are on average shorter than three-wave cycles, the difference is not the duration of a complete wave. Based on reported differences in follicle development, we think that differences in number of waves in natural estrous cycles may rather be due to changes in the mechanisms regulating follicle growth rate, and that the shorter cycle length is rather the result than the cause of the change in wave pattern.

In conclusion, this mathematical model can provide plausible pathways of interactions of follicular and endocrine dynamics that contribute to bovine fertility. Our aim is not to develop a model as simple as possible, but a model that, although with a high level of abstraction, includes all the main processes that are considered important from a physiologic point of view, in order to obtain a model that improves insight in these processes.

References

1. Boer HMT, Veerkamp RF, Beerda B, Woelders H (2010) Estrous behavior in dairy cows: identification of underlying mechanisms and gene functions. Animal 4(3):446–453
2. Vetharaniam I, Peterson AJ, McNatty KP, Soboleva TK (2010) Modelling female reproductive function in farm animals. Animal Rep Sci 122:164–173
3. Boer HMT, Stötzel C, Röblitz S, Deuflhard P, Veerkamp RF, Woelders H (2011) A simple mathematical model of the bovine estrous cycle: follicle development and endocrine interactions. J Theor Biol 278(1):20–31
4. Selgrade JF, Schlosser PM (1999) A model for the production of ovarian hormones during the menstrual cycle. Fields Inst Commun 21:429–446
5. Reinecke I, Deuflhard P (2007) A complex mathematical model of the human menstrual cycle. J Theor Biol 247:303–330
6. Guglielmi N, Hairer E (2005) RADAR5. /http://www.unige.ch/hairer/software.htmlS. Accessed 7 Sept 2009
7. Boer HMT, Stötzel C, Röblitz S, Deuflhard P, Veerkamp RF, Woelders H (2010) A simple mathematical model of the bovine estrous cycle: follicle development and endocrine interactions. Technical Report 10–06, Konrad-Zuse-Zentrum für Informationstechnik Berlin
8. Deuflhard P (2004) Newton methods for nonlinear problems. Affine invariance and adaptive algorithms. Springer Series in Computational Mathematics, vol 35. Springer, Berlin

9. Townson DH, Tsang PCW, Butler WR, Frajblat M, Griel LC Jr, Johnson CJ, Milvae RA, Niksic GM, Pate JL (2002) Relationship of fertility to ovarian follicular waves before breeding in dairy cows. J Animal Sci 80(4):1053–1058
10. Revah I, Butler WR (1996) Prolonged dominance of follicles and reduced viability of bovine oocytes. J Rep Fert 106(1):39–47
11. Bleach ECL, Glencross RG, Knight PG (2004) Association between ovarian follicle development and pregnancy rates in dairy cows undergoing spontaneous oestrous cycles. Reproduction 127(5):621–629
12. Boer HMT, Röblitz S, Stötzel C, Veerkamp RF, Kemp B, Woelders H (2011) Mechanisms regulating follicle wave patterns in the bovine estrous cycle investigated with a mathematical model. J Dairy Sci 94 In Press

Chapter 36
Reducing Systems Biology to Practice in Pharmaceutical Company Research; Selected Case Studies

N. Benson, L. Cucurull-Sanchez, O. Demin, S. Smirnov, and P. van der Graaf

Abstract Reviews of the productivity of the pharmaceutical industry have concluded that the current business model is unsustainable. Various remedies for this have been proposed, however, arguably these do not directly address the fundamental issue; namely, that it is the knowledge required to enable good decisions in the process of delivering a drug that is largely absent; in turn, this leads to a disconnect between our intuition of what the right drug target is and the reality of pharmacological intervention in a system such as a human disease state. As this system is highly complex, modelling will be required to elucidate emergent properties together with the data necessary to construct such models. Currently, however, both the models and data available are limited. The ultimate solution to the problem of pharmaceutical productivity may be the virtual human, however, it is likely to be many years, if at all, before this goal is realised. The current challenge is, therefore, whether systems modelling can contribute to improving productivity in the pharmaceutical industry in the interim and help to guide the optimal route to the virtual human. In this context, this chapter discusses the emergence of *systems pharmacology* in drug discovery from the interface of pharmacokinetic–pharmacodynamic modelling and systems biology. Examples of applications to the identification of optimal drug targets in given pathways, selecting drug modalities and defining biomarkers are discussed, together with future directions.

N. Benson (✉) • L. Cucurull-Sanchez
Modelling and simulation, Department of Pharmacokinetics, Dynamics and Metabolism, Pfizer Worldwide Research, Pfizer Ltd., Sandwich CT13 9NJ, UK
e-mail: neil@xenologiq.com

O. Demin • S. Smirnov
Institute for Systems Biology, Leninskie Gori, Moscow 11992, Russia
e-mail: demin@insysbio.ru; smirnov@insysbio.ru

P. van der Graaf
Pfizer, Pharmacometrics, Global Clinical Pharmacology, Walton Oaks KT20 7NS, UK
e-mail: piet.van.der.graaf@pfizer.com

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_36, © Springer Science+Business Media, LLC 2012

1 Introduction

Recent reviews of the productivity and processes of the pharmaceutical industry have concluded that the current business model is not optimal, or even unsustainable [1]. More specifically, detailed analyses have highlighted that arguably the most significant challenge facing the pharmaceutical industry is compound attrition (the proportion of first in human trial registered drugs that are subsequently approved, a success rate measure that is currently about 10%) [2]. In turn this reflects the failure of preclinical efficacy and safety model data to translate into human proof of mechanism/concept. Various strategies have been proposed for tackling this attrition issue, including increased outsourcing, use of biomarkers, personalised medicine, adaptive trial design, open innovation and a greater direct input from academia [3]. However, one topic that is often not evaluated is the cause of the apparent lack of productivity. No doubt a major component of this is the complexity of questions arising in drug discovery; there are $\sim 25k$ genes in the human genome potentially giving rise to an estimated 1.8 million protein species [4]. There are more than 300 cell types, 4 types of tissue and 12 organ systems. Together these give rise to the organism and its behaviours over timescales ranging from msec to decades, where interactions with the environment influence outcome, be it disease or non-disease. Clearly, the potential complexity of this is staggering; for example, theoretically the number of paired combinations of the $25k$ genes alone yields $\sim$300 million interactions and one conclusion is that predicting the impact of pharmacological intervention is likely to be extremely difficult and non-intuitive.

The notion that cost savings can be achieved via strategic re-organisation may be true, but without a shift in the fundamental understanding of disease biology there is no reason to believe that attrition will improve relative to that observed in the past decades. In this context, experimenting with approaches to improve the understanding of complex biology prior to engaging significant resource would appear logical. In other disciplines such as engineering, finance and environmental science, mathematical modelling is used successfully and indeed the application in biology is growing dramatically [5]. In the age of rapidly developing information technology and computational capabilities, the potential to integrate, share and visualise vastly increasing sources of biological data is without precedent. Indeed this exciting progress at least introduces the idea of generating a virtual human, with for example the virtual physiological human attracting of much current attention [6]. However, the gap between the promise and the reality of understanding a system such as the human body are enormous; undoubtedly most of the data needed to construct models with which we can confidently predict outcome do not exist yet and Cohen's recent statement that "Despite the eloquent pleas that have been made for model-based drug development, it is clear that in many cases the basic data to do this simply are still lacking" remains true [7]. Although it would appear clear that collecting data on the behaviour of human proteins, cells and tissues and integrating these to generate a better understanding is a desirable objective, reducing

this to practice in an efficient way is difficult. Potentially industrial/academic pre-competitive consortia could be a fruitful way to realise synergies and tackle the tough technical, scientific and logistical challenges. However, it is likely to be many years, if at all, before the virtual human is available. The immediate challenge is, therefore, whether systems modelling can contribute to improving productivity in the pharmaceutical industry in the interim.

But why should an improved, but far from complete understanding of system complexity benefit drug discovery? A common response to this question is to conclude that effort to understand such a complex and large system a human disease state is likely to be futile and instead to use surrogates in which to observe emergent properties; i.e., in *vitro* systems and in *vivo* animal experiments where the influence of drugs can be evaluated in complex models of disease states. However, we should recall this paradigm has been in place for decades and culminated in the attrition observed, with far too many encouraging preclinical discoveries turning out to be false positives when evaluated in human disease.

An alternative approach is a strategy that aims to build a contextual understanding of the system. This is currently being employed within companies such as Pfizer, by combining and integrating the ideas and approaches from the disciplines of pharmacokinetic–pharmacodynamic (PKPD) modelling with systems biology and understanding the actions and adverse effects of drugs by considering targets in the context of the biological networks in which they exist. This trend has been labelled "systems pharmacology" to highlight the integration of disciplines that have until recently been distinct [8–10]. Deliverables of systems pharmacology include the identification of testable hypotheses and likely controlling parameters in the known system; a premise of this thinking is that by measuring and therefore defining these parameters, our knowledge will be improved and hence such *biomeasures* (i.e., quantitative information about system properties such as, for example, target expression levels and turnover dynamics) are key enablers for success in this area in a similar way as *biomarkers* have been for PKPD. By generating successive cycles of modelling and biomeasures, the confidence in models can be increased. These experimentally tested models can then be used for project selection and progression, facilitating the assessment of the relative risks of projects. Furthermore, optimal experiments to de-risk programmes ahead of Phase 2 can be designed and projects with a low probability of success terminated early (see Fig. 36.1). Although the added value of systems pharmacology would be greatest if novel targets can be identified with high confidence, there are also benefits from terminating work early on targets that will ultimately lack the requisite safety or efficacy in disease. The resource saved could be re-used to expedite the remaining programs or to generate data and models to improve our systems understanding.

Some examples of useful application in pharmaceutical research follow, but in general the work involves three phases. The first is to engage with a project team to clearly define the question posed. Although this seems trivial and obvious, in reality this fundamental requirement can be overlooked or indeed be difficult to define. Nevertheless, several years experience of such projects has shown repeatedly

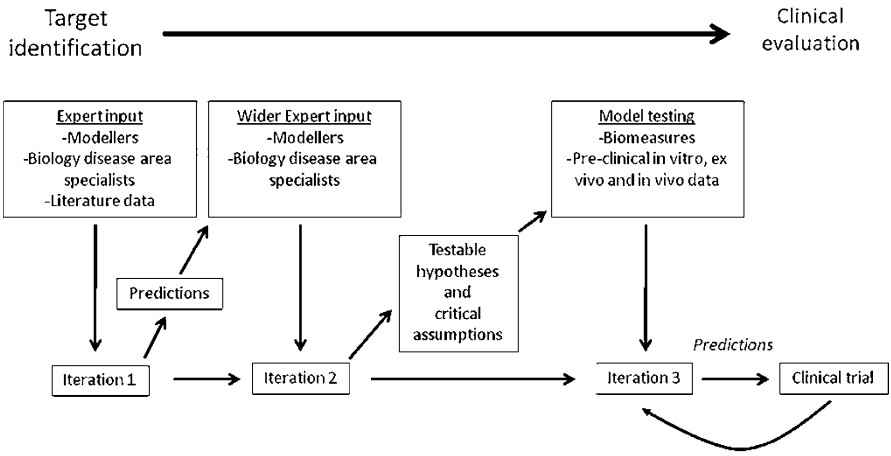


Fig. 36.1 Schematic showing an example of how systems pharmacology could be implemented in drug discovery projects. From an initial evaluation of the literature and input from disease biology experts, with a mathematical model can be constructed (Iteration 1). For example this could be of a particular pathway of interest (e.g. the NFKB pathway). The predictions of this model are subject to feedback from a wider expert panel and subsequently this iteration (Iteration 2), via for example sensitivity analyses, is used to identify critical assumptions and testable hypotheses. At this stage, the determination of key system parameters such as target molecule concentrations (biomeasures) is likely to be critical. An iteration of the model is then produced that is consistent with these data (Iteration 3). This model can be used to predict clinical outcome and contribute to trial design e.g. initial dose prediction. Finally, the actual result can be compared to predictions and the conclusions incorporated in subsequent projects

that having this clarity is critical to a successful outcome. The second phase is to establish the relevant facts via a detailed survey of the available internal and external data. In this regard, the use of state of the art text mining [11] can be very helpful. Finally, once a quantitative and holistic assessment of the available data has been made, a hypothesis that addresses the question can be proposed and described in a mathematical model (usually based on ordinary differential equations). The use of a mathematical model has many advantages. Firstly, with more than a handful of components, it quickly becomes very difficult, if not impossible, to picture intuitively the emergent properties of a given system. In contrast, the mathematical model enables a methodical interrogation of this. Secondly, the model can succinctly summarise the structure of a hypothesis, the parameters included and the sources of parameters used to make predictions. In turn this allows assumptions to be made explicit and facilitates the identification of non-intuitive experiments that can validate or invalidate the model. Together, this can then enable communication between the groups in a given drug discovery project. Finally, the model is not static and can be revised as data becomes available. Hence, it can develop with a project as a tool to aid decision making. Of course, the tool should be used in addition to standard decision making input such as preclinical animal model and in vitro data and considered in this context.

2 Examples

Medicinal chemistry strategy for avoiding CNS penetration. In certain circumstances, excluding a drug from the CNS could be perceived as an advantage. An example is in the case of opioid agonists, where such drugs have negative effects in the CNS but potentially positive (analgesic) benefits in the periphery. In order to prevent a small molecule from entering the CNS, two obvious medicinal chemistry options are apparent, either designing a molecule with poor blood–brain barrier (bbb) permeability or designing in attributes that render the molecule a substrate for P-glycoprotein (PGP). PGP is expressed in the bbb and thought to be a major contributor to the poor CNS penetration of small molecules [12] with export against the concentration gradient achieved via active transport. The question in this instance is whether one of these strategies is superior to the other. In order to address this, a model of the bbb was constructed (cf. Fig. 36.2 and [13]) incorporating drug target binding kinetics. Given that the drug was required to be dosed to a pseudo steady state to achieve the clinical endpoint, we concluded that because the surface area of the brain is large (200,000 cm^2) and the clearance of the drug across the bbb is the product of the surface area and permeability, then success for the non-PGP option was likely to require very low permeability. In this, the quantification of "low" was critical and the model highlighted the need for better methods for determining the permeability of the bbb that give a true parameter estimate that can be used in a prognostic sense, rather than in a relative sense as is current practice [14]. In summary, the model highlighted a risk with the non-PGP option that is not immediately apparent without a tool to explore our understanding of this complex system. Hence, in cases such as this, the systems model view suggests that developing a drug that is a PGP substrate is the superior strategy.

Inhibition of the NFkappaB (NFKB) pathway by small interfering (si) RNA; in this case example the project team posed two questions; firstly, what are the optimal targets within the NFKB pathway and secondly what fractional inhibition is required to significantly impact the key endpoint of the system? Using a published model [15], a sensitivity analysis showed that I kappa B kinase (IKK) was the most tractable influential drug target in the pathway of interest. By assuming, as a base case, that the area under the curve (AUC) of the oscillation of NFKB in the nucleus represented a measure of outcome and a 90% decrease in AUC equated to a positive drug response, the model also showed that inhibition of $>95\%$ of the activity of IKK would be required to achieve this goal. In this particular case, the drug modality proposed was siRNA; in general with this technology, the maximum reduction in protein expression that can be achieved is $\sim 80\%$. Hence, we were able to predict that siRNA was unlikely to yield the necessary efficacy and that, at least potentially, a conventional small molecule carried a higher probability of success. It was subsequently found, using human in vitro cell based assays, that although some inhibition of inflammatory endpoints such as IL8 could be achieved (30%), it was less than required (personal communication from S. Moschos). This result highlights the potential for the systems pharmacology approach to set clear criteria for lead compound selection prior to engaging significant resource.

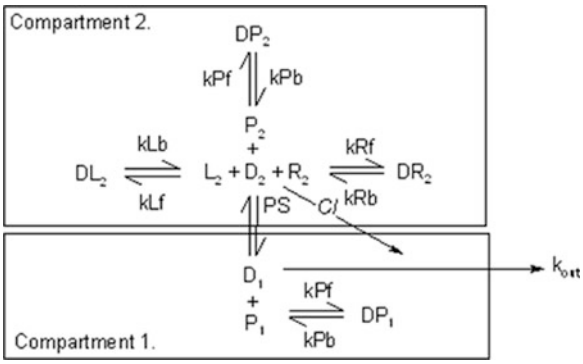


Fig. 36.2 Schematic description of the blood–brain barrier model incorporating target binding kinetics. The model was used to inform decisions on medicinal chemistry strategy. Compartment 1 is the central compartment (e.g., the plasma) and compartment 2 the peripheral (e.g., the central nervous system). D1 and P1 are, respectively, the drug and protein concentration in the central compartment; DP1 is the concentration of the drug–protein complex. D2, P2, R2 and L2 are the concentrations of drug, protein, receptor and lipid in the peripheral compartment; DP2, DR2 and DL2 are the concentrations of the respective complexes of drug and protein, receptor or lipid. PS is the permeability-surface area product and k_{out} is the rate constant for removal of the drug from the plasma

Hepatitis C Virus (HCV) life cycle modelling; in HCV drug research a key question is why the standard of care [Interferon-α (IFN-α) plus ribavirin] is effective for only approximately 50% of patients and, subsequently, what alternative treatments are plausible. By integrating a selection of published models [16–19] together with expressions for the PKPD behaviour of (IFN-α) and incorporating other literature data, we were able to explore this and develop new hypotheses. Notably, one of the key characteristics determining susceptibility of HCV patients to IFN-α therapy (responder vs. non-responder) is the EC_{50} of IFN-α and indeed it has been concluded from analyses of clinical data that the EC_{50} of non-responder patients substantially exceeds that measured for responder patients [20]. The two main experimentally observed phenomena contributing to the variable EC_{50}'s of IFN-α treatment are desensitisation and refractoriness. Desensitisation is defined as an IFN-α independent decrease in the sensitivity of hepatocytes with respect to IFN-α which results from an interaction of virus particles with signalling proteins/receptors. Refractoriness is defined as an IFN-α dependent decrease in the sensitivity of hepatocytes to IFN-α̇ treatment which is based on negative feedback observed in IFN-α signalling pathway.

To capture these quantitative characteristics of signalling pathways mediating the response of the hepatocyte to interferon and virus, we have reconstructed key signalling pathways participating in an IFN-α based cell response in a mathematical model of HCV combining virus/cell dynamics with a quantitative description of IFN-α dependent JAK/STAT mediated signalling pathways (see for example [21]). Molecular mechanisms explaining both desensitisation and refractoriness and,

consequently, the status of HCV patients (responder vs. non-responder) in terms of dynamic and regulatory properties of signalling and gene regulatory pathways have been identified. The model has also been applied to identify possible biomarkers indicating the patient status before IFN-α therapy is applied. Taking into account the values of the biomarkers, the model was also able to predict the optimal dosage and administration regime strategy for each particular patient.

Another example of application of this HCV life cycle model concerns identification of alternative plausible treatments. For example, our models indicated that a viral entry blocking drug, with realistic pharmacological and pharmacokinetic properties, in combination with the standard of care, could achieve a cure in the non-responder population. Moreover, the model enabled a rational interrogation of the drug pharmacological and PK parameters required for success (in this case to elicit a cure within six months). We found that, for example, a hypothetical drug blocking viral entry (e.g., via antagonism of the HCV:host cell interaction) with ∼0.1 nM potency and exhibiting a PK half life ($t_{1/2}$) of twenty one days would meet these criteria at a feasible dose, when combined with the standard of care. These properties are consistent with those often exhibited by IgG monoclonal antibody drugs, where picomolar or even femtomolar affinities can be achieved and $t_{1/2}$'s are typically around 21 days [22]. Notably, there are some data showing that anti-CD81 antibodies can prevent re-infection in vivo [23].

This kind of model based input is very useful in drug discovery projects, where often the range of possibilities at the outset is vast. For example, it is often difficult to decide whether a small molecule or biological drug is likely to be optimal, or in either case what is necessary for success in terms of drug pharmacology, PK and dose. By providing potential answers to such questions, the strategy can be focused, minimising the time and resource required to deliver a drug. The caveats in this case are the assumptions inherent in the model, and ultimately the model prediction will need to be tested clinically.

3 Conclusions

In our experience, the systems pharmacology approach has contributed to improved decision making in projects, via identification of those carrying a high risk of failure. In addition, insight into likely requirements for success has been provided, for example in terms of pharmacology and PK required for a drug. A further deliverable has been the generation of experimentally testable hypotheses that can be readily evaluated early in drug discovery. Integrating these biomeasure data to improve models and/or make decisions in a timely way requires a dynamic interaction between model building, data generation and feedback from disease biology experts. In all cases, a key component for success was the formation of clear questions at the outset of modelling. However, although there are promising signs that systems pharmacology can add value in drug discovery, there are many challenges. For example, given the early stage of research into the application of

systems pharmacology, all current models contain assumptions, any or all of which could invalidate conclusions. Only by accumulating a body of data on the outcome of model predictions will we begin to learn about the strengths and weaknesses of this approach and where the most critical gaps in data and understanding exist. Conceivably one of the scientific challenges may be that truly contextual biomeasures are required, for example from human tissue. Clearly this raises questions from a practical and ethical perspective that will need to be addressed. Furthermore, the interoperability challenges in terms of sharing and integrating models require continued effort [24]. Irrespective of these challenges we envisage, further integration of mechanistic modelling into drug discovery processes in the future. This could include a spectrum of modelling and simulation approaches such as further mechanistic PKPD, systems biology, systems pharmacology and network biology [25]. Ultimately, integrating these disciplines effectively will deliver the best possible understanding of complex biology. Furthermore, systems pharmacology interfaced with disease biology expertise will generate new insights into aspects of diseases biology that will add value in the context of drug discovery (see for example [26]). Having access to the expertise to identify and implement these ideas in pharmaceutical research will be critical to success in the future. To this end, the provision of educational programmes that provide the optimal training will be crucial.

References

1. Munos B (2009) Lessons from 60 years of pharmaceutical innovation. Nat Rev Drug Discov 8(12):959–968
2. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3:711–716. doi:10.1038/nrd1470
3. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R& D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9:203–214. doi:10.1038/nrd3078
4. Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol 8:33–41
5. Kohl P, Crampin EJ, Quinn TA, Noble D (2010) Systems biology: an approach. Clin Pharmacol Ther 88(1):25–33. Epub 2010 June 9
6. Kohl P, Noble D (2009) Systems biology and the virtual physiological human. Mol Syst Biol 5:292. Epub 2009 July 28
7. Cohen A (2008) Pharmacokinetic and pharmacodynamic data to be derived from early-phase drug development-designing informative human pharmacological studies. Clin Pharmacokin 47:373–381
8. Benson N, vander Graaf PH (2011) Systems pharmacology: bridging systems biology and pharmacokinetics–pharmacodynamics (PKPD) in drug discovery and development. *Pharm Res.* 2011 Jul;28(7):1460–4. Epub 2011 May 11.
9. Swat MJ, Kiełbasa SM, Polak S, Olivier B, Bruggeman FJ, Tulloch MQ, Snoep JL, Verhoeven AJ, Westerhoff HV (2011) What it takes to understand and cure a living system: computational systems biology and a systems biology-driven pharmacokinetics–pharmacodynamics platform. Interface Focus 1:16–23

10. Wist AD, Berger SI, Iyengar R (2009) Systems pharmacology and genome medicine: a future perspective. Genome Med 1(1):11
11. Ananiadou S, Kell DB, Tsujii J-i (2006) Text mining and its potential applications in systems biology. Trends Biotechnol 24(12):571–579
12. Löscher W, Potschka H (2005) Blood–brain barrier active efflux transporters: ATP-binding cassette gene family. NeuroRx 2(1):86–98
13. Peletier LA, Benson N, van der Graaf PH (2010) Impact of protein binding on receptor occupancy: a two-compartment model. J Theor Biol 265(4):657–671. Epub 2010 June 2
14. Feng MR (2002) Assessment of blood–brain barrier penetration: in silico, in vitro and in vivo. Curr Drug Metab 3(6):647–657
15. Ihekwaba AE, Broomhead DS, Grimley RL, Benson N, Kell DB (2004) Sensitivity analysis of parameters controlling oscillatory signalling in the NF-kappab pathway: the roles of IKK and ikappabalpha. Syst Biol 1(1):93–103
16. Dahari H, Shudo Emi, Cotler SJ, Layden TJ, Perelson AS (2009) Modeling hepatitis C virus kinetics: the relationship between the infected cell loss rate and the final slope of viral decay. Antivir Ther 14(3):459–464
17. Dahari H, Lo A, Ribeiro RM, Perelson AS (2007) Modeling hepatitis C virus dynamics: liver regeneration and critical drug efficacy. J Theor Biol 247(2):371–381. Epub 2007a, Mar 14
18. Dahari H, Ribeiro RM, Perelson AS (2007) Triphasic decline of hepatitis C virus RNA during antiviral therapy. Hepatology 46(1):16–21
19. Dixit NM, Layden-Almer JE, Layden TJ, Perelson AS (2004) Modelling how ribavirin improves interferon response rates in hepatitis C virus infection. Nature 432(7019):922–924
20. Snoeck E, Chanu P, Lavielle M, Jacqmin P, Jonsson EN, Jorga K, Goggin T, Grippo J, Jumbe NL, Frey NA (2010) Comprehensive hepatitis C viral kinetic model explaining cure. Clin Pharmacol Ther 87(6):706–713. Epub 2010 May 12
21. Maiwald T, Schneider A, Busch H, Sahle S, Gretz N, Weiss TS, Kummer U, Klingmüller U (2010) Combining theoretical analysis and experimental data generation reveals IRF9 as a crucial factor for accelerating interferon-induced early antiviral signalling. FEBS J 277(22):4741–4754
22. Igawa T, Tsunoda H, Kuramochi T, Sampei Z, Ishii S, Hattori K (2011) Engineering the variable region of therapeutic igg antibodies. MAbs. 3(3). Epub ahead of print
23. Meuleman P, Hesselgesser J, Paulson M, Vanwolleghem T, Desombere I, Reiser H, Leroux-Roels G (2008) Anti-CD81 antibodies can prevent a hepatitis C virus infection in vivo. Hepatology 48(6):1761–1768
24. Finney, A., Hucka, M., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C., Kitano, H. (2006). "Software Infrastructure for Effective Communication and Reuse of Computational Models". *Systems Modeling in Cell Biology: From Concepts to Nuts and Bolts.* MIT Press. pp. 369-378.
25. Rosenfeld S (2011) Mathematical descriptions of biochemical networks: stability, stochasticity, evolution. Progr Biophys Mol Biol 106(2):400–409. doi:10.1016/j.pbiomolbio.2011.03.003
26. Valeyev NV, Hundhausen C, Umezawa Y, Kotov NV, Williams G, Clop A, Ainali C, Ouzounis C, Tsoka S, Nestle FO (2010) A systems model for immune cell interactions unravels the mechanism of inflammation in human skin. PLoS Comput Biol 6(12):e1001024

Chapter 37
System-Scale Network Modeling of Cancer Using EPoC

Tobias Abenius, Rebecka Jörnsten, Teresia Kling, Linnéa Schmidt, José Sánchez, and Sven Nelander

Abstract One of the central problems of cancer systems biology is to understand the complex molecular changes of cancerous cells and tissues, and use this understanding to support the development of new targeted therapies. EPoC (Endogenous Perturbation analysis of Cancer) is a network modeling technique for tumor molecular profiles. EPoC models are constructed from combined copy number aberration (CNA) and mRNA data and aim to (1) identify genes whose copy number aberrations significantly affect target mRNA expression and (2) generate markers for long- and short-term survival of cancer patients. Models are constructed by a combination of regression and bootstrapping methods. Prognostic scores are obtained from a singular value decomposition of the networks. We have previously analyzed the performance of EPoC using glioblastoma data from The Cancer Genome Atlas (TCGA) consortium, and have shown that resulting network models contain both known and candidate disease-relevant genes as network hubs, as well as uncover predictors of patient survival. Here, we give a practical guide how to perform EPoC modeling in practice using R, and present a set of alternative modeling frameworks.

1 Introduction

The molecular exploration of cancer is still in its infancy. In the next few years, consortia such as the Cancer Genome Atlas project (TCGA), the Cancer Genome Project (CGP), and the international cancer genome consortium (ICGC) will

T. Abenius • R. Jornsten • J. Sánchez
Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, 412 96 Gothenburg, Sweden
e-mail: Tobias.Abenius@chalmers.se; jornsten@chalmers.se; sanchezj@chalmers.se

T. Kling • L. Schmidt • S. Nelander (✉)
Cancer Center Sahlgrenska, Institute of Medicine, Box 425, 415 30 Gothenburg, Sweden
e-mail: teresia.kling@gu.se; linnea.schmidt@gu.se; sven.nelander@gu.se

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_37,

produce comprehensive observations of molecular changes in solid tumors and leukemias [34]. Mathematical models that integrate several levels of the cancer genome data can prove helpful in the study of several key problems in cancer biology, such as,

1. the identification of "disease driving genes" whose altered copy number impact transcription,
2. the construction of molecular features that are predictive of patient survival, and
3. the discovery of possible therapeutic targets by matching the identified network model hub-genes, or their targets, to pharmacological databases.

Key recent examples of advanced integrative analyses in the literature include modular network modeling combined with clustering analyses, resulting in the discovery of regulators *MITF*, *RAB27A*, and *TBC1D16* in malignant malinoma [2, 9], and the association of *cMYC* amplification to wound healing signatures in breast cancer [1]. Network analysis of a set of transcripts known to be related to breast cancer and relating these to 384 genomic regions with altered copy number has identified a candidate regulatory region on chromosome 17 [23]. For a broader overview that also covers possible non-network approaches to modeling cancer, see Sect. 4 below and [15].

1.1 Network Models Reveal Regulation and Prognostic Scores in Glioblastoma

We recently explored the idea to view acquired genetic variation in tumors (copy number aberrations, CNAs, or single nucleotide variations, SNVs) as informative, "endogenous perturbations", which are analyzed jointly with mRNA profiles to derive causal, system-scale network models (Fig. 37.1). CNAs are prevalent in several human cancers. Moreover, these genetic variations tend to appear in a patient-specific, near multifactorial manner in the tumors, thus resembling an optimal experimental design to derive causality [4]. The use of CNAs as informative perturbations is complementary to using, e.g., RNAi or SNP variation to derive models, e.g., [18, 35].

In our recent work, we develop the modeling technique EPoC and apply it to 186 cases of glioblastoma. The resulting model is the first large-scale model of transcriptional effects of copy number aberrations in glioblastoma, and reveals interesting candidates, not detected by other methods. For instance, targeted validations in four glioblastoma cell lines implicate the p53-interacting protein Necdin in suppressing glioblastoma cell growth (Fig. 37.1). A novel prognostic score, based on the singular value decomposition of the network, successfully stratified glioblastoma patients into poor/favorable prognosis, not achieved by, e.g., principal component analysis (PCA) or clustering (Fig. 37.1b). In technical tests, EPoC performs better than existing tools (e.g., ARACNE, eQTL) in terms of robustness, biological accuracy, and speed [15].

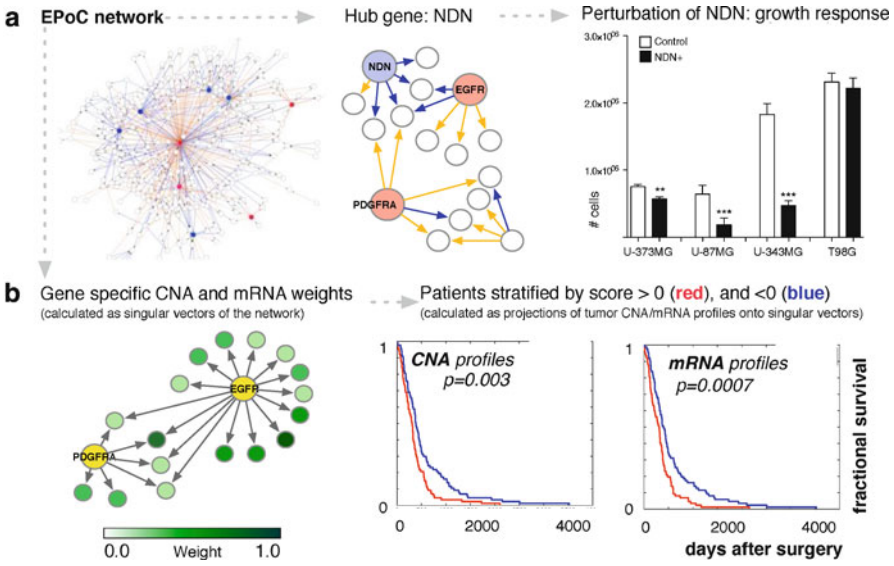


Fig. 37.1 Network modeling of glioblastoma. (**a**) Targeted validations in four glioblastoma cell lines support selected predictions, and implicate the p53-interacting protein Necdin in suppressing glioblastoma cell growth. (**b**) EPoC uses a novel procedure to isolate CNA (yellow) and mRNA (green) prognostic biomarkers, here shown in a network context. Kaplan–Meier curves show that network-derived patient scores achieve prognostic separation [15]

1.2 Making System-Scale Network Modeling Work in Practice: The EPoC Package

Taken together, our data show that large-scale network modeling of the effects of copy number aberrations on gene expression may provide insights into the biology of human cancer. The benefits of our approach become clear when considering that (1) colossal amounts of data are rapidly becoming available, motivating large-scale explorations using this method and, (2) EPoC is distributed as a high-performance software (R / MATLAB) that others can use.

In this chapter, we give a comprehensive presentation of how to perform EPoC modeling of human cancer in practice in the R language. We also cover several alternative modeling approaches for system-scale modeling of cancer, and discuss their key differences in terms of implementation, algorithmic speed, and performance. The first section gives a brief overview of EPoC modeling principles. In the second section, we give a compact tutorial how to build and validate EPoC models based on combined CNA and mRNA data from a set of human tumors. This tutorial should be accessible to all researchers with a working knowledge of R. In the last sections, we discuss other modeling alternatives and outline future challenges.

2 EPoC: Modeling Copy Number-Dependent Transcription in Tumors

2.1 Biological Assumptions and Underlying Mathematical Model

The goal of EPoC modeling, is to construct a global network model that explains the transcriptional consequences of DNA copy number alterations. By assuming a formal model of transcriptional regulation, given in [15], we demonstrate that the CNA-driven transcriptional steady state in a set of tumors can be expressed as two complementary linear matrix equations. We first describe the model for transcriptional regulation via the so-called *transcriptional network*, A:

$$A\Delta Y + \Delta U + R = 0. \tag{37.1}$$

The data matrices ΔY and ΔU represent the mRNA and CNA profiles of glioblastoma, respectively. Each column in ΔY contains the mRNA expression profile for a patient, each row a gene's expression across patient tumors. Similarly, the matrix ΔU contains the copy number profiles of patients, where copy number altered regions have been mapped to the set of genes for which transcription levels have been measured. For the model to make sense, the mRNA profiles should be log-relative expression levels, and be zero-centered, i.e., each row of ΔY should have mean zero [15]. Similarly, the CNA profiles must be log-relative and zero-centered. Finally, the noise term R (defined from parameters of the underlying network model, not shown) is a matrix that captures the effects on transcription of non-CNA perturbations in individual tumors (e.g., SNPs, sequence mutations, or environmental effects). We can interpret the elements of the transcriptional network $A = \{a_{ij}\}$ as follows: the elements a_{ij} represent the net influence from transcript j to transcript i; $a_{ij} > 0$ indicates activation of transcription i by transcript j, $a_{ij} < 0$ inhibition, and the magnitude a_{ij} the strength of the interaction. The transcriptional regulation model can be represented in a second form, where we term the *CNA-driven network* (G):

$$\Delta Y = G\Delta U + \Gamma. \tag{37.2}$$

The elements of $G = \{g_{ij}\}$ consists of CNA–mRNA couplings: $g_{ij} > 0$ indicates CNA-driven transcriptional activation, in the sense that transcription of gene i is increased because the copy number of gene j has been altered. Similarly, $g_{ij} < 0$ indicates a negative coupling between a CNA at gene j and transcription of gene i. For both positive and negative couplings, the magnitude of g_{ij} reflects the strength of the interaction. The network G is related to the transcriptional network as $G = -A^{-1}$ and thus the topologies of the two networks are related. However, while A aims to capture direct transcriptional interaction, corrected for the impact of a transcript's own CNA, G models how the effects of CNA perturbations propagate

through the system to affect transcription indirectly. We therefore conjecture that the alternative network representation G should contain key disease-driving CNAs as hubs, as well as their downstream targets.

2.2 Global Estimation of Network Models Using EPoC

To estimate A and G from the data, EPoC combines methods from lasso regression and non-parametric (Breiman's pseudo-)bootstrap. For each gene, EPoC first estimates the local transcriptional effect of that gene's own copy number aberration. For a particular gene i, this is done by a truncated least squares estimate:

$$d = \max\left(0, \Delta U_i \Delta Y_i^T\right), \tag{37.3}$$

where ΔY_i and ΔU_i are the ith row of the data matrices, respectively (data for gene i across all tumors). For each gene, we then solve a $L1$ regularized regression problem to identify the CNAs that significantly affect the residual transcript after a gene's own copy number has been accounted for:

$$\min_{G_i} \left\| (\Delta Y_i^T - d\Delta U_i^T) - G_i \Delta U_{\setminus i}{}^{\mathsf{T}} \right\|_F^2 + \lambda \sum_{j \setminus i} \mid G_i[j] \mid, \tag{37.4}$$

where $\Delta U_{\setminus i} = \Delta U[\setminus i, \setminus i]$, i.e., the ΔU matrix excluding gene i. G_i here denotes the ith row in G with element i excluded (the diagonal of G). λ is the regularization parameter that controls the degree of sparsity (number of non-zeroes, i.e., network links) in G. $G_i[j]$ denotes the jth element in vector G_i. We solve (37.4) using the cyclic coordinate descent (CCD) algorithm [5, 8].

Details on the estimation of network parameters can be found in [15], but we summarize the main steps in EPoC modeling below:

(1) *Import and standardize the data*. For our model to be meaningful, the mRNA and CNA data needs to be row-centered. Two recommended optional steps are standardization of transcription profile variance (to 1), and pre-filtering of a subset of ca. 25% of the most CNA-altered genes as possible regulators (below).
(2) *Perform a statistical simulation to select* λ. A key feature of EPoC is a novel model validation technique; instead of a minimizing cross-validation prediction error (which is the most commonly used criterion in lasso regression), EPoC maximizes Kendall's W, a score that measures the concordance of networks between data splits (below) [16] . We argue that this criterion produces smaller models, more likely to be structurally consistent between replicate data sets (c.f. [15]).
(3) *Identify hubs and robust links*. EPoC solidifies the results by a 1000-fold pseudo-bootstrapping protocol, in which the network is reconstructed 1000 times, from resampled patient sets. Links that appear at high frequencies are kept as regulatory links.

(4) *Identify prognostic biomarkers.* EPoC derives prognostic scores from the network by the use of a sparse singular value decomposition of the network, as explained below.

In Sect. 3 we describe how to perform each of these steps in practice.

2.3 Biological Assessment of the Derived Networks

The derived networks are exported to Cytoscape for assessment. In our own analyses of EPoC networks, we have thus far focused on three key network features. First, we consider hub genes in the CNA-driven network, i.e., genes which are recurrently altered at the CNA level, and which are inferred by EPoC to control multiple downstream genes (e.g., the genes EGFR, PDGFRA, and NDN in Fig. 37.1). Second, we explore the functional annotation of downstream genes, since this may help elucidate the mechanisms by which CNAs drive the disease process. As an example of this, our EPoC model of glioblastoma contained multiple co-regulated genes involved in early neural development, including PROM1 (CD1333), NKX2.2, and SOX10/11 factors [11, 24, 29]. Third, we use Cytoscape to inspect the intersection between our EPoC model and pharmacological databases (Drugbank, Ingenuity, Chembank, KEGG) to derive compound-target representations, which are analyzed manually to identify targets. This assessment can be followed up by targeted experiments to pursue mechanisms, e.g., growth effects of hub gene perturbation (Fig. 37.1). To further exemplify network analysis, we show results where the A and G networks are compared for our glioblastoma model (Fig. 37.2). In comparison to the G matrix, the A matrix is strongly enriched for genes associated with inflammation, immune response, and blood lineage markers. These results are best explained by the fact that the estimate of A is highly dependent on mRNA–mRNA correlations, and that much of the mRNA variability in our samples is caused by variations in the stromal, blood, and inflammatory components across the tumor samples.

2.4 Network Based Survival Scores

We now describe how to use the estimated EPoC network models to derive prognostic scores for patient survival. The CNA-driven network G can be interpreted as a model for signal amplification. That is, viewing transcriptional regulation from a systems perspective, we think of a copy number profile as *input* into the system G resulting in a mRNA expression *output*. We loosely think of the CNA input as the driver of the disease and the system output as the symptom.

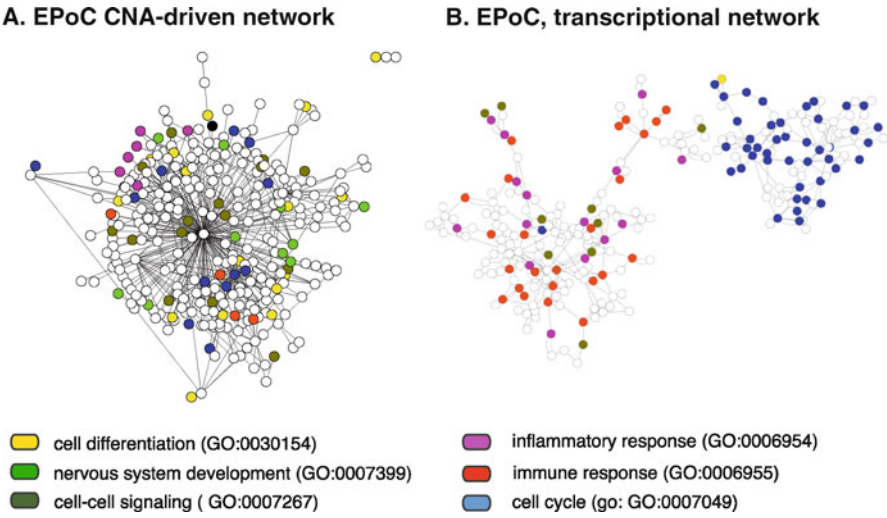


Fig. 37.2 Differences in gene content between EPoC G and A networks. Differences between the CNA-driven (**a**) and transcriptional (**b**) networks are highlighted using the top three enriched GO process terms (the corrected Fisher's test p-values ($<10^{-9}$ for all terms shown) are used as a ranking principle and not as evidence of network links). The CNA-driven network contains genes involved in cell–cell signaling and developmental processes, whereas the transcriptional network contains a large number of genes associated with inflammatory and cell cycle associated processes

To summarize the input–output behavior of a system it is common to compute the main axes of signal gain, defined as the singular value decomposition (SVD) [10, 21, 30]:

$$G = C\Lambda D^T, \tag{37.5}$$

where $CC^T = I$, $DD^T = I$, and Λ is diagonal. The leading left (C) and right (D) SVD components have the following meaning: large elements of the leading components of D represent genes whose CNAs are highly amplified by the system. That is, copy number aberrations for these genes have a substantial and potentially broad impact on mRNA expression. The large magnitude elements of the leading components of C identify the genes whose mRNA expression are most affected by these copy number altered genes. To increase the interpretational strength of the decomposition, we compute C, Λ, and D using *sparse SVD* [38, 40]. The sparse SVD components C and D contain only a small subset of non-zero elements as those with low magnitude are eliminated via a combined $L1$ and $L2$ (elastic net) regularization. The subset of genes present in each leading components facilitate the identification of key disease-driving CNAs (from the sparse D) and their corresponding mRNA targets (from the sparse C).

Once sparse estimates of C and D have been obtained, EPoC computes the level of signal amplification in each tumor by the scalar projection scores:

$$\begin{cases} Z_y = C^T \Delta Y \\ Z_u = D^T \Delta U. \end{cases} \tag{37.6}$$

These scores summarize the total burden of molecular changes consistent with the CNA-driven network, i.e., how well the patient profile aligns with the identified disease-driving CNAs and the corresponding mRNA profile, and should therefore correlate with clinical survival. For the different components of Z_y and Z_u we thus compare patients stratified according to $z > 0$ and $z < 0$ in terms of clinical survival (from date of surgery to date of death); survival difference p-values are obtained by Kaplan–Meier curves and the log-rank test. We have confirmed that both $z = Z_y$ and $z = Z_u$ achieve a clinical separation for glioblastoma patients: for the 186 patients studied in [15], both Z_y and Z_u achieved a significant stratification (results for Z_y in (Fig. 37.1), as analyzed by Kaplan–Meier curves and the log-rank test ($p < 0.001$ for both Z_y and Z_u). Using Pearson correlation, the correlation between Z_y and survival is $\rho = 0.21$, ($p = 0.006$) and the same correlation for Z_u is $\rho = 0.19$, $p = 0.0098$. As a reference, we have also demonstrated that the equivalent scores derived directly from an SVD of the Y (mRNA) or the U (CNA) data matrices do not achieve this clinical separation of patients [15].

3 The `epoc` package

We provide a detailed illustration of how to apply the R package (`epoc`) for the joint analysis of mRNA and CNA data. Note, `epoc` also includes functions that can be used to perform network analysis of mRNA data only. The R package contains convenient summary and plotting commands to aid the extraction of the most important information from the estimated network models. The step-by-step demonstration below summarizes the analysis procedure and generated output. For convenience and ease of illustration we perform the demonstration on a synthetic data set (CNA and expression data simulated from TCGA glioblastoma data). These demonstration data are also provided in the package to allow users to try out the program in a controlled setting. Updated information and releases of the package for older versions of R can be found at: http://sysbio.med.gu.se/epoc.html.

3.1 Data Preparation

To begin analysis, we load the EPoC package and the mRNA data (and CNA if available) we wish to analyze. Below we illustrate how to load the synthetic data

which consists of $N = 186$ tumors and $p = 50$ genes. To analyze another dataset, we would have to read it into R as two data matrices: y is the $N \times p$ mRNA data ($N =$ number of tumors, $p =$ number of genes), and u is the $N \times p$ CNA data.

```
> install.packages('epoc')
> require(epoc)
> data(synth)
> u <- synth$u
> y <- synth$y
> N <- dim(y)[1]
> p <- dim(y)[2]
> print(dim(y))
[1] 186  50
```

We check that the dimensions are such that samples are on the rows (N) and the columns (p) are genes. (NB: If the data are in the format where rows are genes and samples are columns we would need to transpose the data matrices. This is easily done with the transpose commands: `y <- t(y); u <- t(u)`.)

In the current version, the `epoc` package cannot handle missing values (a possible future extension if users express interest in this). If the data set contains missing values we need to apply an imputation method, e.g., the k-nearest neighbor method implemented in the R package `impute` [36] which can be installed from the Comprehensive R Archive Network (CRAN) server using the `install.packages` command as above.

It has been our experience that network analysis benefits from data standardization. In the package, we have left this as a user option. To perform optional standardization for each gene, we simply run the following commands:

```
> stdize <- function(x) x.std <- x/sd(x)
> y <- apply(y, 2, stdize)
> u <- apply(u, 2, stdize)
```

For simplicity of notation we use notation `y` and `u` below regardless if the data are standardized or not.

3.2 The Basic EPoC Call

EPoC estimates networks for a sequence of relative penalty parameter values, λ, ranging from zero to one, where λ closer to one produces more sparse networks. The basic EPoC function (`epocG`) includes a small default set of λ-values ranging from very sparse to very dense networks. However, the user can also provide a set of λ-values focusing on a selected sparsity level (see discussion below).

To estimate a network using the default λ sequence the following command is used:

```
> G <- epocG(Y = y, U = u)
```

The output from `epocG` is an object that contains network estimates for each of the included values of λ. The `summary` function allows for easy extraction of the most important statistics for each network estimate.

```
> summary(G)
Call:
epocG(Y = y, U = u)

Models:
                  R2       Cp        BIC       RSS      links
lambda=1      0.0658 10773.558  220.4464 8642.180        1
lambda=0.8    0.0753 10668.541  136.1846 8554.892        2
lambda=0.64   0.0854 10559.853   64.3761 8461.348        5
lambda=0.512  0.0955 10446.609  -29.3689 8367.350        6
lambda=0.4096 0.1035 10363.954  -71.3253 8293.408       10
lambda=0.3277 0.1116 10295.093  -43.1641 8219.298       21
lambda=0.2621 0.1252 10202.497  127.7470 8093.203       52
lambda=0.2097 0.1463 10090.547  550.1105 7897.490      116
lambda=0.1678 0.1711  9973.619 1109.8054 7668.353      198
lambda=0.1342 0.1953  9882.668 1769.8521 7444.092      290
lambda=0.1074 0.2180  9818.000 2482.3967 7234.743      386

SStot: 9251.292
```

For each λ we obtain the total (across genes) residual sum of squares, `RSS`, and the total error sum of squares corresponding to an empty model, `SStot`. From these, we compute the R^2 statistics `R2`, i.e., the total percentage of the mRNA variation explained by CNAs. The size of the models for each corresponding λ is recorded as the number of network links in `links`.

The `summary` function also generates two model assessment statistics: Mallows C_p (`Cp`) and the Bayesian information criterion BIC (`BIC`). These criteria are commonly used to summarize the tradeoff between the model goodness-of-fit (RSS) and the number of parameters used in the fit (here network size, i.e., the number of links) (see, e.g., [12]). In addition to examining the `summary` output, we can also compare the network fit for different values of λ visually:

```
> plot.modelsel(G)
```

The result is shown in Fig. 37.3. Note, inside `plot.modelsel` we standardize the scales of the C_p and BIC so they can be displayed in the same graph. From the plot we can now identify a candidate region of λ-values that optimizes the criterion of choice, e.g., BIC. We note both from the summary above and Fig. 37.3 that Mallows C_p obtains it minimum for the densest model, which corresponds to penalty parameter $\lambda = 0.1074$. In contrast, the more conservative BIC identifies the `k=5` sparsest model as the best, corresponding to $\lambda = 0.4096$.

We can choose to only view the summary of the best model according to Mallows C_p, BIC, or other chosen network sparsity levels by inputting the optional index parameter `k` into the `summary` function.

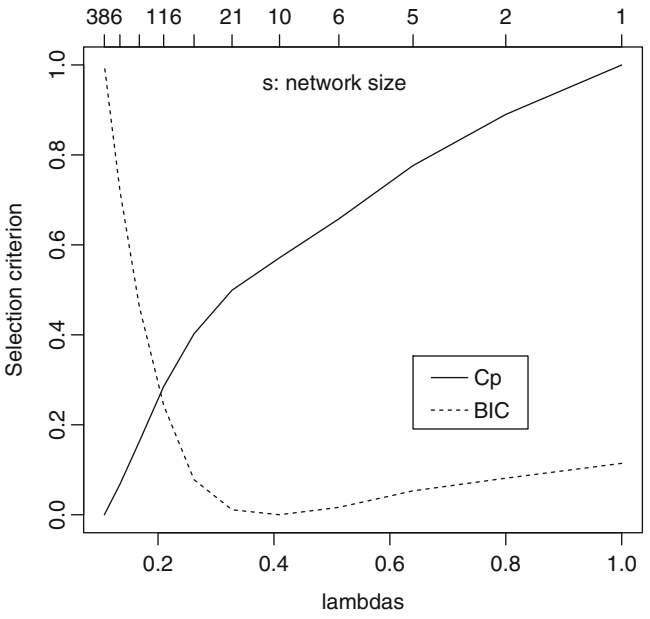


Fig. 37.3 The Mallow's C_p and BIC criteria as functions of the penalty parameter λ. The optimum λ and corresponding network size for each criterion are highlighted with vertical lines

```
> summary(G, k = which.min(G$BIC))
Call:
epocG(Y = y, U = u)

Models:
                    R2         Cp        BIC       RSS  links
lambda=0.4096   0.1035  10363.95  -71.3253  8293.408     10

SStot: 9251.292
```

The `which.min` command extracts the index corresponding to the minimized BIC value (here `k=5`).

Since BIC is minimized for $\lambda = 0.4096$ we explore neighboring penalty values further. In the commands below we first generate a denser sequence of λs centering around the λ that minimizes BIC.

```
> K <- length(G$lambdas)
> uplam <- G$lambdas[max(1, which.min(G$BIC) - 1)]
> lolam <- G$lambdas[min(K, which.min(G$BIC) + 1)]
> lambdas <- seq(lolam, uplam, by = 0.025)
> lambdas
[1] 0.32768 0.35268 0.37768 0.40268 0.42768 0.45268 0.47768
    0.50268
```

We then apply `epocG` for this set of λs. Finally, we output the summary for the network corresponding to the minimum BIC within this finer range λs.

```
> G <- epocG(Y = y, U = u, lambdas = lambdas)

> summary(G, k = which.min(G$BIC))
Call:
epocG(Y = y, U = u, lambdas = lambdas)

Models:
                    R2        Cp         BIC       RSS  links
lambda=0.3777  0.1063  9485.375  -100.2066  8267.693     10

SStot: 9251.292
```

The optimal λ for minimizing BIC is found to equal 0.3777. Note, however, that while C_p and BIC are convenient to use for selection of the penalty parameter, it is well known that C_p tends to overfit (notice in Fig. 37.3 that C_p picked a network with 386 links), and BIC can be too conservative [12]. `epoc` thus includes a separate validation function that uses cross-validation or network concordance as alternative estimation performance indicators. In addition, `epoc` includes a bootstrap procedure to provide more robust network estimates (see sections below).

3.3 Plotting the Network Models

EPoC provides link weight estimates for all networks. `coef(G,k=1)` outputs the links for the most sparse network in the format of a sparse network matrix, and by changing the value of `k` we can extract the denser models. The epoc `plot` command extracts the link weight estimates and displays the network. We can graph any of the estimated network by using the optional index parameter `k` (default `k=1` corresponds to the sparsest model).

```
> plot(G, k = which.min(G$BIC))
```

We here display the model selected by BIC, obtained by using the optional parameter `k=6` (corresponding to $\lambda = 0.3777$) in the `plot` command (result in Fig. 37.4).

While the internal R plotting commands are quite adequate for initial exploration, we recommend that users export network estimation results to Cytoscape [28] for final assessment or when analyzing large (dense) networks. In Cytoscape, we can examine the networks in terms of GO terms and other biological database information. We have made Cytoscape visualization of EPoC results easy by providing an export script, `write.sif`. This function converts the estimated network to a sif-file which can be displayed in, e.g., Cytoscape.

```
> write.sif(G, k = which.min(G$BIC), file = "G.sif")
```

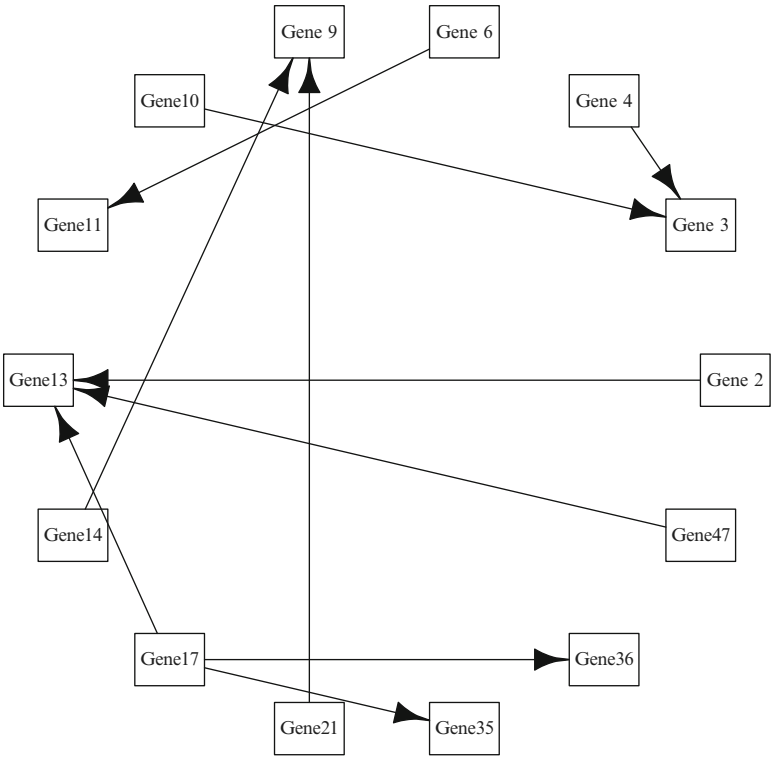


Fig. 37.4 The CNA-driven network, G, for penalty parameter λ minimizing the BIC

3.4 Transcriptional Network Analysis

While EPoC was developed for joint analysis of mRNA and CNA data, we also provide a network estimation function that can be applied to mRNA data only. The function, `epocA`, can be used to estimate *transcriptional networks*, A, as described above, and if no CNA data is available this function generates mRNA–mRNA networks (you simply use the CNA entry empty in the command). The syntax for `epocA` follows that of `epocG`.

```
> A <- epocA(Y = y, U = u)
> summary(A)
> plot(A, k = 3)
```

3.5 Validation

In Sect. 2, we provide an algorithmic overview of EPoC estimation, validation, and robust inference. Here, we illustrate how to use `epoc` to validate the network estimates and select an optimal penalty parameter.

We provide two alternative validation techniques in `epoc`: (1) network concordance using Kendall's W and (2) cross-validation prediction error estimation. These techniques are available as `type` options in the function `epoc.validation`. The validation is performed on B replicates of validation data, generated by random sampling from the original data [15]. Below, we demonstrate how to apply `epoc.validation` to $B = 20$ random subsets of data. Note, as default `epocG` is used as the base function in `epoc.validation`, but by including the optional input `method='A'` you can perform validation for transcriptional network estimation instead.

```
> B <- 20
> W <- epoc.validation(type = "concordance", Y = y, U = u,
                       repl = B)
> P <- epoc.validation(type = "pred", Y = y, U = u, repl = B)
```

The `epoc` package includes plotting commands for visual validation. Below, we plot Kendall's W concordance versus λ.

```
> plot(W)
```

The generated graph displays a smooth fit (`loess`) of W on λ, as well as associated standard error bands (Fig. 37.5). By using a smooth function estimate of the validation criterion's dependency on λ we can use interpolation to estimate the optimum λ. This can make validation considerably faster since fewer values of λ need to be directly compared in the validation procedure. We identify the maximum concordance, W^*, (marked with a horizontal line) in the figure. Finally, to select the optimal value for λ, λ^*, we identify the smallest λ such that the corresponding standard error band contains W^*. We also include a conversion of λ-values to network sizes s in the plot, located on the top axis. These network sizes are obtained from a loess fit of the network sizes from each randomly sampled data set as a function of λ. The optimal network size, s^*, is obtained from the fitted value of this `loess` fit evaluated at $\lambda = \lambda^*$. We note here that the network concordance is optimized for $\lambda^* = 0.4174$ with the corresponding network size $s^* = 17$.

Similarly, using validation `type=''pred"` and applying the `plot` command to the `epoc.valdation` results we obtain a smooth fit of cross-validation (CV) errors as a function of λ.

```
> plot(P)
```

Standard error bands are added and the optimal $\lambda = \lambda^*$ is identified as the smallest λ such that the corresponding error band contains the minimum CV error (Fig. 37.6). The corresponding optimal network size is also provided. As noted before in [15], networks optimized for CV error tend to be larger than networks

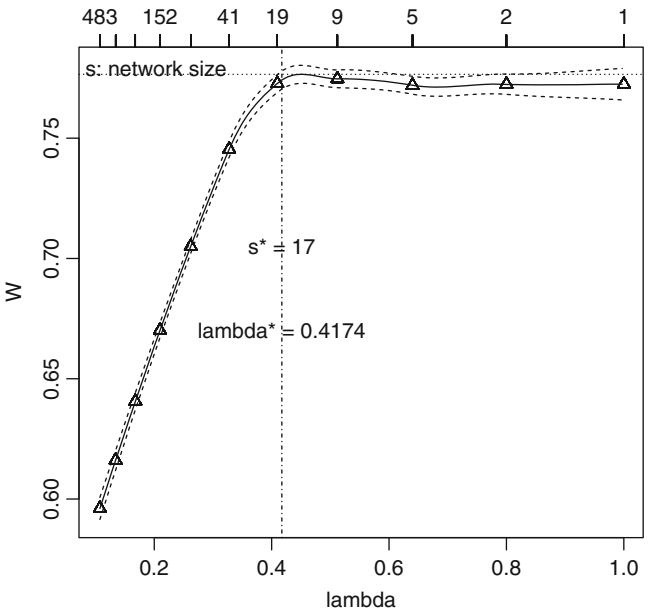


Fig. 37.5 Concordance validation: Kendall's W as a function of λ. The optimum λ and corresponding network size are indicated with a vertical line

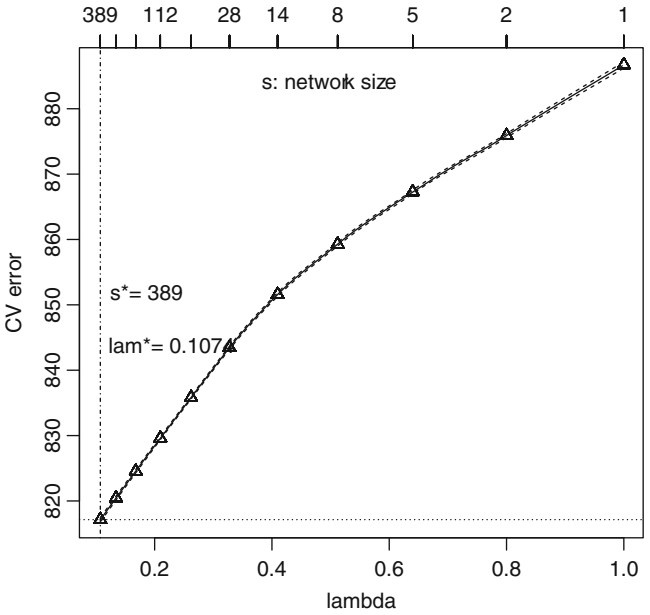


Fig. 37.6 Cross-validation: The CV error, P, as a function of λ. The optimum λ and corresponding network size are indicated with a vertical line

optimized for concordance. Here, the CV error optimized networks are obtained for $\lambda^* = 0.1074$, corresponding to network size $s^* = 389$.

We can access the optimal λ-value (output `lopt` from `epoc.validation`) or optimal network size (output value `sopt`) directly from the `epoc.validation` without plotting (though we strongly recommend that users plot to assess the stability of estimation as a function of the penalty parameter).

```
> W$sopt
[1] 17.2488
> W$lopt
[1] 0.4173742
> P$sopt
[1] 388.5226
> P$lopt
[1] 0.1073742
```

For our small, synthetic data set, we note that C_p and the CV error select similarly sized networks, whereas the concordance statistics and BIC performed quite similarly, though we have noted on real data that these four criteria can differ substantially.

Once the optimal penalty parameter λ^* has been identified (denoted `lopt` in the code above), we can generate a network for the original data set as follows:

```
> G.opt <- epocG(Y = y, U = u, lambdas = W$lopt)


> summary(G.opt)
Call:
epocG(Y = y, U = u, lambdas = W$lopt)

Models:
                     R2     Cp       BIC       RSS  links
lambda=0.4174   0.1028   9442  -63.951   8299.987     10


SStot: 9251.292
```

The estimated optimal parameter values `W$lopt` is not guaranteed to provide a network of size `W$sopt` on the original data as this can depend heavily upon relative sample size between the original and validation data sets, as well as data signal correlation and noise. We thus recommend that users search a range of λ near `W$lopt` that when applied to the original data results in a network of the optimum size `W$sopt`.

```
> ll <- seq(0.3,0.7, by = 0.02)
> G.opt <- epocG(Y = y, U = u, lambdas = ll)
> summary(G.opt)
```

For our synthetic data, we choose $\lambda = 0.34$ which produces a network of size $s = 18$ matching closely the desired network size `W$sopt` which can also be found by examining Fig. 37.3. This figure contains the map from penalty parameters λ to

network size s (top axis) for the original data. From the validation procedure above we find the optimum network size for either network concordance (`W$sopt` = 17) or mRNA prediction (`P$sopt` = 389). A reverse mapping from network size to λ from Fig. 37.3 identifies the optimum penalty parameter value $\lambda = 0.34$ for concordance and $\lambda = 0.11$ for mRNA prediction. (Note, as the validation is performed on randomly resampled sets of data (here $B = 20$ sets) results obtained by users may vary slightly from those documented here.)

3.6 Estimating a Robust Network

To provide robust network estimates, we perform network estimation of pseudo-bootstrap samples [15]. For a given λ, e.g., the estimated optimal penalty parameter from above, we repeatedly generate bootstrap samples and estimate the corresponding networks. We aggregate the generated networks across bootstrap samples and record the proportion of times (across bootstraps) that each link is identified by EPoC. This bootstrap procedure has been implemented in the R function `epoc.bootstrap` which takes as input the data, the number of bootstrap samples, and the identified optimum λ-value.

```
> G.boot <- epoc.bootstrap(Y = y, U = u, nboots = 100,
      method = "epocG", lambda = 0.34)
```

The user can explore the impact of different values of λ near the optimum by providing a vector of values of interest (`lambda.boot` below). The plotting command `plot.bootsize` enables the user to visualize the impact of both values of λ and the bootstrap threshold on the network size.

```
> lambda.boot <- c(0.33,0.34,0.35)
> G.boot <- epoc.bootstrap(Y = y, U = u, nboots = 100,
       method = "epocG", lambda = lambda.boot)
> plot.bootsize(G.boot, lambda.boot, 100, range = c(.05,.7))
```

In Fig. 37.7 we see that the network size stabilizes beyond a threshold of about 30% link appearance frequency (NB: the axis for link prevalence is on a log-scale). If we use this as a cutoff for a link to be retained in the model, we obtain a final network model of size $s = 16$.

We provide a plotting command for `epoc.bootstrap`. We first extract the final network from the bootstrap result using the function `epoc.final`. Input values are the bootstrap estimates `G.boot`, the index of the chosen λ (here `k=2` since $\lambda = 0.34$ is the second element in `lambda.boot`, and finally the threshold value `bthr` (here 30%).

```
> G.final <- epoc.final(G.boot, k = 2, bthr = 0.3)
> epoc.bootplot(G.final)
```

The final network is displayed in Fig. 37.8. (Note, as the bootstrap procedure is random, results obtained by users may vary slightly from those documented here.

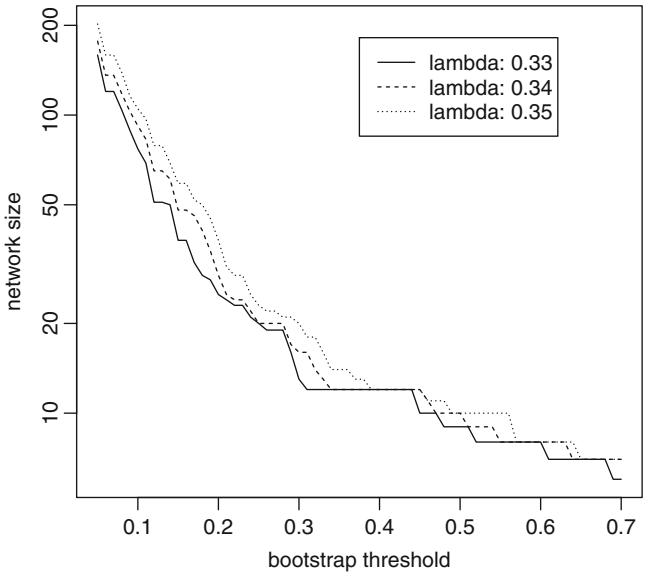


Fig. 37.7 The estimated size network at different bootstrap thresholds

Internal consistency is improved with a larger set of bootstrap samples (`nboots`) and a higher link prevalence threshold (`bthr`) to produce the final network.)

3.7 Survival Scoring

If survival data are available, `epoc` provides the user with functions that reduces the estimated networks to survival scores. As described above, we generate survival scores based on the singular value decomposition (SVD) of the network matrix (G for `epocG` CNA-driven networks, or A for `epocG(method='A')` transcriptional networks). We first generate a sparse SVD decomposition of the network using the function `epoc.svd`.

```
> G.svd <- epoc.svd(G.final, C = 3, numload = c(10, 10, 10))
> print(G.svd)
```

By default, `epoc.svd` outputs only the leading SVD components, but more components can be generated using the optional parameter `C`. The sparsity of the SVD components is controlled via the input parameter `numload`. Values of `numload` determine the number of genes that contribute to each component. To aid in the interpretation of the survival scores we choose to use a small number of genes, e.g., 10 as above, but users can also choose to optimize this number for survival scoring (see below). The output from `epoc.svd` consists of the sparse loadings for the input `G.svd$spload.in` and output `G.svd$spload.out`. If the network

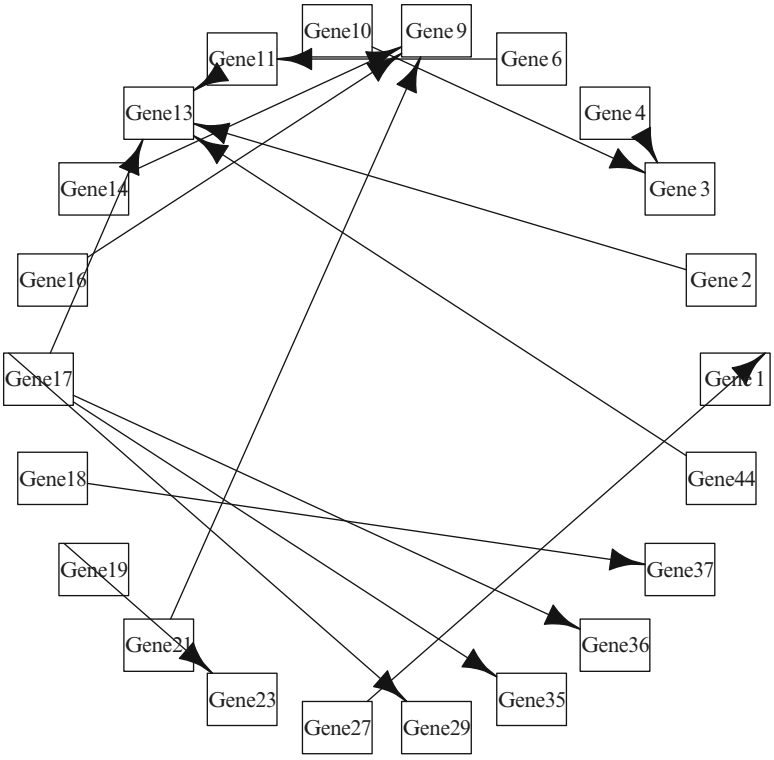


Fig. 37.8 The final estimated network using bootstrap and a 30% threshold

`G.final` above is an `epocG` object, the input is CNA and the output mRNA. If it is an `epocA` object, the reverse is true. If no CNA is provided, both input and output are mRNAs. The `print` command outputs the sparse loadings of the SVD decomposition of the network matrix as well as the list of genes which are present in at least one of the sparse input or output components. For the synthetic data set, only nine out of the 50 genes are present among the first SVD components (object `G.svd$ii` provides this gene list).

The `epoc.svdplot` function displays the subgraph of the `G.final` network above. Only links connecting genes that contribute to either the input or output SVD components are shown, and we thus identify the genes whose CNA are most amplified by the system (listed in `G.svd$sploa.in`) and those mRNA that are main responders in the system (`G.svd$sploa.out`).

```
> epoc.svdplot(G.svd, C = 1)
```

Note in Fig. 37.9, obtained from the first SVD components, that e.g., Gene 13 is a main mRNA responder and perturbation of e.g., Gene 17 is strongly amplified by the system (compare Fig. 37.9 to full system network in Fig. 37.8).

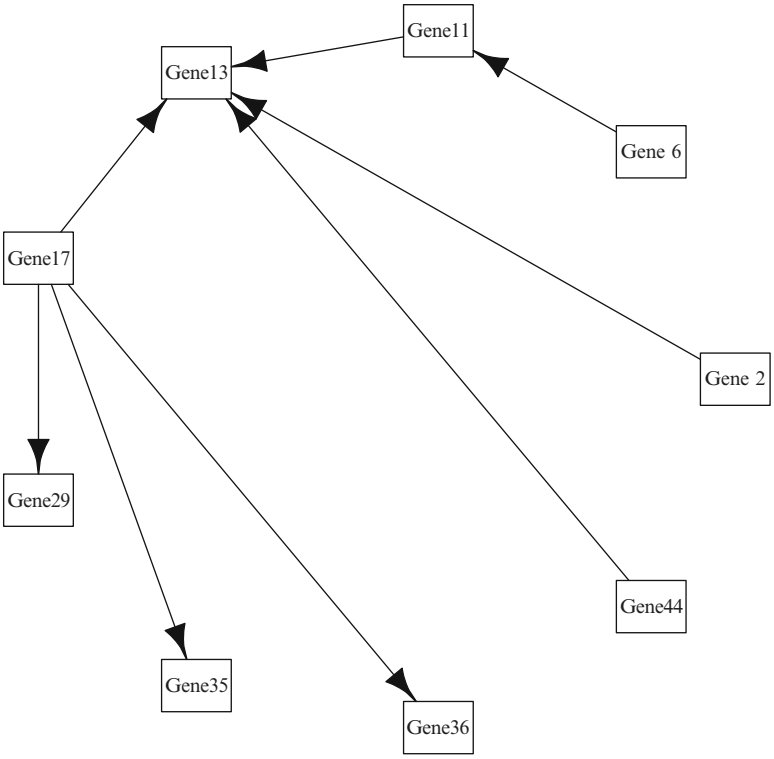


Fig. 37.9 The sub-graph corresponding to the leading SVD in- and out-components. CNAs amplified, mRNA responders

The function `epoc.survival` uses the SVD decomposition components to perform survival comparisons. This function takes as input the selected sparse SVD input and output component and the survival data of the patients in the data set.

```
> surv <- synth$surv
> G.surv <- epoc.survival(G.svd, y, u, surv, C = 1, type = "G")
```

By default, `epoc.surv` applies survival analysis using the first SVD component, but other components can also be used by changing the input value of `C`. Survival scores are generated as described in Subsect. 2.4. A simple non-parametric survival analysis is performed, comparing survival between patients with positive or negative scores (tumor fitness). The `epoc.survival` object contains the summary information from a log-rank test comparing survival (R function `survdiff`) and survival fit objects.

```
> summary(G.surv)

            In
Call:
survdiff(formula = Surv(surv) ~ sign(sc.in))
```

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
sign(sc.in)=-1 93       93     74.5      4.57      7.91
sign(sc.in)=1  93       93    111.5      3.06      7.91

 Chisq= 7.9  on 1 degrees of freedom, p= 0.00493

           Out
Call:
survdiff(formula = Surv(surv) ~ sign(sc.out))

                 N Observed Expected (O-E)^2/E (O-E)^2/V
sign(sc.out)=-1 104      104     80.6       6.8        13
sign(sc.out)=1   82       82    105.4       5.2        13

 Chisq= 13  on 1 degrees of freedom, p= 0.000317
```

For the synthetic data set, we note that both the first input (CNA) and output (mRNA) components can stratify patients in terms of survival. We can also summarize this graphically using the `plot` command which outputs regular Kaplain–Meier curves for patients stratified by positive and negative projection scores onto the SVD components.

```
> plot(G.surv)
```

In Fig. 37.10 we see that those patients with positive scores (solid lines) have lower survival than those with negative scores (dashed), indicating the patients whose CNA and mRNA profiles agree with the network model have poorer prognosis.

3.8 Summary

The `epoc` package can be used for network model exploration using the basic `epocG` function. The internal validation criteria (C_p and BIC) can be used to identify a candidate region of penalty parameter values and network sizes. Cross-validation or network concordance are used to better assess optimal network sizes (`epoc.validation`). We recommend final robust network estimation to retain links in the network that are consistently identified across many bootstrap datasets (`epoc.bootstrap`). Finally, estimated networks can be summarized in terms of input and output (CNA drivers and mRNA responders) using a sparse SVD decomposition (`epoc.svd`). When survival data are available, the networks can be examined for potential patient stratification (`epoc.survival`).

The `epoc` package is still under development. We hope to expand the package to include more visualization and validation options, as well as allow for survival analysis with multiple factors and variables. New and updated package components and their descriptions will be distributed through the package web page http://sysbio.med.gu.se/epoc.html.

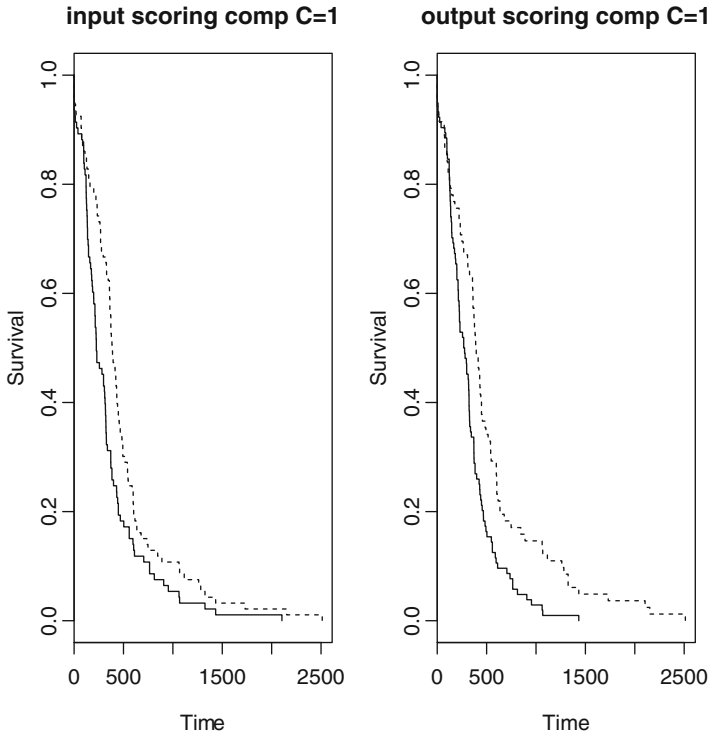


Fig. 37.10 Survival curves for input and output scores with patient groups corresponding to positive and negative scores, respectively

4 Other Packages for Modeling CNA–mRNA Effects

In this section we give an overview of key alternatives to EPoC modeling, also underlining factors such as algorithmic principle, performance, and speed. One common approach to large-scale transcriptional modeling, is to derive models from mRNA profiles only, using e.g., gene–gene (partial) correlations, Bayesian networks, ordinary differential equations, or mutual information [3, 7, 20, 22, 27]. Useful alternative to EPoC include:

- *Partial correlation methods*, such as `glasso` [6] and GeneNet, estimate the (sparse) inverse correlation matrix from a set of data which reflects the direct dependencies. In contrast, the correlation matrix itself includes both direct and indirect interactions. The inverse correlation matrix is related to our A network as follows. Under our model formulation, $\Delta Y \simeq -A^{-1}\Delta U$, and so the correlation matrix of the mRNA expression levels $\Sigma_{YY} = A^{-1}\Sigma_{UU}(A^{-1})^T$, or equivalently, the inverse correlation matrix $\Sigma_{YY}^{-1} = A\Sigma_{UU}^{-1}A^T$. The estimate of Σ_{YY}^{-1} thus generates an undirected version of A.

- *ARACNE* uses a mutual information criterion to identify directly dependent transcripts and distinguish them from those dependent only through other transcripts [20].

A second approach, used in "genetical genomics", is to use the naturally occurring genetic variation in a separating population to study the relationship between genotype and expression phenotype [13, 18, 19, 25, 33, 39]. Useful alternatives to EPoC in this category of tools include:

- *eQTL* is a standard class of methods to associate SNPs to mRNA levels (e.g., [31, 32]). These methods simply involve calculation of all pairwise linear regression models between genotypes and mRNAs, and correction of the nominal p-values to obtain a smaller set of links.
- *remMap*, introduced in [23] is a method for combined CNA–mRNA analysis that involves several steps. First, the CNA data is converted to CNA-intervals using fixed order clustering. Second, a set of mRNA–mRNA interactions are identified by running a partial correlation analysis of the mRNA data (similar to `glasso`). Third, for each mRNA transcript a model is built based on other transcripts and CNA interval data using elastic net regression techniques. The method can therefore be thought of as a hybrid of EPoC *A*/*G* analyzed for genomic regions.
- *LirNet* [18] is designed to derive a transcriptional module network from combined SNP and mRNA data. Given a set of transcript clusters and a set of possible regulators, this algorithm identifies SNP and mRNA regulators for each cluster by elastic net regression (combined lasso-type and standard least squares penalties). An additional feature of the algorithm is that the lasso penalties can be learnt from annotation features (e.g., the position of the SNP inside the gene).

Applying these methods to the same set of data (glioblastoma data from the Cancer Genome Atlas project), see [15], we see drastic differences in terms of method robustness (Fig. 37.11a), matching pathways (Fig. 37.11b), and in terms of speed (Table 37.1). We see generally strong performance of EPoC and glasso, which are both robust and achieve a higher overlap with known pathway interactions than the other methods. Two key differences between EPoC and glasso, however are speed (results for 500 gene networks in Table 37.1, the difference is even more accentuated for larger networks), and the fact that EPoC links are directed whereas glasso links are not. In addition, it is not clear how to extend glasso to incorporate multiple data types like CNA and mRNA. From these comparisons, we conclude that EPoC exhibits excellent performance in terms of model reproducibility, validity, and algorithmic speed.

We compare the estimated network structures by each method using hierarchical clustering, with the concordance measure Kendall's W as the metric of similarity. The result is shown in Fig. 37.12. We see a separation of the methods into two groups: those coupling CNA and mRNA data (LirNet, remMap, eQTL, EPoC-G) and those that derive networks from mainly mRNA–mRNA

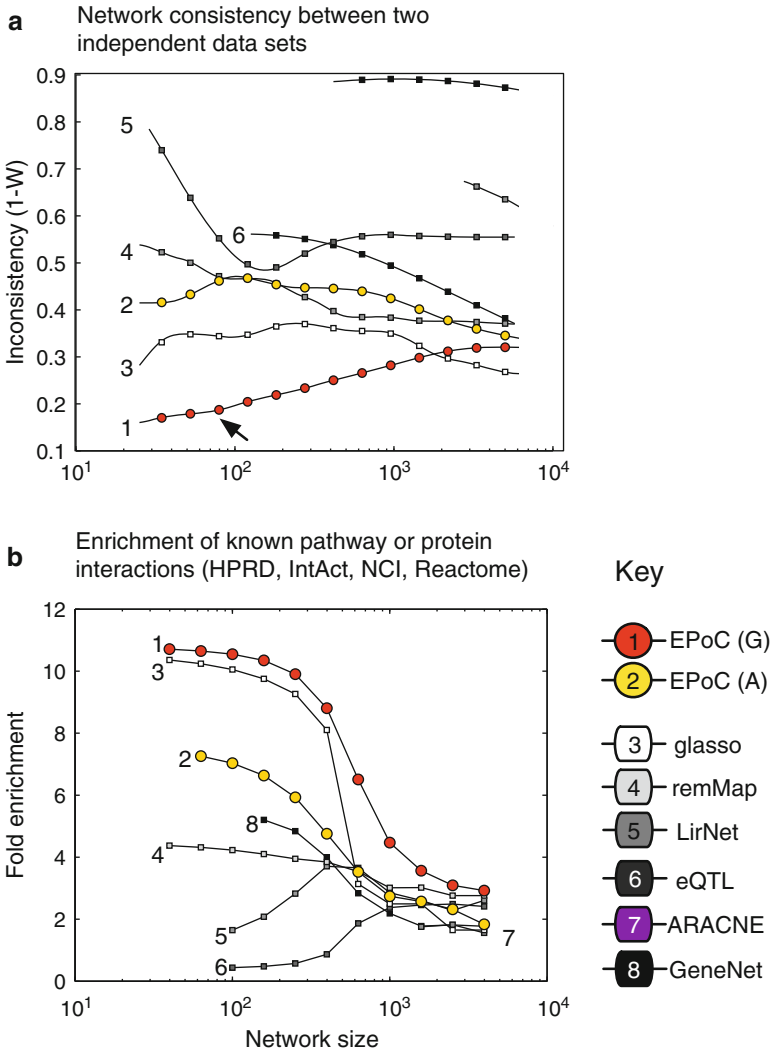


Fig. 37.11 Method comparisons: network consistency and pathway interactions. (**a**) In a first analysis, we test each method's reliability, i.e., its robustness to noise and technological factors. For this, we compare network models derived from two full replicate glioblastoma datasets (146 identical tumors (same patients and samples) but processed at different centers with slightly different technological setups (Affymetrix and Agilent technologies, run at MSKCC, Harvard Medical School and Broad Institute). We subsequently measure the inconsistency $(1-W)$ between the two models. In this test, EPoC estimation of the CNA-driven network G is the best performing method on the TCGA data ($1-W$ lower, arrow $\nearrow$). `glasso` is second best, followed by sparse estimation of the transcriptional network A (EPoC A), and `remMap`. LirNet, eQTL, GeneNet, and ARACNE all exhibit less robust performance compared with EPoC G.

Table 37.1 Methods for modeling CNA/mRNA data, main characteristics

Method	Type	Principle	Speed	Software
EPoC	CNA + mRNA	Regression	3–6 s	R
ARACNE	mRNA cleavage	Information theory	100–200 s	C
glasso	mRNA	Sparse inverse correlation	15–60 s	R
GeneNet	mRNA	Inverse correlation	3–6 s	R
remMap	CNA + mRNA	Clustering, sparse inverse correlation, and elastic net regression	40–60 s	R
LirNet	CNA + mRNA	Clustering, elastic net regression	10–20 s	MATLAB

All run times were obtained on a desktop computer, Mac Pro, 2 × 2.8 GHz quad-core Intel Xeon, for a set of 500 gene problems

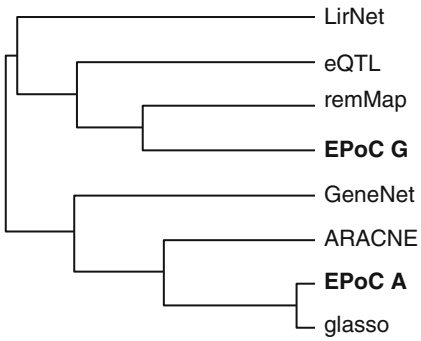


Fig. 37.12 Comparison of networks obtained by different methods. Hierarchical clustering of network solutions (single linkage, 1-fractional network overlap as distance); note that EPoC *A* networks group with transcriptional network methods (ARACNE and GeneNet) and EPoC *G* groups with `remMap` and similar genotype-based methods

dependencies (ARACNE, GeneNet, glasso, EPoC-A). Viewed together with the results in Fig. 37.2, we see that mRNA–CNA and mRNA–mRNA methods complement each other and identifies different dependency structures in the data.

Fig. 37.11 (continued) (**b**) In a second analysis, we test each method's ability to re-discover pathway and PPI links present in databases. For this, we map interactions, found by EPoC and other methods, to molecular links in the pathway repositories HPRD, Reactome, Intact, and NCI-nature. Each interaction is characterized by the number of steps minimally needed to "walk" between the network gene and its target (i.e., the shortest path). A well-estimated network, in our view, should be comprised of identified interactions that either match known interactions in the data bases or are enriched for shorter paths [14]. The figure depicts the enrichment, defined as the relative proportion of interactions that correspond to a shortest path length of 1 or 2 interactions in a pooled network based on the four different pathway databases. EPoC *G* interactions are clearly enriched for short or direct paths in the data bases, followed by `glasso` and EPoC *A*

5 Perspectives

In the light of ongoing efforts currently being undertaken to acquire comprehensive genome-scale data sets for several cancer types (e.g., the Cancer Genome Atlas, and the International Cancer Genome Consortium), meaningful analysis of the data becomes a major challenge. We argue that a priority should be to develop mechanistically and clinically relevant molecular network models of the data. EPoC is one step in this direction, and helps to set the stage for the continued modeling efforts in the context of human cancer genome programs. Future work should be directed at generalizing our methods to enable comparison of regulatory networks between sets of patients, and to make optimal use of all available data (e.g., methylation and miRNA profiles). These extensions will enable more exact and comprehensive model-based analysis of the complex molecular landscape of cancer.

Acknowledgments The authors thank the editors and reviewer for their constructive comments. This project receives funding from Cancerfonden, Barncancerfonden (NB-CNS consortium), Vetenskapsradet (SN,RJ), BioCare (SN).

References

1. Adler AS, Lin M et al (2006) Genetic regulators of large-scale transcriptional signatures in cancer. Nat Genet 38:421–430
2. Akavia UD, Litvin O et al (2010) An integrated approach to uncover drivers of cancer. Cell 143:1005–1017
3. Bansal M, Belcastro V et al (2007) How to infer gene networks from expression profiles. Mol Syst Biol 3:78
4. Fisher R (1926) The arrangement of field experiments. J Ministry Agric Great Britain 33: 503–515
5. Friedman J, Hastie T et al (2007) Pathwise coordinate optimization. Ann Appl Stat 1:302–332
6. Friedman J, Hastie T et al (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
7. Friedman N, Linial M et al (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620
8. Fu WJ (1998) Penalized regressions: the bridge versus the lasso. J Comput Graph Statist 7: 397–416
9. Garraway LA, Widlund HR et al (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 436:117–122
10. Golub GH, Loan CFV (1996) Matrix computations. Johns Hopkins University Press, Baltimore, MD, USA
11. Haslinger A, Schwarz TJ et al (2009) Expression of Sox11 in adult neurogenic niches suggests a stage-specific role in adult neurogenesis. Eur J Neurosci 29:2103–2114
12. Hastie T, Friedman J et al (2009) Elements of statistical learning, 2nd ed. Springer Verlag. Corr. 3rd printing 5th Printing, Springer-Verlag, New York
13. Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145–151
14. Johnson D (1977) Efficient algorithms for shortest paths in sparse networks. J Acm 24:1–13

15. Jörnsten R, Abenius T et al (2011) Large-scale network modeling and prognostic scoring of the effects of DNA copy number aberrations on gene expression in glioblastoma. Mol Syst Biol. Nature Publishing Group, 1(7)
16. Kendall MG, Smith BB (1939) The problem of *m* rankings. Ann Math Stat 10:275–287
17. Kim YA, Wuchty S et al (2010) Simultaneous identification of causal genes and dys-regulated pathways in complex disease. Res Comput Mol Biol (RECOMB) 6044:263–280
18. Lee SI, Dudley AM et al (2009) Learning a prior on regulatory potential from eQTL data. PLoS Genet 5:e1000358
19. Lee SI, Pe'er D et al (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci USA 103:14062–14067
20. Margolin AA, Nemenman I et al (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7
21. Nordling TEM, Jacobsen EW (2009) Interampatteness – a generic property of biochemical networks. IET Syst Biol 3(5):388–403
22. Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol 1:37
23. Peng J, Zhu J et al (2010) Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. Ann Math Stat 53–77
24. Piccirillo SGM, Binda E et al (2009) Brain cancer stem cells. J Mol Med 87:1087–1095
25. Rockman MV (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. Nature 456:738–744
26. Savageau MA (1976) Biochemical systems analysis : a study of function and design in molecular biology; with a foreword by Robert Rosen. Advanced Book Program Addison-Wesley Pub Co, Addison-Wesley Reading, MA, USA
27. Schäfer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21:754–764
28. Shannon P, Markiel A et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
29. Shi Y, Sun G et al (2008) Neural stem cell self-renewal. Crit Rev Oncol Hematol 65:43–53
30. Skogestad S, Postlethwaite I (1996) Multivariable feedback control: analysis and design? Wiley, Chichester and New York
31. Stranger BE, Forrest MS et al (2007a) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853
32. Stranger BE, Nica AC et al (2007b) Population genomics of human gene expression. Nat Genet 39:1217–1224
33. Suthram S, Beyer A et al (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. Mol Syst Biol 4:162
34. TCGA-Consortium (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061–1068
35. Tegner J, Yeung MKS et al (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci USA 100:5944–5949
36. Troyanskaya O, Cantor M et al (2001) Missing value estimation methods for DNA microarrays Bioinformatics 17(6):520–525
37. Verhaak CPRG, Hoadley KA et al (2009) Reproducible Gene Expression Subtypes of Glioblastoma Show Associations with Chromosomal Aberrations Gene Mutations, and Clinical Phenotypes. Manuscript
38. Witten DM, Tibshirani R et al (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10:515–534
39. Zhu J, Zhang B et al (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40:854–861
40. Zou H, Hastie T et al (2006) Sparse Principal Component Analysis. J Comput Graph Stat 2:262–286

Chapter 38
Early Patient Stratification and Predictive Biomarkers in Drug Discovery and Development

A Case Study in Ulcerative Colitis Anti-TNF Therapy

Daphna Laifenfeld, David A. Drubin, Natalie L. Catlett, Jennifer S. Park, Aaron A. Van Hooser, Brian P. Frushour, David de Graaf, David A. Fryburg, and Renée Deehan

Abstract The current drug discovery paradigm is long, costly, and prone to failure. For projects in early development, lack of efficacy in Phase II is a major contributor to the overall failure rate. Efficacy failures often occur from one of two major reasons: either the investigational agent did not achieve the required pharmacology or the mechanism targeted by the investigational agent did not significantly contribute to the disease in the tested patient population. The latter scenario can arise due to insufficient study power stemming from patient heterogeneity. If the subset of disease patients driven by the mechanism that is likely to respond to the drug can be identified and selected before enrollment begins, efficacy and response rates should improve. This will not only augment drug approval percentages, but will also minimize the number of patients at risk of side effects in the face of a suboptimal response to treatment. Here we describe a systems biology approach using molecular profiling data from patients at baseline for the development of predictive biomarker content to identify potential responders to a molecular targeted therapy before the drug is tested in humans. A case study is presented where a classifier to predict response to a TNF targeted therapy for ulcerative colitis is developed a priori and verified against a test set of patients where clinical outcomes are known. This approach will promote the tandem development of drugs with predictive response, patient selection biomarkers.

1 Introduction

Though our abilities to measure and analyze large amounts of complex data have increased significantly over the past decade and have provided valuable

D. Laifenfeld • D.A. Drubin • N.L. Catlett • J.S. Park • A.A. Van Hooser • B.P. Frushour • D. de Graaf • D.A. Fryburg • R. Deehan (✉)
Selventa, One Alewife Center, Cambridge, MA 02140, USA
e-mail: dlaifenfeld@selventa.com; ddrubin@selvanta.com; ncatlett@selventa.com; jpark@selventa.com; avanhooser@selventa.com; bfrushour@selventa.com; dfryburg@selventa.com; rkenney@selventa.com

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_38,

insight into the molecular mechanisms underlying disease, the industry as a whole is lagging in the production of new and innovative therapies. Multiple studies reference the extremely high failure rate (>80%), the length of time to develop (10–15 years through Phase III), and the high cost (at least $800 million) of new therapies [1–3]. A substantial part of this expenditure is attributed to the cost of those projects (investigational drugs) that failed. Phase II, in which efficacy is usually first tested in patients, is the stage of drug development exhibiting an extremely high failure rate. Across multiple therapeutic mechanisms, approximately 80% of novel projects that reach Phase II fail to demonstrate clinically significant efficacy [1]. This emphasizes the need to select the patients most likely to respond to treatment for entry into clinical trials.

It has long been recognized that some patients may respond well to a particular intervention, whereas others may gain little or no benefit. As diseases are classically characterized by their phenotype and not always sub-categorized by the specific mechanisms or genotypes contributing to the phenotype, applying a focused molecular targeted therapy may not be effective in most patients, thus obscuring the benefit to a responder sub-population. For example, although glucose elevation defines the diagnosis of diabetes, it does not explain what pathophysiology caused the glucose to be elevated, nor does it suggest the treatment mechanism by which it could be lowered. Or, in the case study presented here, is ulcerative colitis driven by the same mechanism in all patients (and hence should all receive the same treatment)? Although one possibility for efficacy failure in a group of classically defined patients could be that the investigated mechanism is altogether irrelevant to the disease, an alternative is that there are molecular subpopulations of patients, some of whom might be sensitive to a highly specific and directed therapy. Potentially valuable therapies are likely failing in some cases due to uninformed patient selection.

Ideally, the responsive patient population within a disease group would be identified with the help of predictive biomarkers before enrollment in a clinical trial. However, the current paradigm to develop such biomarkers suffers from the dependency on available datasets bridging potential biomarker measurements with clinical outcome. Significant patient numbers to develop these correlative biomarkers are not available until after a phase II or III clinical trial, at which point millions of dollars have been spent on a program that could fail due to a lack of efficacy. Establishing predictive biomarkers in pre-clinical phases, before outcomes data become available, is a critical factor for selecting the patients most likely to respond to the drug and therefore improving success.

Systems biology, which focuses on complex interactions in biological systems, moving away from a reductionist, hypothesis-driven approach, holds the potential to address the above mentioned key challenges in identification of predictive biomarkers. This approach does not focus on a limited number of molecular components but rather achieves a comprehensive understanding of how large numbers of interrelated components of a system comprise networks whose functional properties emerge as definable phenotypes, using complex, rich data as a substrate. In the context of a patient population, this means that through a systems biology approach, one can define patients not through their phenotype, but rather via the molecular networks that underlie them.

In this review, we present a systems biology approach that exploits high-throughput data collected from diseased patients *before treatment* to develop predictive biomarkers in tandem with drugs as early as the pre-clinical stage. This approach may in many cases allow for a better in-depth understanding of human disease biology, improved success rate, and improved translatability from pre-clinical to clinical studies. Specifically, we will review the following as they relate to patient stratification: (1) the use of biomarkers within the current paradigm of drug development, (2) an approach where high-throughput patient data are used to generate mechanistic biomarkers to predict the most likely candidates for response to treatment, and (3) a case study where a classifier was developed to predict response to a TNF-targeted therapy in ulcerative colitis patients without prior knowledge of clinical outcomes using transcriptomic data from colon biopsies.

2 Challenges Associated with Predictive Biomarker Development Using Clinical Outcomes Data

In addition to selection of the right mechanism to target within a disease population, it is critical that we also select the right patient for targeted therapy treatment. Even a population of patients that appears to be phenotypically similar can exhibit distinct molecular disease profiles due to differences in etiology, environmental factors, co-morbidities, or genetics. For example, in a disease like atherosclerosis, there are multiple elements that may contribute to the observed burden of disease and eventual myocardial infarction (e.g., inflammation, lipid metabolism, anatomic alterations, etc.). The same is also true for many malignancies. A similar clinical diagnosis, therefore, may be the integrated result of multiple molecular disease-driving mechanisms.

Thus, the patient population in a clinical trial for a targeted therapy often represents multiple disease subsets driven by different molecular mechanisms, only a subset of which will respond to a very specific, molecularly-focused treatment. There are several examples of how biomarkers are used to identify the likely-to-respond subjects, best exemplified in oncology. In breast cancer, immunohistochemical detection of the estrogen receptor is used to predict the efficacy of tamoxifen or aromatase inhibitors, and HER2 gene amplification is a positive indication for use of anti-HER2 treatment such as trastuzumab (Herceptin) [4, 5]. In colon cancer, k-ras mutations predict resistance to anti-EGFR therapies [6]. Distinct biomarkers such as these that provide a specific patient stratification are currently packaged as companion diagnostics for targeted therapies, enabling the selection of patients that have a greater chance of responding to receive the drug. As a result, companion diagnostics are currently accepted and even mandated by regulatory agencies [7]. Selecting the patient pool most likely to respond has proven beneficial for obtaining regulatory approval of effective drugs, for example, herceptin and gefitinib [8]. Importantly, in the absence of the ability to select the right patients prior to enrollment, the efficacy of these drugs may have been masked

by a cohort of patients that, while clinically similar, were heterogeneous with respect to disease etiology and pathogenesis, and an unstratified patient population would have yielded a lackluster response to the molecularly precise drug. Lackluster responses may often lead to termination of a program, and a potentially effective approach for some patients will have been discarded.

Given the advantages of patient stratification for both treatment protocol design and targeted therapeutic efficacy, proactively applying patient stratification as early as possible in the drug development paradigm is critical. The current use of patient stratification has too often arisen as a reactive solution to the problems of patient heterogeneity and drug resistance wherein markers of effective response were assessed only subsequent to extensive characterization of clinical trial data (i.e., after a costly Phase II clinical trial). Moreover, if a biomarker strategy to identify likely responders is currently employed, it most often will depend on a single biomarker directly associated with the therapeutic target rather than a more robust multiple-biomarker signature of target activity.

Recent use has been made of high-throughput data such as gene expression and proteomics to add granularity to the biomarker approach. While this post hoc application of large-scale patient data is being realized, the value of the data will be even greater when utilized as the substrate for patient stratification at early drug development phases. Wielding high-resolution readouts of patient biology a priori would facilitate a more proactive approach to drug development, one that leverages patient data directly for drug target discovery and an increased potential for clinical success. Additionally, the identification of candidate therapeutic targets and biomarkers predictive of response would effectively be co-indicated at the pre-clinical stage, affording the ability to pre-select an optimal patient population for response prior to the initiation of clinical trials. However, realization of such an a priori strategy requires a means of effectively interpreting this patient data in terms of the specific biological mechanisms active in individual patients.

3 A Methodology to Develop Predictive Biomarkers in the Absence of Relevant Clinical Outcomes Data

To meet the demand for a priori patient stratification, we developed a unique strategy to develop predictive biomarkers that identify likely responders for a targeted therapeutic within a population of patients across multiple disease areas (e.g., cancer, inflammatory disease, etc.) based upon mechanistic inferences from patient data. The strategy relies upon the hypothesis that patient groups that exhibit either high levels or low levels of target mechanism signaling strength will be more likely to respond to treatment with that targeted therapeutic. This approach can be implemented using molecular profiling data, such as whole genome expression data, from diseased patients at baseline, and does not require a priori knowledge of treatment outcomes. The substrate for this strategy is our causal knowledge base that has stored gene expression signatures of over 2,000 biological perturbations from

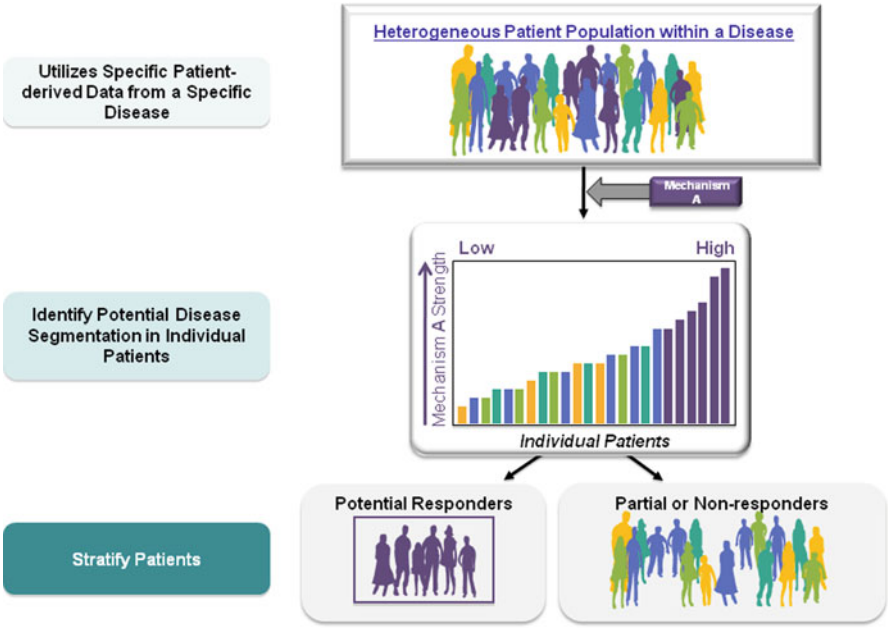


Fig. 38.1 Patients can be stratified by target mechanisms to identify likely responders. The non-responder population can be further stratified to identify disease-driving mechanisms unique to different sub-groups of patients. This elucidates targets that can potentially treat each sub-group, and biomarkers can be co-developed to identify subgroup members using the signaling strength approach

over 46,000 peer-reviewed publications. Each of these mechanisms can represent a potential driver of disease. The perturbation, or "signaling strength" of each mechanism, can be assessed in individual patients within a population. For example, a gene expression signature for MAPK13 activity, based on prior knowledge, can be extracted from our knowledge base. Fold changes in gene expression are calculated for each patient as compared to a common baseline like a non-diseased population or a median patient, and a strength assessment algorithm that takes the hypergeometric mean of the fold change for each gene in the signature of interest is applied. The output of this assessment is a quantitative value that enables the group of patients to be stratified by their levels of signaling strength for each of the 2,000+ mechanisms. Patient stratification by signal strength allows identification of those mechanisms that are most strongly or weakly activated in different subsets of heterogeneous patients and can be used in this way to identify subsets most likely to respond to a molecular-targeted treatment.

The strength algorithm is applied to gene signatures that represent the target mechanism (e.g., a c-Met targeted therapy) and patients can be stratified by their respective levels of pathway activation (Fig. 38.1). If patients can be stratified by the strength of target mechanism signaling, and this signaling can be considered

a surrogate response to treatment, we can determine whether a patient is in the "likely responder" or "likely non-responder" category based on their individual target mechanism/pathway signaling levels. Patients with high versus low signaling strength may be predicted to be the responders depending on the disease–target pair under study (see case study below for an example).

The gene signatures used to stratify patients by target mechanism signaling strength can range in size from four to over a thousand genes and be derived from multiple tissues. With respect to development of content for a biomarker, it is useful to identify a small, targeted number of genes to be measured. Therefore, we can develop classifiers to predict whether a patient is a "likely responder" or a "likely non-responder." The population of likely non-responders can be analyzed further to identify the disease driving mechanisms active in these patients by researching known mechanisms in the literature, or using the 2,000+ gene signatures to identify mechanisms that mostly saliently distinguish different groups of patients. This can illuminate potential therapeutics that may target the different subsets from patients, and enable portfolio optimization to provide holistic treatment solutions for an entire disease population.

4 Case-study: Identifying Ulcerative Colitis Patients that Respond to Infliximab in the Absence of Clinical Outcomes Data via TNF Pathway Activation Levels

We tested this approach to predict response to therapy while remaining naive to clinical response by generating a gene expression classifier to identify patients most likely to respond to the TNF targeted therapy infliximab, and testing it in a patient population where response to infliximab is known. This example was chosen because two datasets with baseline gene expression profiling data and response to therapy are published providing for training [9] and test [10] datasets. Based on the previously published work of others, we hypothesized that patients with high levels of TNF activation were less likely to respond to a TNF targeted therapy [11–13] than those with lower TNF activation, and developed a TNF signaling strength-based classifier to identify patients with "high" versus "low" TNF pathway activation. The response to therapy calls available in training and test sets were used only for: (1) validation of the hypothesis that patients with high TNF pathway activation were less likely to respond to treatment and (2) validation of the predictive capacity of the classifier that were not used in any way during classifier generation.

To detect TNF signaling in colon, a 256-gene signature was culled from our casual knowledge base and applied to colon samples from a training set of 43 patients with inflammatory bowel disease (24 patients with UC and 19 with Crohn's colitis). Six healthy control subjects along with the training set patients were stratified by their individual levels of TNF pathway activation (Fig. 38.2). The

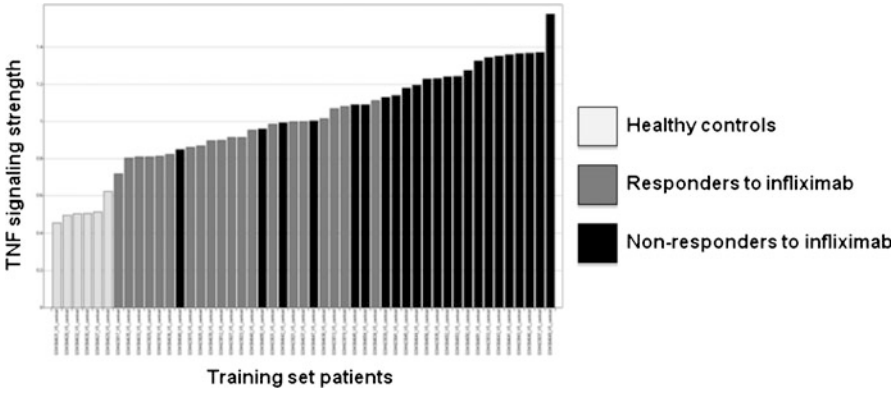


Fig. 38.2 Stratification of diseased training set patients and healthy controls by their levels of TNF signaling pathway strength. Patients that responded to infliximab are shown in *orange*, non-responders in *blue*, and healthy controls in *grey*. Patients with lower levels of TNF pathway activation were more likely to respond to infliximab

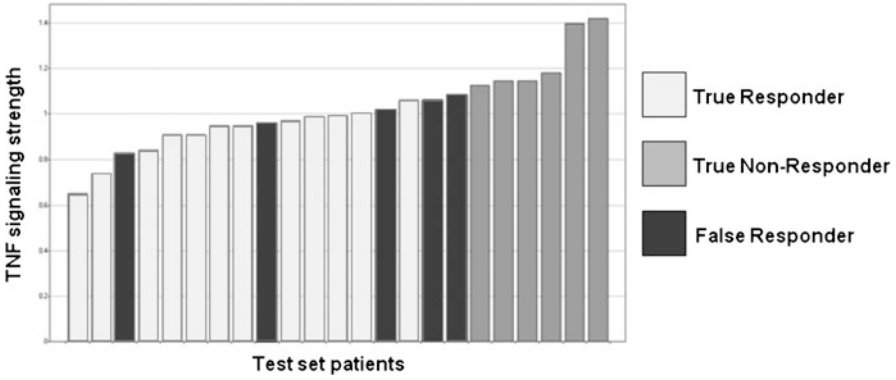


Fig. 38.3 Predicting infliximab response through TNF activation levels. The TNF pathway activation classifier predicts low TNF signaling and response to infliximab. True responders and non-responders are shown in *yellow* and *grey*, respectively. False responders are shown in *purple*

healthy controls had the lowest levels of TNF pathway activation, and low levels of activation in treated patients correlated with response (Fig. 38.2, $p = 3e^{-8}$), confirming our hypothesis.

Standard classifier development methods were applied on data from patients with the highest 20% and lowest 20% TNF activation level to develop a 20-gene classifier [14]. The TNF pathway activation classifier, using detection of TNF pathway amplitude as a surrogate marker of response, performed with a 70% responder predictive value and a 100% non-responder predictive value in an independent test set of 23 UC patients where outcomes to infliximab were known (Fig. 38.3).

This example with infliximab in UC is one validation of how our approach for patient stratification by disease-driving mechanisms and pathway activation can be used to predict response to a targeted therapy. Once patient populations are identified, biomarkers can be generated for each subset driven by a distinct pathway, as we did here with TNF. These biomarkers may then be further developed as a therapeutic diagnostics for selecting appropriate patient populations for entry into clinical trials or for postmarketing use.

5 Conclusions

While previous post hoc biomarker development has demonstrated the benefit of applying biomarkers to preselect a patient population more likely to respond to a particular therapy, the resources required for the generation of data coupled with clinical outcomes are extensive. Here we have demonstrated a means to identify biomarkers with exclusively untreated patient data, bypassing the cost associated with post hoc biomarker development.

The biomarker development strategy presented here is driven by patient data and requires a highly represented field of patients to best capture the heterogeneity of disease. Currently, efforts to molecularly describe cancers of various tissues have been undertaken and are embodied by efforts such as the government-sponsored, The Cancer Genome Atlas [15]. Databases such as these are an ideal substrate by which to generate patient biological profiles facilitating identification and stratification for specific drug targets: they are extensive, with a large amount of samples and multiple large-scale data modalities including gene expression, and they are publicly available. Furthermore, they serve as a blueprint for how we should move forward with other diseases that remain wanting for rich and plentiful patient datasets.

The strategy of using patient data in the early phases of the drug development process holds tremendous promise for bringing increased value to portfolios, both by assigning new indications for existing drugs and as the engine for future pipeline development. Our approach described in this chapter using high-throughput data analysis and interpretation is an ideal conduit for bringing the wealth of information from attainable patient data to the process of drug discovery in order to realize the success of an a priori drug development paradigm.

References

1. Brown D, Superti-Furga G (2003) Rediscovering the sweet spot in drug discovery. Drug Discov Today 8(23):1067–1077
2. DiMasi J (2002) The value of improving the productivity of the drug development process: faster times and better decisions. Pharmacoeconomics 20(Suppl 3):1–10

3. DiMasi J, Hansen R, Grabowski H (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22(2):151–185
4. Early Breast Cancer Trialists' Collaborative Group (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomized trials. Lancet 365:1687–1717
5. Hofmann M, Stoss O, Gaiser T, Kneitz H, Heinmöller P, Gutjahr T, Kaufmann M, Henkel T, Rüschoff J (2008) Central HER2 IHC and FISH analysis in a trastuzumab (Herceptin) phase II monotherapy study: assessment of test sensitivity and impact of chromosome 17 polysomy. J Clin Pathol 61(1):89–94
6. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes J, Chalchal H, Shapiro JD, Robitaille S, Price TJ, Shepherd L, Au HJ, Langer C, Moore MJ, Zalcberg JR (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 359:1757–1765
7. Tan P (2009) Divide and conquer: progress in the molecular stratification of cancer. Yonsei Med 50(4):464–473
8. Naylor S, Cole T (2010) Overview of companion diagnostics in the pharmaceutical industry. Drug Discov World Spring Edition: 67–79
9. Arijs I, De Hertogh G, Lemaire K, Quintens R, Van Lommel L, Van Steen K, Leemans P, Cleynen I, Van Assche G, Vermeire S, Geboes K, Schuit F, Rutgeerts P (2009) Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. PLoS One 4(11):e7984
10. Arjis I, Li K, Toedter G, Quintens R, Van Lommel L, Van Steen K, Leemans P, De Hertogh G, Lemaire K, Ferrante M, Schnitzler F, Thorrez L, Ma K, Song XY, Maranoo C, Van Assche G, Vermeire S, Geboes K, Schuit F, Baribaud F, Rutgeerts P (2009) Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. Gut 58(12):1612–1619
11. Kang CP, Lee KW, Yoo DH, Kang C, Bae SC (2005) The influence of a polymorphism at position-857 of the tumour necrosis factor alpha gene on clinical response to etanercept therapy in rheumatoid arthritis. Rheumatology 44(4):547–552
12. De Vries N, Tak PP (2005) The response to anti-TNF-alpha treatment: gene regulation at the bedside. Rheumatology 44(4):547–552
13. Shetty A, Forbes A (2002) Pharmacogenomics of response to anti-tumor necrosis factor therapy in patients with Crohn's disease. Am J Pharmacogenom 2(4):215–221
14. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
15. McLendon R et al (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061–1068

Chapter 39
Biomedical Atlases: Systematics, Informatics and Analysis

Richard A. Baldock and Albert Burger

Abstract Biomedical imaging is ubiquitous in the Life Sciences. Technology advances, and the resulting multitude of imaging modalities, have led to a sharp rise in the quantity and quality of such images. In addition, computational models are increasingly used to study biological processes involving spatio-temporal changes from the cell to the organism level, e.g., the development of an embryo or the growth of a tumour, and models and images are extensively described in natural language, for example, in research publications and patient records. Together this leads to a major spatio-temporal data and model integration challenge. Biomedical atlases have emerged as a key technology in solving this integration problem. Such atlases typically include an image-based (2D and/or 3D) component as well as a conceptual representation (ontologies) of the organisms involved. In this chapter, we review the notion of atlases in the biomedical domain, how they can be created, how they provide an index to spatio-temporal experimental data, issues of atlas data integration and their use for the analysis of large volumes of biomedical data.

R.A. Baldock (✉)
MRC Human Genetics Unit, MRC Institute of Genetic and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK
e-mail: Richard.Baldock@hgu.mrc.ac.uk

A. Burger
MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK

Department of Computer Science, Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: Albert.Burger@hgu.mrc.ac.uk; A.G.Burger@hw.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_39, © Springer Science+Business Media, LLC 2012

1 Introduction

Biomedical research has always relied on visual observation and imaging is a primary mechanism for recording data from the sub-cellular through to whole-organism level. In particular, imaging is used to capture the spatial organisation of biological entities, such as sub-cellular organelle and chromosomal organisation, cellular morphology, tissue organisation and organ histology and morphology. At the highest levels of resolution imaging is being used to capture molecular structures, synaptic organisation and molecular flux within the cytoplasm. Modern imaging techniques have been extended to capture 3D data not only at all ranges of resolution, but also to include the option of capture through time. Figure 39.1 shows a range of imaging modes that illustrate the nature of the spatio-temporal data produced for biomedical research.

In many cases image data are used to support simple observations. For example, gene X is expressed in the ventral half of the left ventricle, or the cells of the epithelial layer show an elongated appearance. As more data is collected, the trend is to use manual and automated means to extract numerical information from the

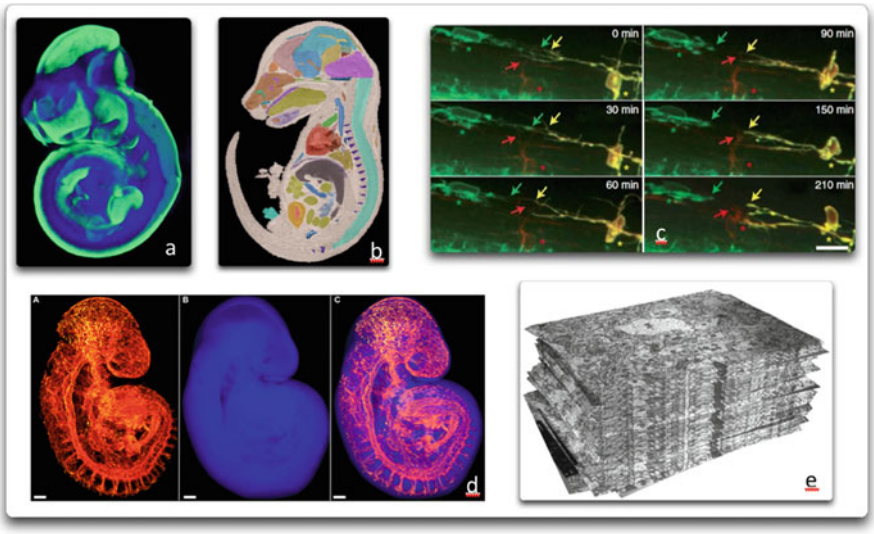


Fig. 39.1 (**a**) Optical Projection Tomography (OPT) image of mouse 10.5 dpc embryo in-situ hybridisation expression of Crabp1; (**b**) Caltech μMRI (Magnetic Resonance Imaging) image from the Caltech Mouse Atlas; (**c**) Time-lapse confocal images of oligodendrocyte development, adapted by permission from Macmillan Publishers Ltd., Nat Neurosci [21] copyright 2007; (**d**) Transgenic expression of vascular development in the mouse embryo [34]; (**e**) serial-section EM (Electron Microscopy) reconstruction of a neuropil structure courtesy of SynapseWeb, Kristen M. Harris, PI, http://synapses.clm.utexas.edu/

images to provide objective numerical analysis in terms of spatial patterns, signal intensities, shape and morphology, cell densities ablation recovery times, etc. As data is captured at higher rates and volumes, the requirements for image archiving and analysis are demanding far greater automation. The focus of this paper is image data captured to show information at the organ or whole-organism level of biological organisation. In our case this is with respect to embryo development and can include gene-expression patterns, lineage tracing, physiology and cellular activity, morphometric and mutant phenotype. At this level of biological organisation a key requirement is to be able to compare spatial and temporal patterning and to be able to collate information from across all the different imaging modalities. At the genomic and molecular biology level the natural framework for capturing data relationships is the genome, at the organ/organism level the appropriate framework is provided by explicit spatio-temporal atlases [9]. To some extent the spatial aspects of the information can be captured by annotation using an anatomy ontology, but this does not have the resolution or computational capability of an explicit coordinate framework provided by a digital atlas.

Atlases provide the integrating framework for spatial data of tissues, organs and whole organisms. For genomic and molecular level data, information between species can be compared by "mapping" of the sequence data. Such sequence mapping provides detailed comparative analysis of the evolution of the genome and enables the use of model organisms (e.g., mouse) to support research into human disease and abnormalities for translational purposes. By analogy the basic information captured at the whole-organ level can be compared across species including through to human for direct medical research and ultimately clinical application. If we take the "layer cake" view of biology passing from the lowest levels of organisation at the base through to tissue, organ and whole-organism level at the top, then the spatio-temporal data mapped to the atlases at the top serve as the target for a systems biology understanding of the high level biology. In addition, the basic research data captured, for example, for model organisms such as zebrafish and mouse, serve for comparative analysis and provide basic understanding to physiological and disease modelling applied for translational research into the human condition. This can extend through to medical and clinical data sets and ultimately through to individuals. At this end of the atlas range we envisage the use of a personal "myAtlas" to capture and record the clinical history of an individual and perhaps to support patient–doctor consultation. This view of the role of explicit spatio-temporal atlases in the context of biological and medical research and potentially clinical practice is illustrated in Fig. 39.2.

In this paper, we outline informatics aspects of atlasing frameworks in the context of biomedical research and illustrate these with procedures and examples from the eMouseAtlas project EMAP and EMAGE [4] (Edinburgh Mouse Atlas Project and Edinburgh Mouse Atlas Gene Expression database).

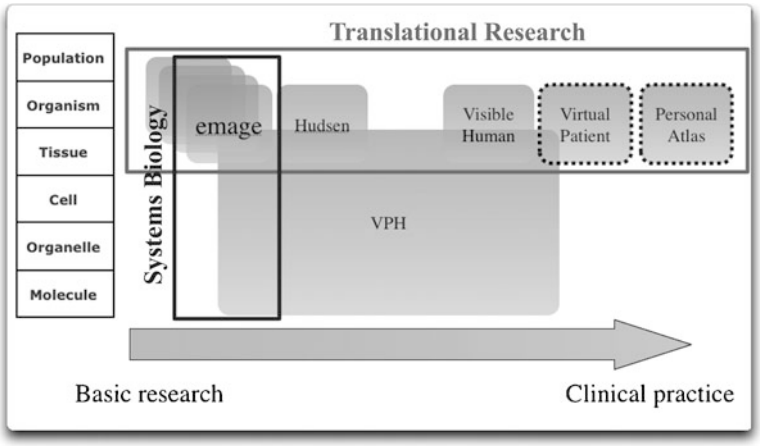


Fig. 39.2 Atlases in the context of systems biology and translation biomedical research. Hudsen is the human development atlas and data resource, the visible human indicates the adult level atlas and virtual patient and personal atlas indicate resources under development or envisaged. VPH refers to the international Virtual Physiological Human programme to develop computational and predictive models of adult human physiology

2 Atlas Systematics

In the scientific world, there often is a general understanding of the meaning of widely used key terms, such as, "gene" or "ontology", but a lack of agreement on their precise definition. This applies particularly to the use of the term "Atlas" in biomedical research. Here we develop a classification of resources that describe themselves as atlases and argue that a proper use of the term should imply an overt spatial representation used to express the spatial relationships in the data.

For most people the definition of an atlas relates back to the familiar geographic atlases and maps and is typically an overt depiction of a coordinate space, e.g., the surface of the earth. This is supplemented with the representation of features and regions which in the geographic example could be cities and countries. The Oxford English Dictionary (OED) defines the term *atlas*: *A collection of maps (or illustrative plates) in a volume*, where a *map* is defined as *A diagram or collection of data showing the spatial distribution of something or the relative positions of its components*. For us the equivalent of the collection of maps is a collection of 2D or 3D images, which define the space we need to represent for the mapping of data with spatial relationships. Some technologies, e.g., Optical Projection Tomography (OPT) [32] allow the generation of the 3D model directly, but from which 2D section images can be generated. Most atlases we know of use actual images, such as, generated by microscopy or MRI, instead of symbolic depictions (e.g., drawings). In either case, the visual representation in the form of sets of pixels and voxels, described in an appropriate coordinate framework, forms the first essential

component of our notion of a biomedical atlas. Although it is not stated so explicitly in the OED definition implying "spatial distribution" or "relative position" requires some sort of labelling or *mapping* of the artefacts in the context of the map. In geography we expect the regions of countries and cities on a map. Similarly, in the context of a biomedical atlas, we expect labels describing the components in the visual representation, e.g., the label "heart" refers to an image region depicting the heart in the image model. This implies that there is a mapping between the term and the image model.

The terms may simply consist of a controlled vocabulary such as the names of anatomical structures, or form a part of a formally specified ontology. This formalisation can be fairly lightweight, using languages such as SKOS [27], or rather detailed and precise, using languages such as OWL,[1] to describe it. The higher the level of formalisation, the more automated reasoning it will allow, but the more difficult it is to get widespread acceptance of the ontology as a standard within the biomedical community. This has implications for interoperability (see Sect. 5).
With this discussion we can identify components that could be part of an atlas:

Representation of space: a visual representation such as an image with the image coordinates allowing location of specific features or regions. For biomedical atlases this is typically a selected representative image or an averaged image over a number of samples.

Spatial reference terms: in biomedical atlases this is typically anatomy.
Mapping: locations or regions of the spatial reference terms in the context of the spatial representation.

Direction: definition of directions in the context of the underlying object. In a geographic atlas this is usually simple to identify North or to plot lines of latitude and longitude. In biomedicine it may require a much more complex mapping of left–right, dorsal–ventral and anterior–posterior axes at each location within the map.

Data: this is the association of data such as gene-expression or physiological state with different parts of the spatial representation.

Some "Atlas" resources only include the spatial representation in an implicit way by referring only to the anatomical terms. Examples of such atlases are the Human Protein Atlas [5] and Gene Expression Atlas [18], where data is annotated with anatomical tissue and cellular terms but there is no explicit spatial representation or mapping. All spatial association is via the spatial understanding of the user. At the other extreme are the full 3D, spatially mapped anatomical atlases, used as frameworks for capturing digital image data. Examples here are the mouse gene-expression databases eMouseAtlas [4, 7] and the Allen Brain Atlas [23]. Between these there are traditional "paper" atlases such as the Atlas of Mouse Development' by Kaufman [19] and those with digital content, e.g., the Paxinos Rat Brain Atlas [29].

[1] www.w3.org/TR/owl2-overview/.

In order to realise the power of a digital atlas to provide an objective spatial analysis and provide tools for spatial data mining an explicit spatial representation is essential, and we therefore define the minimal requirements for a biomedical atlas to be:

$$\text{Spatial Representation} + \text{Terms} + \text{Mappings} = \text{Atlas}$$

If an atlas also includes a specification of biological directions then more sophisticated query and analysis in biological terms becomes possible. The construction of biomedical atlases, the use of atlases to index spatio-temporal experimental data, the integration across atlases and other resources, and the use of atlases in the context of the analysis of large quantities of biomedical data are discussed in the following sections.

3 Atlas Construction and Spatial Annotation

Biomedical atlases that include an explicit coordinate framework can be constructed in many ways, including "simple" graphical modelling to depict the primary structures that are to be represented. In practice, atlases developed for biomedical research are typically based on one or more representative individuals using imaging that enables full 3D reconstruction. This can be a direct 3D imaging technique, such as, μMRI [11], μCT [1] (Computed Tomography), block-face imaging [35, 36] or OPT [32] or, if resolution and contrast are critical, then 2D imaging of microtome sections followed by 3D reconstruction. When the key requirement is to be able to capture spatial patterns for subsequent comparison and analysis, for example anatomical labelling or syn-expression grouping, then it is sufficient for the atlas to be a *representative* individual. Such an atlas can also be used to capture morphological variation of experimental sets by capturing both the mapped data and the spatial transformation from which variation in the original data set can be established [8]. If, however, the key purpose is to be able to assess the morphology of a new sample, it is more convenient to create a probabilistic atlas [13, 25].

For the mouse embryo models of the eMouseAtlas database we have used a combination of OPT 3D imaging to capture the original shape of the embryo followed by wax-embedding and microtome sectioning so that the individual section could be stained to reveal the cellular detail. These histological sections are imaged at high-resolution and reconstructed using the OPT image as a morphological template. When the 3D model has been reconstructed, it is then segmented into anatomical regions which provide the link between the spatial representation of the embryo (image coordinates) and the anatomy ontology. This process is illustrated in Fig. 39.3. The embryo atlas we have developed in Edinburgh is comprised of a series of 3D reconstructions, an anatomy ontology to describe the developing anatomy, plus a set of delineated regions or domains that link the ontology terms (at some level of resolution) to the 3D image model.

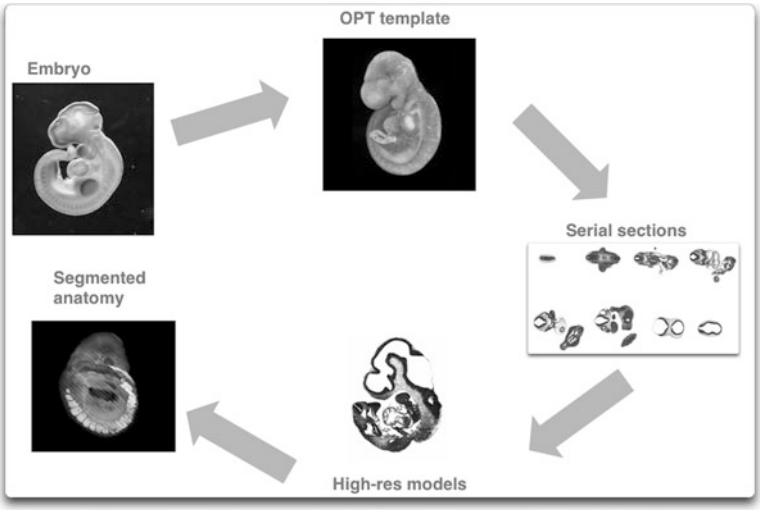


Fig. 39.3 Reconstruction process used to build the high-resolution 3D models of the mouse embryo for EMAP

The 3D image can be used directly as an atlas framework. In some cases it is possible to supplement the image coordinate frame with more biologically meaningful coordinates such as the stereotaxic coordinates used in neuroscience studies of the brain [16]. This is not always required and the key requirement for an atlas to be useful is a mechanism by which data can be spatially transformed or *mapped* into the atlas space. This is termed spatial registration or spatial mapping and in general is a complex non-linear transformation from the original (source) coordinate frame to the atlas coordinate frame (target). Image registration has been studied very thoroughly, especially for clinical imaging where comparison between modalities and for disease progression are important. Techniques that have been established typically define a deformation field across the volume of image space enclosing the source image of interest. This deformation field is established by manual definition of points of correspondence or automation and the full field defined via a mathematical model such as radial-basis function interpolation or physical modelling of the deformation. In either case we have found that the embryo presents special problems because of the extreme deformations that arise due to the flexibility and variability of presentation and pose. In this situation the standard warping techniques fail and we therefore have established warping based on the constrained distance transform [17], which is a combination of rapid manual alignment to correct the primary deformation due to pose followed by an automated process using the open-source software ANTs [2] to fine-tune the alignment. The WlzWarp software tool we use for the manual registration is illustrated in Fig. 39.4.

The procedure to transform experimental samples into the model space can be considered as *spatial annotation*. Each location within the sample acquires a

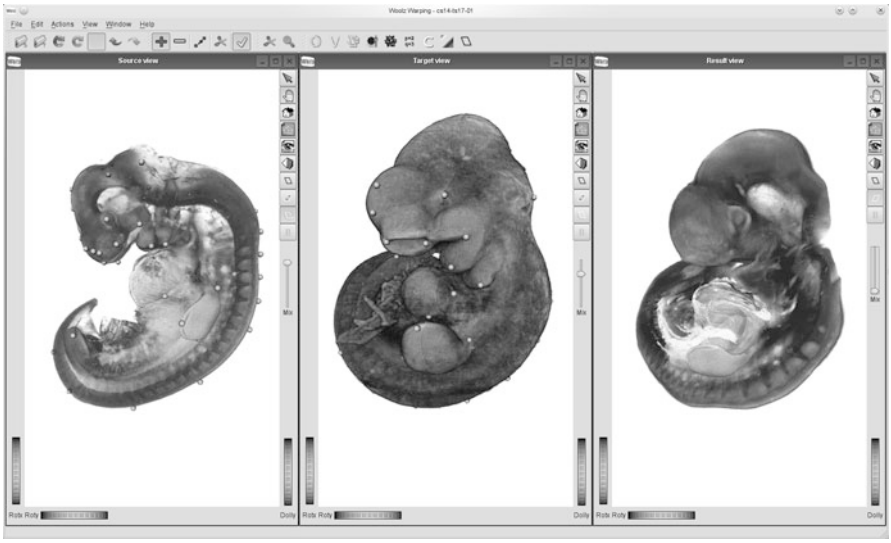


Fig. 39.4 Spatial mapping of a 3D image of a human Carnegie stage 14 embryo onto the EMAP Theiler stage 17 embryo model using the WlzWarp software tool developed to deal with the complex mappings required with embryos. *Left-hand frame* original human embryo; *middle frame* the target mouse embryo; *right-hand frame* the warped human embryo to match the mouse. The marked locations show locations of point-correspondences (Note: Carnegie and Theiler stages are classification systems for how far human and mouse embryos, respectively, have progressed in their development.)

mapping into the model. In this way, data from the model can be presented in the context of the sample, or data from the sample such as gene-expression signal intensity, can be transformed into the space of the atlas models. It is then possible to analyse the data either in terms of the atlas, e.g., to establish anatomical regions that show gene-expression, or to compare with other data directly, such as, other gene-expression patterns. In analogous fashion to a text-based annotation, spatial annotation enables search for patterns but directly in terms of the atlas space, e.g., queries, such as, "what genes are expressed at this locations?" or "what gene show expression in a similar pattern?" are now possible.

In the mouse atlas EMAP and associated gene-expression database EMAGE [7] the spatial annotation is a standardised procedure to map the source image, and to segment the signal into pre-defined strengths of expression.[2] The mapped signal patterns are held in the database and a query against the database results in direct comparison of image data. This is an image-processing operation and executed in

[2] http://www.emouseatlas.org/emage/about/data_annotation_methods.html.

an image server linked to the RDBMS (Relational Data Base Management System) which manages the metadata. For efficiency the spatial indexing and similarity calculations are encoded using the Woolz image-processing library.

4 Experimental Data and Atlases

Atlas frameworks can be used to capture, compare and analyse any spatial data, which can range from cellular signalling and gene-expression patterns through clonal distributions to long-range neuronal connectivity and physiological function. Here we will use the data captured in the context of the eMouseAtlas models to illustrate the issues of mapping and interoperability of atlas-based resources. The primary data for which the mouse embryo atlas was established is gene-expression patterns as revealed by in situ hybridisation to mRNA and immunohistochemistry with protein antibodies. In addition, we have mapped anatomy terms to the 3D space and explored direct mapping of cellular clonal data following lineage tracing experiments.

4.1 In situ Data

Transforming a gene-expression pattern from an in situ experiment involves two steps. The first is to establish the spatial transformation from the experimental data images to the atlas model. The second is to segment the signal in the context of the original data and to use the spatial mapping to transform the pattern to the atlas model context. Our experience with mouse embryo data indicates we need to deal with a number of different presentations of the information:

2D data: Intrinsically 2D data captured from the embryo. The prime example is a whole-mount view which is effectively a projection of the underlying 3D data onto 2D and in principle the original 3D location of the signal cannot be recovered. For this data we have adopted a simple approach of mapping the data to a projection of the atlas model, i.e., maintain the 2D character of the original data. Within EMAGE this implies that the data is segregated and a spatial query is currently against either the wholemeal data or the 3D data.

2D images of 3D data: These are microtome sections of the sample embryo which has been physically sectioned and stained. In principle, the section image can be mapped back into a 3D location within the atlas model. In practice this can be difficult, because distortion of the tissue section could mimic a re-location of the image within the 3D framework with a consequential ambiguity. This can be resolved by capturing more than one section and using the adjacent data to correctly align the sections that have been treated to show the in situ pattern. Most high-throughput data, such as generated for the Allen Brain Atlas [23] and Unexpress [12], is of this form – sparse 3D data.

In EMAGE we have adopted two strategies for the data. The first is simply to find the best matching section for each sample and to use a mapping tool such as Maxint to transform (warp) the image onto the atlas. The same tool then allows a segmentation of the signal into a series of domains to represent *strong*, *moderate*, *weak*, *possible* and *not-detected* expression strengths. An additional domain *not-observed* ensures that a null-return from the data base can distinguish data that shows no-detectable expression from no-data.

A second strategy is to project the 3D data onto 2D and to treat it in a similar fashion to wholemeal samples. This of course loses the 3D information and reduces the ability to discriminate patterns, but can be useful for a first-pass automated mapping to be followed up with a more detailed 3D mapping later.

Full 3D data: This is data from a 3D imaging technique such as OPT or confocal LAM (Laser Scanning Microscopy) or could be serial sectioning that can be reconstructed to a full 3D data set. This type of data provides the most complete view of the overall expression pattern, but is typically at a lower resolution and does not deliver the cellular detail of real histological sections. A benefit is that the process of 3D mapping is very much faster than the section-by-section mapping of sparse data, but does require sophisticated mapping tools such as the WlzWarp tool based on the constrained distance transform and potentially significant compute power for the automated fine-alignment phase.

4.2 Sparse Cell Data

In some experiments the observation may be a set of cells that exhibit a particular stain. A particular instance of this is a clonal set of cells arising from a single progenitor cell. This could be marking using a vital dye [22] or by a random recombination event in a tamoxifen-inducible cre-transgenic line [24]. The issue with this data is that the individual cells that comprise the clones cannot in general be identified in the target model. The mechanism for mapping is therefore to map by direct marking of the estimated best match for each clone cell. With serial section data this can be very time-consuming. This could be done by direct matching of a serial section series which encompasses the clone, but this is similarly time-consuming.

4.3 Anatomy and Physiology

Classically an atlas depicts the physical geography overlaid with coloured regions depicting countries and national boundaries. In biological atlases the closest analogy is anatomy overlaid on a histological image and rather like the geographic case the "country" boundaries are subject to disagreement and dispute! In the EMAP mouse

Fig. 39.5 The EMAP anatomy browser. The user can select arbitrary section views through the 3D model and show selected anatomical components overlaid on the histology. In this case we are showing the limb atlas material from DeLaurier et al. [10]

atlas the anatomy delineations are available for download and can be visualised in a number of applications. Figure 39.5 shows a screenshot of the anatomy viewer provided for visualisation in the context of a standard web-browser. In this case we are showing a view through the limb atlas of DeLaurier et al. [10].

The atlas can of course also capture physiological data such as calcium concentration and ion-channel status in the heart or functional imaging of the brain. This type of data will clearly include detailed temporal and behavioural information, but the spatial aspects of the observations can be mapped to the atlas and compared with other data. An example we have been exploring is discussed in the next section; it is the use of the atlas approach to integrate a detailed physiological model of the heart with a statistical model of dynamic heart morphology over the cardiac cycle. The basic concept is to map both models to a common atlas model which can then also bring in other data from for example the EMAGE gene-expression database into the same analysis.

5 Integration of Biomedical Atlases

Computational frameworks, such as the atlases described in this paper, are in the first instance mechanisms for the management of data and knowledge, initially simply to capture and store it, but subsequently also to query it and to perform complex analysis studies (see Sect. 6). Typically, we find more than one atlas covering the same or related data, and we usually want to link data in an atlas to that in a

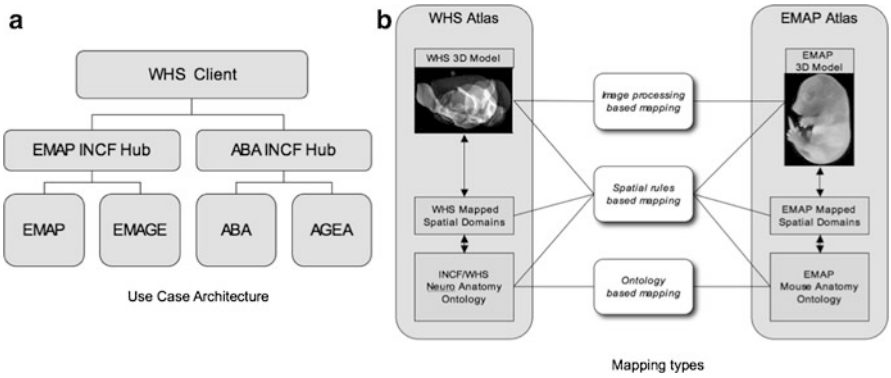


Fig. 39.6 (**a**) For the gene expression use case, a client application specifies a point in Waxholm Space (WHS) in order to access relevant gene-expression data in EMAGE and AGEA, (**b**) mapping between atlases can be achieved by applying image-processing techniques, ontology-based mappings and the specification of locations using spatial rules which are based on the 3D models as well as the ontological description of the atlas anatomies

non-atlas resource, e.g., entries for gene-expression experiments in EMAGE have links to Ensemble (www.ensemble.org) for further information about the gene under consideration. All this creates a challenging integration problem for biomedical atlases. As always, the desired interoperability between atlas and related resources depends to a large extent on agreed standards. In this section, we illustrate these interoperability issues, drawing on our experience of a use case study linking the EMAGE data set to the Allen Brain Atlas [28] using the emerging Waxholm Space standard [16] which is a 3D reference atlas for the adult mouse brain.

The basic architecture for this use case is shown in Fig. 39.6a. It is based on the INCF Digital Atlas Infrastructure (INCF-DAI), which is currently being developed by the INCF (International Neuroinformatics Coordination Facility, www.incf.org). The INCF hubs for EMAP and ABA (Allen Brain Atlas) are responsible for mapping the point of interest in Waxholm Space (WHS) into corresponding locations in their respective atlas spaces – an alternative approach where a central spatial transformation INCF hub will assume responsibility for all spatial transformations is being considered – and then retrieve the relevant gene-expression data for return to the client. At this stage, the hubs simply return URLs to html pages containing the gene-expression query results. The client displays these in two separate browser windows, but does not merge the results. To facilitate the latter, a standard for gene-expression query results needs to be agreed first. This is a key point, as it applies to many different types of data. Achieving interoperability between different spatio-temporal reference frameworks does not guarantee the interoperability of the data that is indexed by these frameworks. Standards, such as for gene-expression data, are required in addition, if the analysis of the data across multiple atlases is to be maximally supported by software.

As discussed in Sect. 2, although there is no single definition of what biomedical atlases should consist of, it is usually the case that they have an image component as

well as textual labels for identifiable regions of the image space. In some cases the textual labels are part of comprehensive anatomy ontologies. This leads us to the following three spatial mapping types: (1) based on image processing, (2) based on ontology mapping and (3) based on spatial rules; see Fig. 39.6b for an overview diagram in the context of our use case. The first of these uses image-processing algorithms to transform pixels and voxels from one space to another. In our examples we use a constrained distance transform to link between the WHS atlas and the EMAP atlas spaces. The second type is based on mapping anatomical concepts from one ontology to another, e.g., the concept `Cerebellum` in the ABA maps to the `Cerebellum` in EMAP. The third type uses spatial relations, such as, `contained_in` and `next_to`, known about identified regions in the atlas to describe a spatial location. Whilst the image-processing solution can potentially achieve very good accuracy, it does so only for atlases that are morphologically not too different. Ontology-based mappings deal with such differences more easily, but do not achieve the same level of precision. The use of spatial rules is a compromise solution that aims at reasonable accuracy in spite of some morphological differences.

We know that the level of formalisation of the terms used by atlases has a significant impact on their interoperability. In principle, a more detailed ontology leads to better integration possibilities, but only if this ontology is widely shared and used by the biomedical atlasing community. Herein, however, lies the dilemma, since the more detail one specifies in the ontology, the more difficult it becomes to obtain community acceptance. There exists, of course, a large body of work on the topic of biomedical ontologies, and a detailed discussion of it lies outside the scope of this paper. For a collection of relevant papers, we refer the reader to [6]. An area of biomedical ontologies that has not been explored very much thus far is their use in the context of spatial rules-based mappings, which will require, amongst other things, some level of standardisation of the meaning of directional terms, such as, “lateral to”, “close to”, etc. The challenge in biomedical atlasing is the lack of a single frame of reference, such as is available in the geo-sciences; there is only one Earth, but there are many instances of organisms such as human and mouse.

It is important to remember that in the context of integrated spatial queries, several mappings across different spaces may be involved. Figure 39.7 illustrates how we distinguish between four categories of spaces. Initially, experiments, such as for in-situ hybridisation gene-expression analysis, are carried out in the context of specific animal experiments resulting in 2D and 3D image data for their particular *experiments space*. These results are typically mapped into a standard spatial or spatio-temporal repository framework, the *respository space*, such as EMAP, through which they can then be queried. To integrate across two or more repositories may involve a *mediator space* such as Waxholm Space (WHS), and if the data that has triggered the original query of interest is based on a particular experiment, we require a mapping from the *query space* to the mediator space. Where labs produce their own data for their reference space, the distinction between repository and experiments space may not explicitly exist. So, an integrated query to EMAGE and AGEA brain gene-expression data, using WHS as the mediator, would involve at least 4 spatial mappings and potentially all three mapping types.

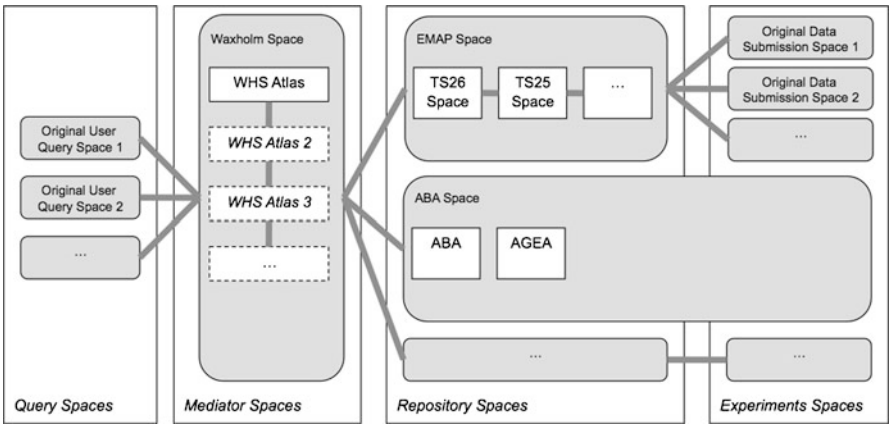


Fig. 39.7 Types of spaces

As the number of mappings across spaces increases, the accuracy of the results for a query is likely to diminish. Intuitively, one might argue to expect the overall accuracy to be determined by the "weakest link", i.e., the least accurate spatial mapping involved in the query. However, there may also be an accumulative effect resulting in even worse accuracy. There is also an issue of giving an end user the impression of high precision, for example, because his/her query space was very carefully mapped into the mediator space, but that the actual results are by far less accurate due to other mappings involved.

The above discussion has focused on the integration of atlases as spatio-temporal frameworks for experimental data, but as more and more such data becomes available, we also see an increase in computational models which firstly help explain the underlying biomedical processes resulting in this data, and secondly include predictive capabilities for scenarios that have not yet been studied experimentally. The integration of "data atlases" with the spatio-temporal frameworks of computational models is critical for the development and calibration, as well as the validation and verification, of the models. As part of the European Union's *Virtual Physiological Human* (VPH) research programme (www.vph-noe.eu), the RICORDO project (www.ricordo.eu) investigates this data model integration for volumetric data. Amongst other work, it has developed a spatial mapping from the computational heart developed at the University of Auckland to the EMAP atlas in Edinburgh. To the best of our knowledge this is the first example of mapping volumetric, computational VPH model sections to the corresponding location in a 3D framework for molecular data (gene expression). Although it is outwith the scope of this paper to discuss the technical details of this mapping, it illustrates one example of this extended requirement for atlas integration. Based on the increasing amounts of experimental data and related models, we predict that the importance of this type of integration will significantly increase over the next five years.

6 Using Atlases for Data Analysis

Atlas frameworks provide a straightforward context for spatial comparison and analysis. The types of analysis depend on the nature of the data collected and can be characterised by the nature of the input data and the output results. For example, if the input is a point or region defined within the atlas space, the result could be atlas-based, such as an image of the overall gene-expression intensity, or a numeric result, such as the similarity to another expression pattern, or just a list of assay results that match the query. Similarly the input could be a list of genes for which co-expression hot-spots are required in which case the output would be a heat-map type image with a gene-list associated with each point. In this section we illustrate the use of atlases for data analysis in the context of the embryo and atlas databases that we have integrated with the eMouseAtlas resource. These include the human embryo atlas and database Hudsen [20], GUDMAP [26], EurExpress [12] and Chick Atlas[3] databases.

6.1 Annotation and Query

Mapping data onto the atlas framework provides a means to specify a query on the database in graphical terms. This could be as simple as a single vertex or as an arbitrary point-set representing a complex region within the domain of the atlas model. In addition, an atlas within which the anatomical tissues have been delineated makes it possible to query using the anatomical terms. These provide two simple examples of data analysis. The first is annotation. By mapping the expression pattern onto the atlas model and comparing the mapped pattern with the anatomical domains delineated within the model, it is possible to generate an anatomical description of the gene-expression pattern. This is illustrated in Fig. 39.8 in the context of the EMAGE database. As well as establishing the list of tissues that show gene-expression, it is possible to calculate the relative proportion of each tissue that shows expression.

The second example is to use the spatial location or region as a means of finding genes expressed at the given location or area of interest. To process this query the given location is compared to each stored pattern in the database to establish if it is contained within the mapped region. In this case the spatial "index" of a mapped gene-expression pattern is represented internally as an image region or binary image. The query region as a single point or a second image region is compared with the expression pattern using a simple image operation of domain-intersection. This is equivalent to the intersection of two point-sets, but executed as an efficient image-processing algorithm. If the resulting intersection domain is non-empty then

[3] http://www.echickatlas.org/.

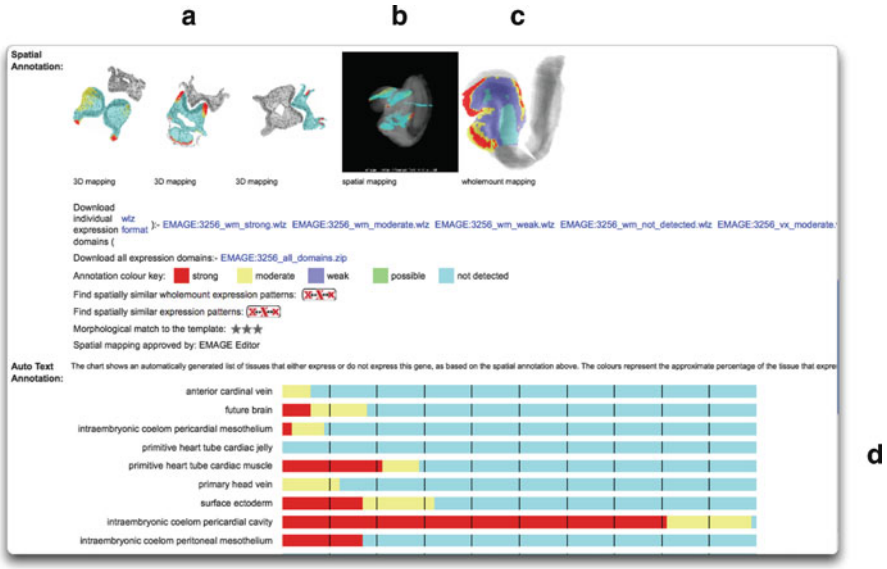


Fig. 39.8 Spatial analysis. The mapped expression pattern for the gene Bone morphogenic protein 5 (Bmp5) is mapped onto the Theiler stage 12 embryo atlas (**a**). The mapped section data is shown in 3D (**b**) and in the context of the wholemount embryo (**c**). The bar chart (**d**) shows the expression analysis in the terms of anatomical tissues that is automatically generated by comparing the expression domain with the delineated anatomy domains

the two patterns overlap. This is repeated for each pattern that could form a match. The result in the context of EMAGE is a list of assays that show overlap with the query region (see Fig. 39.9).

6.2 Similarity and Correlation

In addition to using the pattern to simply test spatial overlap or containment, the patterns can be compared for spatial similarity. For potentially dispersed and non-contiguous patterns we have discovered that the Jaccard index, which is a simple set-based measure of similarity, provides a suitable first-pass numerical value of similarity which is robust to the variation and noise found in typical gene-expression patterns. Here it is implemented in the context of a spatial region of interest defined by dilating the query pattern by the equivalent of about 300 μm. The tool is the LOcal Spatial Similarity Search Tool (LOSSST) and is described by Venkataraman et al. [33]. Figure 39.9 shows the result of using LOSSST to query the EMAGE database using the expression pattern of Bmp5 at Theiler stage 12. Using the option of mapping the query region through to other embryonic stages, it is possible to extend the query to return temporal data.

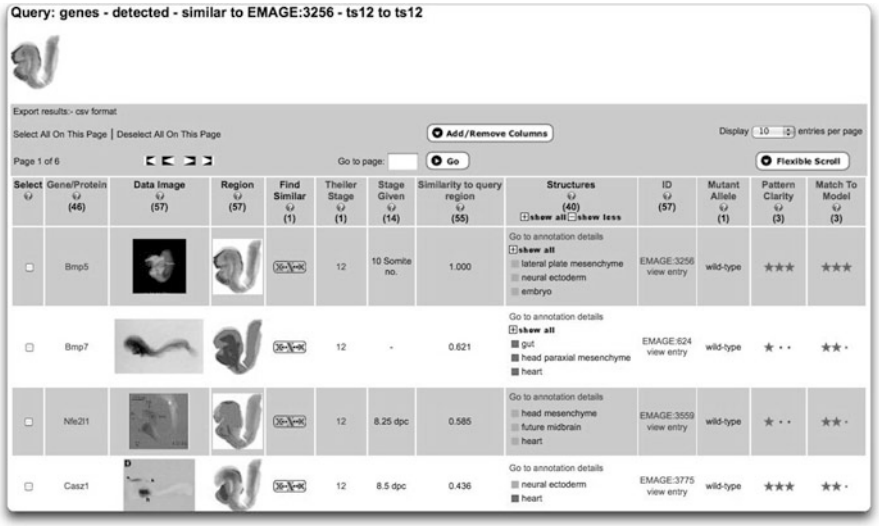


Fig. 39.9 Result of a spatial query on the EMAGE gene-expression database. In this screen shot the data has been sorted by similarity with the expression of Bmp5. The Bmp5 pattern is returned at the top if the list with a maximum similarity match of 1.0, the next most similar if Bmp7 from the same gene-family. Note real interface uses colour to show pattern strength

The use of similarity provides a sorted query result bringing to the top syn-expression patterns for any given gene. The same query can be posed on text-annotated data such as in the EurExpress database. The two annotation options are complementary. Text annotation can provide a more accurate and focussed return for very sparse and isolated tissue and cell-type specific expression patterns. Spatial annotation delivers accurate analysis of more complex and in particular near ubiquitous, but non-uniform, expression patterns.

A second measure that becomes available with spatially mapped data is expression correlation. With spatial similarity the query is to find genes with similar spatial expression patterns. It is also possible to test the expression correlation between spatial locations, i.e., to query for similar expression profiles between different locations. A good example of this is the interface provided by the Allen Brain Atlas [28] (see Fig. 39.10). For a given seed point, selected by "clicking" the required point on the screen, the system returns a correlation map which of course will have value 1 at the seed and typically includes the local region. Regions that have similar expression signatures are not always spatially connected and this may well indicate similar function or similar developmental lineage.

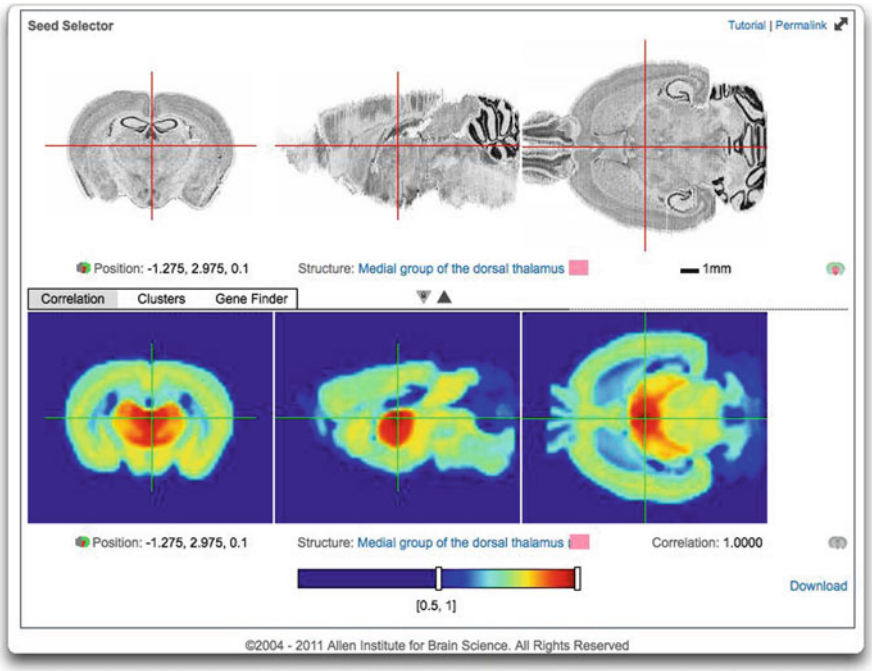


Fig. 39.10 Allen Brain atlas AGEA interface showing the correlation map for gene-expression with respect to the selected seed point

6.3 Data Mining

Atlas-based data with a mapping either onto the spatial framework of the atlas or the ontological framework, such as anatomy, becomes accessible for data mining. The simplest data mining approach is clustering based on a measure of spatial similarity or annotation. An example of this clustering is provided by the EurExpress data set [12], and the downloaded data can be visualized using standard cluster viewing packages, such as, TreeView and MeV. The results can be displayed in two ways: the first is as a set of gene-expression patterns that show similar spatial distributions and the orthogonal clustering will reveal the set of atlas regions that show similar expression profiles. These are related to the search options described above, but are not directed, and therefore provide a more objective overview of the structures implicit in the data.

Data mining can also be used to extract more detailed information from the data by using one set of data to train a classifier that can then be used to infer new relationships with a measure of confidence. An example here is the automated annotation of gene-expression patterns with anatomical terms by using an annotated set of images to train a classifier which can then be applied to new image samples. Han et al. [15] use this approach to develop tissue classifiers in the context of the EurExpress gene-expression data set.

6.4 Morphometric Variation

Atlases provide a natural framework in which to capture spatial patterns, such as, gene-expression, cell morphologies and behaviour. They can also be used to capture morphological variation even though the atlas may not be an "average" or "correct" model in the sense of representing the average size and shape for a given stage of development. In fact, in the context of mouse development a standard embryo is very difficult to define given the dynamic nature and heterochronicity of development even for pure strains. Nevertheless, if an experiment collects a standardised set of embryos with a protocol that will preserve size and shape, then morphological variation can be captured. The key to understanding this is that the data set that needs to be preserved is the non-linear mapping from each experimental instance as well as any data that may be mapped. If the mapping structures are available, i.e., the detail of how each point in the source experimental embryo is mapped, then it is straightforward to establish the average mapping from the experimental set to the atlas, and by applying the inverse of this transform to the atlas, an average embryo can be established. The meaning of this is that for each point within the average the sum of the displacements from each of the source experimental embryos will be zero. With this average in place other quantities that relate to morphological variation can be established. For example, the mean size and shapes of any given structure, say the heart, are the transformed version of the same structure in the original atlas. To establish the variation of any given feature, it is simply a matter of defining that feature in the atlas, e.g., the volume of the ventricular space in the brain, applying the inverse transform to the original sample and then re-measure the volume.

In addition, spatial patterns of variation, such as parts of tissues, that exhibit most volume variation can be displayed as a heat map in the context of the atlas and overlaid with other information. In this way it may be possible to associate morphological variation with gene-expression. Cleary et al. [8] showed how μMRI can be used to capture this type of data for late stage mouse embryos. Their methodology would not work for earlier stages of development because it lacks resolution and tissue contrast, but the principle is clear. Atlases deliver the necessary framework to associate complex morphological variation with other patterns and phenotypes in the biology.

7 Discussion and Conclusion

In the context of the developing mouse embryo, we have illustrated aspects of a new "bioinformatics" that can capture and manage data associated with higher levels of biological organisation. This approach in biology was pioneered by the Edinburgh Mouse Atlas Project (EMAP) [3,30] and demonstrates the use of explicit biomedical atlases to collate, compare and analyse spatially organised data. For the associated computer science infrastructure we coin the term "atlas informatics"

which covers the underlying theory and practice of using spatio-temporal atlases as the organising framework for spatial data. The geo-sciences have been working with geographic information systems for many years, but biology and medicine require significant extension of the techniques and capabilities, because of the variability of the underlying data sources and the complexity of the structures.

It is clear that atlases can provide the key integrating framework for data associated with individual model organisms and with the development of the methods and services for atlas integration many of the aspects of comparative analysis that are taken for granted at the genomic level become possible at the tissue and whole-organism level. This can be based on "simple" image-based mapping but a much richer semantic mapping is possible by using the underlying biology to define spatial location and direction. Developing this atlas semantic context and the logic and algebra that can use these *natural coordinate* systems is an immediate challenge for atlas informatics.

Finally, in a special issue of Science[4] dedicated to the "data deluge" a paper entitled "The disappearing third dimension", Rowe and Frank [31] discuss the difficulties of publishing 3D data, citing examples of tissue and palaeontology samples which may be difficult to replicate. They compare the image context with genomics which has a natural framework on which to associate data where re-use of experimental data is the norm. They conclude:

> Funding agencies can rejoice in the unexpected longevity and growing value in voxels they have already produced. But they must first secure the basic tenet of science by ensuring that researchers have the means to archive, disclose, validate and re-purpose their primary data.

Image repositories such as the Open Microscopy Environment [14] are essential to address part of this problem, but to retrieve, compare and analyse "re-purposed" data, a spatial framework is required, which is the role of atlases. Atlas frameworks are the key component of any informatics strategy to manage and analyse such data. In this paper we have illustrated some of the atlas informatics issues in the context of the mouse embryo but the underlying informatics model applies across biological, medical and natural sciences.

References

1. Aggarwal M, Zhang J, Miller MI, Sidman RL, Mori S (2009) Magnetic resonance imaging and micro-computed tomography combined atlas of developing and adult mouse brains for stereotaxic surgery. Neuroscience 162(4):1339–1350
2. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54(3): 2033–2044
3. Baldock RA, Bard J, Kaufman MH, Davidson D (1992) A real mouse for your computer. BioEssays 14:501–502

[4]Science 6 March 2009.

4. Baldock RA, Bard JBL, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, Davidson DR (2003) EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics 1(4):309–325
5. Berglund L, Björling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA-K, Persson A, Ottosson J, Wernérus H, Nilsson P, Lundberg E, Sivertsson A, Navani S, Wester K, Kampf C, Hober S, Pontén F, Uhlén M (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteom: MCP 7(10):2019–2027
6. Burger A, Davidson D, Baldock R (eds) (2008) Anatomy ontologies for bioinformatics: principles and practice. Springer, Dordrecht, The Netherlands
7. Christiansen JH, Yang Y, Venkataraman S, Richardson L, Stevenson P, Burton N, Baldock RA, Davidson DR (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. Nucl Acids Res 34(Database issue):D637–D641
8. Cleary JO, Modat M, Norris FC, Price AN, Jayakody SA, Martinez-Barbera JP, Greene NDE, Hawkes DJ, Ordidge RJ, Scambler PJ, Ourselin S, Lythgoe MF (2011) Magnetic resonance virtual histology for embryos: 3D atlases for automated high-throughput phenotyping. NeuroImage 54(2):769–778
9. Davidson D, Baldock R (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. Nat Rev Genet 2:409–418
10. Delaurier A, Burton N, Bennett M, Baldock R, Davidson D, Mohun TJ, Logan MP (2008) The mouse limb anatomy atlas: an interactive 3D tool for studying embryonic limb patterning. BMC Dev Biol 8:83
11. Dhenain M, Ruffins SW, Jacobs RE (2001) Three-dimensional digital mouse atlas using high-resolution MRI. Dev Biol 232(2):458–470
12. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, Masi CD, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nürnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, Grabbe F, Sieverding C, Fischer B, Schrader K, Brockmeyer M, Dettmer S, Helbig C, Alunni V, Battaini M-A, Mura C, Henrichsen CN, Garcia-Lopez R, Echevarria D, Puelles E, Garcia-Calero E, Kruse S, Uhr M, Kauck C, Feng G, Milyaev N, Ong CK, Kumar L, Lam MS, Semple CA, Gyenesei A, Mundlos S, Radelof U, Lehrach H, Sarmientos P, Reymond A, Davidson DR, Dollé P, Antonarakis SE, Yaspo M-L, Martinez S, Baldock RA, Eichele G, Ballabio A (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol 9(1):e1000582
13. Duchateau N, Craene MD, Piella G, Silva E, Doltra A, Sitges M, Bijnens BH, Frangi AF (2011) A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. Med Image Anal 15(3):316–328
14. Goldberg IG, Allan C, Burel J-M, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK, Swedlow JR (2005) The open microscopy environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol 6(5):R47
15. Han L, van Hemert JI, Baldock RA (2011) Automatically identifying and annotating mouse embryo gene expression patterns. Bioinformatics (Oxford, England) 27(8):1101–1107
16. Hawrylycz M, Baldock RA, Burger A, Hashikawa T, Johnson GA, Martone M, Ng L, Lau C, Larson SD, Larsen SD, Nissanov J, Puelles L, Ruffins S, Verbeek F, Zaslavsky I, Boline J (2011) Digital atlasing and standardization in the mouse brain. PLoS Comput Biol 7(2):e1001065
17. Hill B, Baldock RA (2006) The constrained distance transform: interactive atlas registration with large deformations through constrained distances. In: Proc DEFORM'06 – workshop on image registration in deformable environments, MRC Human Genetics Unit.
18. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A (2010) Gene expression atlas at the European bioinformatics institute. Nucl Acids Res 38(Database issue):D690–D698

19. Kaufman MH (1992) The atlas of mouse development. Elsevier Academic Press, London, revised 1995 edition
20. Kerwin J, Yang Y, Merchan P, Sarma S, Thompson J, Wang X, Sandoval J, Puelles L, Baldock R, Lindsay S (2010) The HUDSEN Atlas: a three-dimensional (3D) spatial framework for studying gene expression in the developing human brain. J Anat 217(4):289–299
21. Kirby BB, Takada N, Latimer AJ, Shin J, Carney TJ, Kelsh RN, Appel B (2006) In vivo time-lapse imaging shows dynamic oligodendrocyte progenitor behavior during zebrafish development. Nat Neurosci 9(12):1506–1511
22. Lawson KA, Meneses JJ, Pedersen RA (1986) Cell fate and cell lineage in the endoderm of the presomite mouse embryo, studied with an intracellular tracer. Dev Biol 115(2):325–339
23. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, Lin Chen, Li Chen, Chen T-M, Chin MC, Chong J, Crook BE, Czaplinska A, Dang CH, Datta S, Dee NR, Desaki AL, Desta T, Diep E, Dolbeare TA, Donelan MJ, Dong H-W, Dougherty JG, Duncan BJ, Ebbert AJ, Eichele G, Estin LK, Faber C, Facer BA, Fields R, Fischer SR, Fliss TP, Frensley C, Gates SN, Glattfelder KJ, Halverson KR, Hart MR, Hohmann JG, Howell MP, Jeung DP, Johnson RA, Karr PT, Kawal R, Kidney JM, Knapik RH, Kuan CL, Lake JH, Laramee AR, Larsen KD, Lau C, Lemon TA, Liang AJ, Liu Y, Luong LT, Michaels J, Morgan JJ, Morgan RJ, Mortrud MT, Mosqueda NF, Ng LL, Ng R, Orta GJ, Overly CC, Pak TH, Parry SE, Pathak SD, Pearson OC, Puchalski RB, Riley ZL, Rockett HR, Rowland SA, Royall JJ, Ruiz MJ, Sarno NR, Schaffnit K, Shapovalova NV, Sivisay T, Slaughterbeck CR, Smith SC, Smith KA, Smith BI, Sodt AJ, Stewart NN, Stumpf K-R, Sunkin SM, Sutram M, Tam A, Teemer CD, Thaller C, Thompson CL, Varnam LR, Visel A, Whitlock RM, Wohnoutka PE, Wolkey CK, Wong VY, Wood M, Yaylaoglu MB, Young RC, Youngstrom BL, Yuan XF, Zhang B, Zwingman TA, Jones AR (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445(7124):168–176
24. Marcon L, Arqués CG, Torres MS, Sharpe J (2011) A computational clonal analysis of the developing mouse limb bud. PLoS Comput Biol 7(2):e1001071
25. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Goualher GL, Feidler J, Smith K, Boomsma D, Pol HH, Cannon T, Kawashima R, Mazoyer B (2001) A four-dimensional probabilistic atlas of the human brain. J Am Med Inf Assoc (JAMIA) 8(5):401–430
26. McMahon AP, Aronow BJ, Davidson DR, Davies JA, Gaido KW, Grimmond S, Lessard JL, Little MH, Potter SS, Wilder EL, Zhang P, GUDMAP project (2008) GUDMAP: the genitourinary developmental molecular anatomy project. J Am Soc Nephrol (JASN) 19(4):667–671
27. Miles A, Bechhofer S (2008) SKOS simple knowledge organization system reference. W3C Recommendation
28. Ng L, Bernard A, Lau C, Overly CC, Dong H-W, Kuan C, Pathak S, Sunkin SM, Dang C, Bohland JW, Bokil H, Mitra PP, Puelles L, Hohmann J, Anderson DJ, Lein ES, Jones AR, Hawrylycz M (2009) An anatomic gene expression atlas of the adult mouse brain. Nat Neurosci 12(3):356–362
29. Paxinos G, Watson C (2004) The rat brain in stereotaxic coordinates – the new coronal set. 5th edition, Academic Press, Amsterdam, The Netherlands
30. Ringwald M, Baldock RA, Bard J, Kaufman MH, Eppig JT, Richardson JE, Nadeau JH, Davidson D (1994) A database for mouse development. Science (New York, NY) 265: 2033–2034
31. Rowe T, Frank LR (2011) The disappearing third dimension. Science (New York, NY) 331(6018):712–714
32. Sharpe J, Algren U, Perry P, Hill B, Ross A, Hecksher-Serensen J, Baldock R, Davidson D (2002) Optical projection tomography for 3D microscopy and gene expression studies. Science (New York, NY) 296:541–545 (The initial OPT paper).

33. Venkataraman S, Stevenson P, Yang Y, Richardson L, Burton N, Perry TP, Smith P, Baldock RA, Davidson DR, Christiansen JH (2008) EMAGE–Edinburgh mouse atlas of gene expression: 2008 update. Nucl Acids Res 36(Database issue):D860–D865
34. Walls JR, Coultas L, Rossant J, Henkelman RM (2008) Three-dimensional analysis of vascular development in the mouse embryo. PloS One 3(8):e2853
35. Weninger WJ, Geyer SH, Mohun TJ, Rasskin-Gutman D, Matsui T, Ribeiro I, Costa LDF, Izpisúa-Belmonte JC, Müller GB (2006) High-resolution episcopic microscopy: a rapid technique for high detailed 3D analysis of gene activity in the context of tissue architecture and morphology. Anat Embryol 211(3):213–221
36. Weninger WJ, Mohun T (2002) Phenotyping transgenic embryos: a rapid 3-D screening method based on episcopic fluorescence image capturing. Nat Genet 30(1):59–65

Index

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*,
Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1,