

> **Computational Systems Biology**

EDITED BY

Andres Kriete

Roland Eils



Elsevier Academic Press

30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

525 B Street, Suite 1900, San Diego, California 92101-4495, USA

84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. 

Copyright © 2006, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333,

E-mail: permissions@elsevier.co.uk. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>), by selecting

"Customer Support" and then "Obtaining Permissions."

Cover design: Computer simulation revealing patterns of concentrations based on the Gray-Scott model (courtesy of Systems Biology Group, Drexel University).

Library of Congress Cataloging-in-Publication Data

Computational systems biology / edited by Andres Kriete, Roland Eils.
p. cm.

Includes bibliographical references and index.

ISBN 0-12-088786-X (alk. paper)

1. Biological systems—Computer simulation. I. Kriete, Andres. II. Eils, Roland.

QH324.2.C638 2005

570'.1'13—dc22

2005020831

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 13: 978-0-12-

ISBN 10: 0-12-088786-X

For all information on all Elsevier Academic Press publications visit our Web site at www.books.elsevier.com

Printed in the United States of America

05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

Contents

Preface	ix
1. Introducing Computational Systems Biology <i>Andres Kriete and Roland Eils</i>	1
I. Enabling Information and Integration Technologies for Systems Biology	
2. Databases for Systems Biology <i>Jürgen Eils, Christian Lawerenz, Kathy Astrahantseff, Martin Ginkel, and Roland Eils</i>	15
3. Natural Language Processing and Ontology-enhanced Biomedical Literature Mining for Systems Biology <i>Xiaohua Hu</i>	39
4. Integrated Imaging Informatics <i>Bahram Parvin, Qing Yang, Gerald Fontenay, and Mary Helen Barcellos-Hoff</i>	57
5. Simpathica: A Computational Systems Biology Tool within the Valis Bioinformatics Environment <i>Bud Mishra, Marco Antoniotti, Salvatore Paxia, and Nadia Ugel</i>	79
6. Standards, Platforms, and Applications <i>Herbert M. Sauro</i>	103
II. Foundations of Biochemical Network Analysis and Modeling	
7. Introduction to Computational Models of Biochemical Reaction Networks <i>Frank J. Bruggeman, Barbara M. Bakker, Jorrit J. Hornberg, and Hans V. Westerhoff</i>	127
8. Biological Foundations of Signal Transduction and the Systems Biology Perspective <i>Ursula Klingmüller</i>	149
9. Reconstruction of Metabolic Network from Genome Information and Its Structural and Functional Analysis <i>Hong-Wu Ma and An-Ping Zeng</i>	169

10. Integrated Regulatory and Metabolic Models <i>Markus W. Covert</i>	191
III. Computer Simulations of Dynamic Networks	
11. A Discrete Approach to Network Modeling <i>Reinhard Laubenbacher and Pedro Mendes</i>	205
12. Gene Networks: Estimation, Modeling and Simulation <i>Seiya Imoto, Hiroshi Matsuno, and Satoru Miyano</i>	229
13. Computational Models for Circadian Rhythms: Deterministic Versus Stochastic Approaches <i>Jean-Christophe Leloup, Didier Gonze, and Albert Goldbeter</i>	249
IV. Multi-Scale Representations of Cells and Emerging Phenotypes	
14. Multistability and Multicellularity: Cell Fates as High-dimensional Attractors of Gene Regulatory Networks <i>Sui Huang</i>	293
15. Spatio-Temporal Systems Biology <i>Avijit Ghosh, David Miller, Rui Zou, Bahrad Sokhansanj, and Andres Kriete</i>	327
16. Cytomics—from Cell States to Predictive Medicine <i>Günter Valet, Robert F. Murphy, J. Paul Robinson, Attila Tarnock, and Andres Kriete</i>	363
17. The IUPS Physiome Project: Progress and Plans <i>Peter Hunter, Kelly Burrowes, Justin Fernandez, Poul Nielsen, Nic Smith, and Merryn Tawhai</i>	383
Subject Index	395

Preface

Computational systems biology, a term coined by Kitano in 2002, is a field that aims at a system-level understanding by analyzing biological data using computational techniques. The explosive progress of genome sequencing projects and the massive amounts of data generated by high-throughput experiments in DNA microarrays, proteomics, and metabolomics advances this field in a bidirectional but dependent fashion. As the need for a complete quantitative part list in biology is recognized, the understanding develops that living systems cannot be understood by studying just individual parts. Under the guiding vision of systems biology, sophisticated computational methods are currently being developed to analyze the data generated by this new technology in a systematic fashion, unraveling complex and networked biological phenomena, and modeling processes that take place in cells, tissues, and organisms. With recent advances in information technology, fast and inexpensive computer power, global networking infrastructure, and comprehensive databases, mathematical modeling and simulation of complex biological processes have become increasingly important and feasible. Modeling and simulation methods involve the use of different system analysis tools such as discrete mathematics and stochastics, differential equations, complex system simulation, as well as model-database integration architectures. The construction and testing of quantitative representations and models will be possible through the collaborative input of experimental and theoretical biologists working together with system analysts, computer scientists, mathematicians, engineers, physicists, and physicians to contend creatively with the hierarchical and nonlinear nature of cellular systems, while bioengineers will maintain a focus on directing the research results toward developing and improving cell-based, biotechnological processes.

This book has a distinct focus on computational issues related to systems biology. As such, it presents a timely, multi-authored compendium representing state-of-the-art computational technologies and methods developed in this area. This includes a review of enabling information and data integration technologies that have not been covered elsewhere in this depth. Modeling of gene, signaling and metabolic networks, being the main thrust in current computational efforts, is comprehensively covered. Contributions have been selected and compiled to introduce the different methods, including methods of abstraction, modeling of dynamical properties, and biological perspectives. A comprehensive coverage of computer representations of the multiple scales within a cell in relation to emergent properties in biological systems is also provided.

Beside the 17 primary authors and their respective teams who have dedicated their time to contribute to this book, there are many other individuals whose support was instrumental in making book a reality. We would like to in particular

thank Chip Coward, for editing, and reviewing and for making many useful suggestions, as well as Joel Beaudouin, Hauke Busch, Donald Coppeck, Rainer König, Sven Mesecke, Leo Neumann, Avijit Ghosh, Hannah Schmidt-Glenewinkel, Bahrad Sokhansanj, Markus Ulrich, Jörg Weimar and Ivayla Vacheva.

Both editors thank the team at Elsevier, in particular Luna Han, who supported this project from on the beginning. Without her generous support this book wouldn't have come into existence. We also thank Pat Gonzalez for an excellent technical supervision of the production.

It is often mentioned that biological systems in its entirety present more than a sum of its parts. To this extent, we hope that the chapters in this book, not only give a contemporary and comprehensive overview about recent developments, but that this volume advances the field and encourages new strategies, interdisciplinary cooperation, and research activities.

Andres Kriete and Roland Eils
Philadelphia and Heidelberg, May 2005

Introducing Computational Systems Biology

Andres Kriete and Roland Eils*

School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, Pennsylvania and Coriell Institute for Medical Research, Camden, New Jersey, USA

**Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, and Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg, Germany*

Chapter 1

Science is built up with facts, as a house is with stones. But a collection of facts is no more science than a heap of stones is a house.

—Jules Henri Poincarè (1854–1912)

I. INTRODUCTION

Contemporary biological information resides in some thousand public databases providing descriptive genomics, proteomics and enzyme information, gene expression, gene variants, and ontologies, supplemented by millions of scientific publications. Refined explorative tools, new genotyping techniques, and genome consortia efforts such as the ongoing international Haplotype Mapping project—along with the emergence of new profiling tools such as protein and cell arrays—constantly feed into this data pool and accelerate its growth (see Figure 1.1). Given the enormous and heterogeneous amount of data, computational tools have become indispensable in mining, analyzing, and connecting such information, which is often only interpretable under stringent consideration of how experiments were conducted. The aggregate of statistical bioinformatics tools for collecting, storing, retrieving, and analyzing complex biological data has repeatedly proven useful in biological decision support and discovery, a notable hallmark being the deciphering of the human genome as led by the Genome International Sequencing Consortium. Cataloging the basic building blocks of life is a necessary step in biological research, but this provides only limited knowledge in terms of understanding and predictability. In particular, the human genome project stirred the

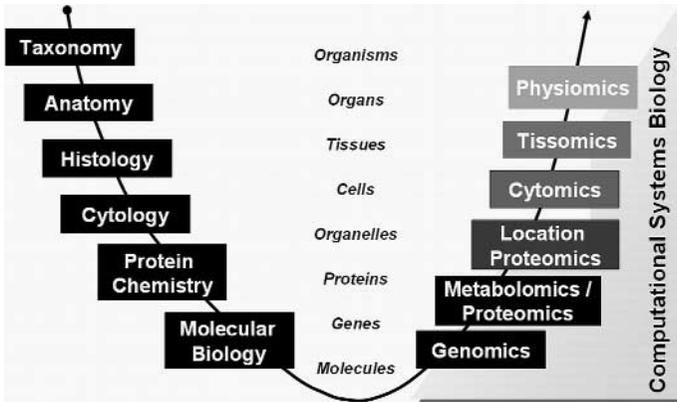


Figure 1.1. By the evolution of sub-disciplines in biology over time, ever-smaller structures have come into focus and more detailed questions have been asked. With the availability of high-throughput sequencing techniques in genetics, a turning point was reached regarding research on the molecular basis of life. The investigations extended to hypothesis-free data acquisition of biological entities, with genomics becoming the first in a growing series of “-omics” disciplines. Although functional genomics and proteomics are far from being completed, new type of approaches addressing the phenotypical cellular, tissue, and physiological levels constitute new specialized disciplines, filling up an otherwise sparse data space. Computational systems biology provides methodologies for combining, modeling, and simulating entities on diverse (horizontal) levels of biological organization, such as gene regulatory and protein networks, and between these levels by using multiscale (vertical) approaches. (After R. Eckner, Vienna.)

public expectation for a rapid increase in the identification of disease genes and development of more effective drugs and cures. However, these days it is well recognized that the many mechanisms involved in the proliferation of complex diseases such as cancer cannot be understood solely on the basis of knowing all of their molecular components.

As a consequence, a lack of system-level understanding of cellular dynamics has prevented any substantial increase in the number of new drugs available to the public and any increase in drug efficacy or eradication of any specific diseases. In contrast, pharmaceutical companies are currently lacking criteria for selecting the most valuable targets, research-and-development (R&D) expenses skyrocket, and new drugs rarely hit the market and often fail in clinical trials, while physicians face an increasing wealth of information that needs to be interpreted intelligently and holistically (Hood 2004).

Analysis of this dilemma reveals primary difficulties due to the enormous bio-molecular complexity, structural and functional unknowns in a large portion of gene products, and a lack of understanding of how the concert of molecular activities transfers into physiological alterations and disease. Exploring how cells interact with the environment, perform their tasks, and sustain homeostasis—or homeodynamics (Yates 1992)—as well as the role of inherited and epigenetic factors and evolu-

tionary constraints are at the forefront of urgent questions and at the heart of computational systems biology. We are at a very important turning point in biology, in that the ever-increasing quality and quantity of molecular data now provides the basis for building mathematical models of biological processes with increasing complexity.

An envisioned digital blueprint of complex diseases but also of biological development, aging, and immunity should not solely consist of descriptive charts as widely found in scientific literature or in genomic databases. They should instead be based on rigorously quantitative data-based mathematical models of metabolic pathways, signal transduction cascades, cell-to-cell communication, and so on. The general focus of biomedical research needs to change from a primarily steady-state analysis at the molecular level to a systems biology level capturing the characteristic dynamic behavior. Such concepts will likely transform current diagnostic and therapeutic approaches to medicine (Hood et al. 2004). How soon we will be able to predict physiology from the molecular capacity of cells remains an unknown, but it will be an important cornerstone in this grand scientific challenge.

For this ambitious program to succeed, computational systems biology is a key, providing data integration, network analysis and multiscale modeling (Kitano 2002). Additional components will require special attention, as pointed out by Bad Mishra in his contribution (Chapter 5). These include computational environments, research and pedagogic modeling tools that can be used by a novice user, rapid development of new biotechnological approaches, and creation of a catalog of illustrating examples by which these methodologies prove their power unambiguously. This book documents diverse ongoing attempts in these areas. In the following, we will broadly review the content of the chapters as they appear in this book, along with specific introductions and outlooks.

II. AREAS OF COMPUTATIONAL SYSTEMS BIOLOGY

A. Enabling technologies

Chapter 2—by J. Eils, C. Lawerenz, K. Astrahantseff, M. Ginkel, and R. Eils—discusses databases for systems biology to aggregate information about the responses of biological systems to genetic or environmental perturbations. As researchers try to solve biological problems at the level of entire systems, the very nature of this approach requires the integration of highly divergent data types. Moreover, computational systems biology also deals with models, simulations, and predictions. The concept of an integrative database as presented is therefore uniquely designed to tightly couple the three general areas of data generated in systems biology: experimental data, elements of biological systems, and mathematical models and their derived simulations.

In addition to some 800 public bioinformatics databases, information also resides in over 12 million abstracts accessible in PubMed, supplemented by an increasing

number of freely available full texts. These constitute an underutilized information resource. In Chapter 3, T. Hu describes the possibilities of natural language processing and ontology-enhanced biomedical literature mining for systems biology. It is important to develop efficient and effective technologies that automatically search large collections of biomedical literature and that extract and mine the important biological relationships (such as protein-to-protein interaction and function) so that domain experts can analyze this information to form new hypotheses, conduct new experiments, and facilitate new discoveries in systems biology research.

Data is not only generated by genomics sequencing and structural proteomics but increasingly by image-based spatial and time-lapse microscopic observations. B. Parvin's Chapter 4, on integrated image informatics, describes an imaging bioinformatics framework for cataloging protein localization and subcellular responses as a function of experimental factors. The underlying data model leverages new standards, assay development, and experimental designs. The presentation layer is web-based and utilizes a graphical interface to navigate through the annotation, data, and quantitative representations, whereas novel computational components enable multiscale representation of images.

In Chapter 5, B. Mishra, M. Antonioti, S. Paxia, and N. Ugel explain a computational systems biology tool within a bioinformatics environment. The group introduces Simpathica, used for modular and hierarchical modeling, simulation, and reasoning. The chapter discusses the construction of Simpathica in the rapid prototyping environment Valis, and its use in understanding signaling pathways.

Standards, platforms, and applications as presented by H. Sauro in Chapter 6 conclude this first part of this book. One of the trends indicative of cooperation within the systems biology community to emerge in recent years is the development of model interchange standards that permit biologists to exchange models between different software tools. Two exchange standards, SBML and CellML, are described by Sauro. Also discussed is the development of extensible software frameworks, including SBW, BioSPICE, and BioUML. Finally, the rich set of computational tools is introduced that is emerging as systems biology becomes a mainstream science.

B. Biological discovery by analysis and modeling of biochemical networks

The multitude of computational tools needed for systems biology research can roughly be classified into two categories (Kitano 2001): *system identification* and *behavior analysis*. Once the system has been identified and a model constructed, the system behavior can be studied, for instance, by numerical integration or sensitivity analysis against external perturbations. In molecular biology, system identification amounts to identifying the regulatory relationships between genes, proteins, and small molecules, as well as their inherent dynamics hidden in the specific kinetic and binding parameters.

System identification is arguably one of the most complicated problems in science. Whereas behavior analysis is solely performed on a model, model construction is a process tightly connected to reality. In many disciplines, model construction is interpreted as an iterative process. The modeling cycle begins with a reductionist approach, creating the simplest possible model. The modeling process generates an understanding of the underlying structures, as components are represented with mathematical and statistical concepts. The minimal model then grows in complexity, driven by new hypotheses that may not have been apparent from the phenomenological descriptions. Then, an experiment is designed using the biological system to test whether the model predictions agree with the experimental observations of the system behavior. The constitutive model parameters may be measured directly or may be inferred during this validation process. However, the propagation of error through these parameters presents significant challenges for the modeler. If data and predictions agree, a new experiment is designed and performed. This process continues until sufficient experimental evidence in favor of the model is collected.

Modeling approaches can be divided into bottom-up and top-down. In the bottom-up approach, we use a reductionist approach and study basic components and integrate these to find relevant patterns and functions, such as pathways. However, a bottom-up data-driven strategy as currently performed is limited in its capacity to translate the effect of perturbations in these pathways onto the cell as a whole. This approach is not effective in modeling multicellular entities (e.g., tissues) or organisms. In top-down mode, we start with the intact system and decompose it into its parts and interactions. Hereby, we establish our knowledge of the system, and attempt to disassemble it into functional modules. Breakdown of cellular function into computable entities uses the principle of modularity (Hartwell et al. 1999). A decomposition of the many cellular components into groups allows for modeling and simulations within reasonable time frames and may mimic an evolved biological property (Kitano 2004). The critical difference between these approaches occurs when components and interactions are not all known.

The section in this book dedicated to biochemical network analysis and modeling introduces this field with a contribution by F. J. Bruggemann, J. J. Hornberg, B. M. Bakker, and H. V. Westerhoff on the basics of computational models of biochemical reaction networks (Chapter 7). With the notion that cells are highly organized biochemical reaction networks consisting of interacting gene, metabolic, and signaling networks, systems biology focusses understanding on the functioning of cells in terms of the properties of and interactions between their constituent macromolecules. This chapter provides an overview of the methods available for analyzing structural, regulatory, and kinetic models of biochemical reaction networks composed of gene, metabolic, and signaling networks, as well as simulations of biochemical reaction networks and metabolic control analysis. Examples are included to illustrate the reviewed types of models and analyses.

In Chapter 8, U. Klingmüller reviews the biological foundations of signal transduction and the systems biology perspective. It is revealed that a deeper understanding of complex biological responses cannot be achieved by traditional approaches but requires a tight combination of experimental data and mathematical modeling. By combining computer simulations with experimental verification systems, the properties of signaling pathways such as cycling behavior or threshold response can be successfully identified. However, to analyze complex growth and maturation processes at a systems level, and to quantitatively predict the outcome of perturbations, further advances in both experimental and theoretical methodologies are demanded.

Chapter 9, by H.-W. Ma and A.-P. Zeng, provides an overview of the reconstruction of metabolic networks from the structural and functional analysis of genome information. Existing databases for gene-enzyme and enzyme-reaction relationships needed for the reconstruction of metabolic networks are introduced. Distinct mathematical representation of metabolic networks is explained, and results of structural analysis of large-scale metabolic networks are summarized. The comparative metabolic network analysis of a large number of fully sequenced organisms has revealed several fundamental organizational principles, such as the power law connection degree distribution and the “bow-tie” global connectivity structure. The authors present an example of how structural analysis can be used for functional modular analysis of metabolic networks.

M. Covert’s Chapter 10, on integrated regulatory and metabolic models, describes the reconstruction of functional metabolic and transcriptional regulatory networks and a modeling approach that allows simulation of network behavior for each network separately, as well as for the two networks combined. This process is placed in the context of model-driven biological discovery, and is illustrated with a case study of a genome-scale model, which was reconstructed and used in conjunction with experimental data to elucidate the regulatory and metabolic networks in *Escherichia coli*.

C. Model selection and simulation of dynamic cellular processes

Time-discrete dynamic systems models have long been used in biology. Biologic computer simulations require careful consideration as to the level of detail necessary for a representative model, because unnecessary detail will lead to models so complex that detailed numerical study would become highly cumbersome or impossible. Circadian rhythms provide a particularly interesting case study for showing how computational models can be used to address a wide range of issues extending from molecular mechanisms to physiological disorders.

Chapter 11 by S. Imoto, H. Matsuno, and S. Miyano explores the estimation, modeling, and simulation of gene networks. Important computational topics related to gene networks are outlined, including computational methods for estimating gene networks from microarray gene expression data—a contemporary problem. Subsequently, a software tool for modeling and simulating gene networks (based on

the concept of Petri nets) is introduced. The authors demonstrate the utility of this software for the modeling and simulation of a gene network for controlling circadian rhythms.

A discrete approach to top-down modeling of dynamic biochemical networks is reviewed by R. Laubenbacher and P. Mendes in Chapter 12. This chapter focuses on methods of constructing discrete dynamic models of biochemical networks from high-throughput experimental data sets, in terms of a reverse-engineering approach to accommodate the accelerating flux of new experimental observations. A time-discrete dynamic system description over a finite-state set serves as a framework. Modeling methods having their origin in computer algebra and the theory of Groebner bases provide a compact description of the entire space of possible models. The approach determines from that space a model that is minimal in the sense that it contains no components that vanish on the data set used to construct the model.

Deterministic versus stochastic approaches to computational models for circadian rhythms are explored by J.-C. Leloup, D. Gonze, and A. Goldbeter in Chapter 13. This chapter demonstrates requirements for models that possess a minimum degree of complexity. Autonomous chaos was obtained in a 10-variable model for circadian rhythms in *Drosophila* incorporating the formation of a PER-TIM complex, but not in the five-variable model based on PER alone. In this mammalian clock model, the addition of feedback loops demonstrates multiple sources of oscillatory behavior.

D. Multiscale representations of cells and emerging phenotypes

The term *complexity* is often associated with “unpredictability.” However, biological systems such as cells are quite robust and functionally stable (Buchanan 2002). As such, complexity in biology is on one hand related to the large diversity in elements (e.g., genes, proteins, and cells). Characterizing these elements can reveal variety in state space, as does protein activation or cell cycle. Furthermore, a multitude of interactions, nonlinearities, and feedback loops over levels of biological hierarchy contribute to an intricate network that appears to be a complex in terms of being not just complicated but emergent. Emergent behavior in complex systems arises if all constituents of the system observed on one level cannot explain the system properties on a coarser or higher level (Walleczek 2000). Proteins provide an illustrative and intuitive example in that they obtain function not only by their sequence of amino acids but through a process of folding that gives rise to particular 3D structures responsible for their functional capabilities.

It is commonly recognized that biological complexity is due to progressive evolution that brought along an increasing complexity of cells and organisms over time (Adami et al. 2000). This judgment coincides with the notion that greater complexity is “better” in terms of complex adaptive systems and the capacity for self-organization. Computer-based analysis and representation of emergent properties are new but essential fields in systems biology. The goal is to conceptualize and

abstract basic principles and to model biological structures, including higher levels of organization such as cells, tissues, and organs.

Modeling efforts have largely focused at a single level or scale, such as genomic or proteomic, cellular, tissue, organ, organ system, whole body, behavior, and population. Little current research is devoted to the development of tools, techniques, algorithms, and mathematical theory needed to integrate the continuum from the micro- to the macroscale in a seamless fashion. Multiscale modeling encompasses concepts in state space and across time scales. Different organizational levels—such as gene regulatory networks, modules, and pathways—may be nested in a hierarchical fashion (Oltvai and Barabasi 2002). Computer models representing spatiotemporal relationships are not limited to a specific resolution but can integrate multiscales, including abstractions suitable to functional physiological simulations (Kriete 1999; Bassingthwaite 2000; Noble 2002; Hunter 2003). Different scales may also be connected through parameters or coupling coefficients, novel numerical methods such as implicit solvers, and model coupling. The following chapters illustrate the necessity for embracing pathway details and spatiotemporal observations, and for simplifying abstractions in computational systems biology.

S. Huang's Chapter 14, on multistability and multicellularity cell fates as high-dimensional attractors of gene regulatory networks, explains how the high number of combinatorially possible expression configurations collapse into a few configurations characteristic of observable cell fates. The latter have been proposed to be high-dimensional attractors in gene activity state space. The biology of cell fate regulation from a systems perspective is reviewed. Two gene network models (small systems of differential equations and high-dimensional Boolean networks) are discussed to illustrate how molecular interactions produce multistability and attractors.

A. Ghosh, D. Miller, R. Zuo, B. Sokhansanj, and A. Kriete examine in Chapter 15 the extension of systems biology into the spatiotemporal realm. This contribution rests on the notion that models of the intricate networks in cells, so far described in a dynamic but otherwise dimensionless and "well-stirred" biochemical approximation, are limited. Yet, spatial and temporal heterogeneity of the cell, and processes such as diffusion, have to be considered in model construction. Both limitations and extensions of current modeling and computational approaches are investigated, and implemented in a newly developed software package.

Chapter 16, on cytomics from cellular states to predictive medicine, is a contribution by G. Valet, A. Tarnok, B. Murphy, P. Robinson, and A. Kriete. Cytomics is the systematic study of biological organization and behavior at the cellular level. It has developed out of computational imaging and flow cytometry. This approach is suited to the population of the data space at the cellular level, which appears to be rather sparse compared to genomics and proteomics information. The ability to perform high-content and high-throughput imaging and analysis to reveal complex cellular phenotypes will not only further our understanding of how cells and tissues carry out their functions but will provide insight into the mechanisms by which those

functions are disrupted. Cytomics not only provides a new framework for a spatiotemporal systems biology but may enrich personalized medicine.

The IUPS Physiome project—as described by P. Hunter, K. Burrowes, J. Fernandez, P. Nielsen, N. Smith, and M. Tawhai in Chapter 17—aims to facilitate the understanding of physiological function in healthy and diseased mammalian tissues by developing a multiscale modeling framework that can link biological structure and function across spatial scales. This requires an open-source internationally collaborative effort, including XML standards for encapsulating models, web-accessible model databases, and computational tools for authoring and visualizing models and running model simulations. Current progress and future plans for several target organs are discussed.

III. CHALLENGES IN COMPUTATIONAL SYSTEMS BIOLOGY

A. Advanced topics in computing and biological computing

Progress in computational systems biology is bound to our ability to develop advanced computing environments and methodologies that solve problems efficiently. Computational biology offers the most difficult algorithmic challenges and optimization problem in science today, involving large solution spaces and multiple goals (Karp 2002). Because many real-world problems are Non-deterministic Polynomial-time (NP) hard or even difficult to approximate, there are several important classes of “softer” meta-algorithms that although they offer no mathematical guarantee of performance apply well to practical problem instances. Applications include meta-heuristics to study gene expression data sets from DNA chips. Because the problem of estimating gene networks is NP-hard and exhibits a search space of super-exponential size, researchers are increasingly using heuristic algorithms for this task to reduce the search space to a biologically meaningful subspace to find optimal solutions in linear time. Related algorithmic problems are represented in this volume in the chapters by Imoto, Laubenbacher, Goldbeter, and Ghosh.

It appears that distinct computational disciplines—including a joint application of statistical bioinformatics, methods in computational neuroscience, and medical informatics (Wiemer et al. 2003)—will contribute to the progress in systems biology, as they help to more easily select crucial components on any level or between levels of biological organization as they change by disease or treatment. This is particularly true for high-throughput experimentation. Eventually the field will move to more automated learning and discovery strategies known in artificial intelligence (Weber et al. 2005), as models are defined more automatically and decisions and refinements of the best match and/or best prediction have to be made.

There is a great deal of science to be done in elucidating the mechanisms by which living cells store and process information. New biochemical tools and techniques based on these mechanisms are therefore also gaining attention. In partic-

ular, these findings will inevitably suggest new modes of biomolecular computing. Work is ongoing, both in assessing the theoretical and computational issues of molecular computing and in studying and improving the practical aspects of the biomolecular systems themselves. Although efforts can be traced back to early concepts in Turing machines (Turing 1952), this field has been motivated by a paper by Adleman (Adleman 1994), in which he showed how to use DNA to encode and solve a seven-city traveling salesman problem. The traveling salesman problem is a member of problems known as “NP-complete,” for which there are no known efficient algorithms on conventional computers. From the result, molecules that represented solutions to the problem could be isolated. The small volume of DNA used (100 microliters), the speed of computation (approximately 10^{14} operations/second), and the extremely small energy used (2×10^{19} operations/joule) were promising. Work is progressing in characterizing and improving the biochemical operations that can be performed, and in designing new architectures for biomolecular computers.

The ever-increasing mass of data being generated with heterogeneous technologies at different sites over the world requires entirely new strategies in computation on this data. As an example, genome-wide cell-based screens as RNAi knockdowns or overexpression of proteins in combination with cell arrays (Conrad et al. 2004; Erfle et al. 2004) typically produce several terabytes of data. Thus, it is no longer feasible to transfer this data over the Web for local computing. In many cases, it might be more useful to bring the computational process to the data. This, however, will impose enormous problems in terms of computational resources at the site of data. A solution to this problem is offered by the data grid. The data grid is the next generation of computing infrastructure providing intensive computation and analysis of shared large-scale databases, from hundreds of terabytes to petabytes, across widely distributed scientific communities. For such large data volumes, traditional infrastructure components for data management can no longer be applied. Presently, new concepts for large-scale data input/output and data management are being developed in various international efforts, such as the EU-funded EGEE project (Enabling Grid for E-science; www.eu-egee.org), the German D-grid consortium (www.d-grid.de), the U.S. DOE science grid (www.science.org), and the NSF-funded BIRN project (www.birn.net). These efforts facilitate collaborative scientific workloads, grid computing pipelines, and distributed file sharing.

B. De novo experimental designs for systems biology

Defined biological systems can provide the basis for developing quantitative frameworks in systems biology. Unicellular or engineered unicellular organisms are perhaps the most desirable subjects for developing models, as more rapid progress can be obtained when substantial genomic data are combined with the ability to carry out assays to completely define phenotypes with controlled environmental conditions under which proliferation and gene expression occur. Recently, “minimal” cells were obtained by either reducing the genome and silencing parts

of the functional machinery, as demonstrated in bacteria (Luisi 2002) or by taking a bottom-up bioengineering approach starting with cell-free extracts encapsulated in vesicles (Noireaux and Libchaber 2004). It has been recognized from these experiments that a higher number of active genes for the fine-tuning of essential functional processes leads to more efficient and stable cells, but that the regulation of gene activities is an increasingly important requirement for stability. The notion of regulation coincides with previous considerations of stability theories in physiology (Yates 1994). These minimal cells and bioreactors are ideal test cases for computational systems approaches. From this point of view, computational systems biology can greatly contribute to the newly emerging field of synthetic biology, in particular by a partnership between biology and engineering (Brent 2004).

C. Computational systems biology and the scientific community

Systems biology is in need of an ambitious interdisciplinary effort. This paradigm shift in biomedical research cannot be achieved by a few isolated research teams but requires a concerted action of many experts and departments at the local, national, and international levels—in such diverse areas as bioinformatics, molecular biology, cell biology, biochemistry, applied mathematics, theoretical physics, engineering, and biomedical research (Kitano 2005). Specifically, systems biology as a “synthetic science” fosters new collaborations between computational and/or modeling experts who have traditionally focused their models on the same system but at different scales, or collaborations between computational and/or modeling experts and experimentalists currently working on a single experimental scale. Major biology-oriented modeling activities are now supported at most federal agencies under titles such as Computational Biology, Bioinformatics, Quantitative Systems Biology, Biocomplexity, Modeling at the Nanoscale, and Multiscale Modeling.

The necessity for cooperation is evidenced in the formation of entire institutions devoted to systems biology (e.g., the independent Institute of Systems Biology in Seattle) and in the reorientation of research departments toward systems biology as recently put into practice at Harvard Medical School and MIT along with several other research institutions in the United States and worldwide (such as BioX in Zurich, Switzerland; Bioquant in Heidelberg, Germany; and the Systems Biology Institute of Tokyo, Japan). The need for cooperation within the systems biology field is even more greatly reflected in the creation of buildings and workplaces designed to encourage collaboration. Increasing interest in communication has also been demonstrated by the multitude of recently established conferences, workshops, and new Journals. Current conferences include ICSB 2005, RECOMB 2005, ISMB 2005, SysbioECAL 2005, IBSB 2005, and FOSBE 2005. Special journal issues include *Science* (March 2002), *Nature Biotech* (February 2004), *ChemBioChem* (October 2004), and *FEBS Letters* (April 2005). New journals include *IEESysBio*, *Nature Molecular Systems Biology*, and the *IEEE Journal of Systems Biology*.

IV. OUTLOOK

Many of the current efforts in systems biology look to integrate the results of today's scientific technologies responsible for the ubiquitous "data overload." The difficulty resides in converting data into information that provides insight and represents knowledge, as addressed by S. Brenner in his Nobel lecture (Brenner, 2003). The initial transition requires data cleansing and data coherency, but turning information into knowledge requires interpreting what the data actually means.

The ultimate goal of systems biology is the development and analysis of high-resolution quantitative models that recapitulate cellular behavior in time and space. Such models are the key to detailed understanding of biological functions, the diagnosis of diseases, the identification and validation of therapeutic targets, and the design of drugs and drug therapies (Lappe and Holm 2004). Experimental techniques yielding quantitative genomic, proteomic, and metabolomic data needed for the development of such models are now evolving. To meet the complexity of the accumulating data, to extract knowledge on the underlying cellular behavior, and eventually to construct predictive models, a broad spectrum of computer tools is required. Computer representations describing the underlying mechanisms may not always be able to provide complete accuracy due to limited computational, experimental, and methodological resources. Increases in data quality and coherence, availability within integrated databases, and approaches (such as fuzzy logic) that can manage experimental variability are less frequently considered but may be essential to the robust growth of in-silico representations (Mendes et al. 2004; Sokhansanj et al. 2005).

The enormous complexity of biological systems has given rise to additional cautionary remarks. First, it may well be that our models and future supermodels correctly predict experimental observations, but may still prevent a deeper understanding due to complexities, nonlinearities, or stochastic phenomena. This notion may initially sound quite disappointing, but is the daily experience of all those who employ modeling and simulations of large-scale phenomena, predominantly in physics. However, it shows the relevance of computational approaches in this area. In addition, recent progress in systems biology enables the discovery of common motifs—such as in regulatory gene networks and fundamental building blocks (Milo et al. 2002, 2004; Csete and Doyle 2004) of networks that have evolved over eons—and the discovery of principles of robustness and tolerance (Albert et al. 2000; Stelling et al. 2004). These findings also underpin the strength of computational models.

Secondly, systems biology should follow strict standards and conventions, but progress in theory and computational approaches will always demand new models that can elicit new insights if applied to an existing body of information. Once established, new models can be reimplemented in existing platforms to be more broadly available. In the long run, the aim is to develop user-friendly, scalable, open-ended platforms that also handle methods for behavior analysis and model-based disease diagnosis, that support scientists in their everyday practice of deci-

sion making and biological inquiry, and that support physicians in clinical decision making.

Systems biology has risen out of consensus in the scientific community, initially driven by visionary scientific entrepreneurs. Now, as its strength becomes obvious, it is recognized as a rapidly evolving mainstream endeavor that ties together various disciplines in a way that will move toward a formal, quantitative, and predictive framework of biology.

REFERENCES

- Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity. *PNAS* **97**(9):4463–4468.
- Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems. *Science* **266**(5187):1021–1024.
- Albert R., Jeong, H., and Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**:378–382.
- Bassingthwaighe, J. B. (2000). Strategies for the Physiome Project. *Annals of Biomedical Engineering* **28**:1043–1058.
- Brenner, S. (2003). Nobel lecture: Nature's gift to science. *Biosci Rep.* **23**(5/6):225–237.
- Brent, R. (2004). A partnership between biology and engineering. *Nature Biotech.* **22**(10):1211–1214.
- Buchanan, T. G. (2002). The community of the self. *Nature* **420**:246–251.
- Conrad, C. (2004). Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.* **200**(14):1130–1136.
- Csete, M. E., and Doyle, J. C. (2004). Bow ties, metabolism, and disease. *Trends in Biotechnology* **22**(4):446–450.
- Erfle, H., Simpson, J. C., Bastiaens, P. I., and Pepperkok, R. (2004). siRNA cell arrays for high-content screening microscopy. *Biotechniques* **37**(3):454–458.
- Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **2**(402/6761 Suppl.):C47–C52.
- Hood, L., Heath, J. R., Phelps, E. W., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventive medicine. *Science* **306**:640–643.
- Huang, S. The practical problems of post-genomic biology. *Nature Biotech.* **18**:471–472.
- Hunter, P. J., and Borg, T. K. (2003). Integration from proteins to organs: The Physiome Project. *Nature* **4**:237–243.
- Ideker T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**:343–372.
- Karp, R. (2002). Mathematical challenges from genomics and molecular biology. *Notices of the AMS* **94**(5):544–553.
- Kitano, H. (2001). *Foundations of Systems Biology*. Cambridge, MA: MIT Press.
- Kitano, H. (2004). Biological robustness. *Nature* **5**:826–837.
- Kitano, H. (2005). International alliances for quantitative modeling in systems biology. *Molecular Systems Biology* doi: 10.1038/msb4100011.
- Kriete, A. (1998). Form and function of mammalian lung: Analysis by scientific computing. *Adv. Anat. Embryol. Cell Biol.* **145**:1–105.

- Lappe, M., and Holm, L. (2004). Unraveling protein networks with near-optimal efficiency. *Nat. Biotech.* **22**:98–103.
- Mendes P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**(2):1122–1129.
- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of designed and evolved networks. *Science* **303**:1538–1542.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298**:824–827.
- Editorial. (2000). Can biological phenomena be understood by humans? *Nature* **408**:345.
- Noble, D. (2002). Modeling the heart: From genes to cells to the whole organ. *Science* **295**:1678–1682.
- Noireaux, V., and Libchaber, A. (2004). A vesicle bioreactor as a step toward an artificial cell assembly. *PNAS* **101**(51):17669–17674.
- Oltvai, Z. N., and Barabasi, A. L. (2002). Life's complexity pyramid. *Science* **298**:763–764.
- Sokhansanj, B. A., Fitch, J. P., Quong, J. N., and Quong, A. A. (2004). Linear fuzzy gene network models obtained from microarray data by exhaustive search. *BMC Bioinformatics* **5**(1):108.
- Stelling J., Sauer, U., Szallasi, Z., Doyle, F. J. III, and Doyle, J. (2004). Robustness of cellular functions. *Nature* **409**(15):860–921.
- Turing, A. (1952). The chemical basis for morphogenesis. *Phil. Trans. R. Soc. London B*(237):37–72.
- Walleczek, J. (2000). *Self-organized Biological Dynamics and Nonlinear Control*. New York: Cambridge University Press.
- Weber, R., Proctor, J. M., Waldstein, I., and Kriete, A. (2005). Case-base reasoning for modeling complex systems. In Muñoz-Avila, H., and Ricci, F. (Eds.) *Case-Based Reasoning Research and Development*. LNCS 3620:625–639, Berlin: Springer.
- Wiemer, J., Schubert, F., Granzow, M., Ragg, T., Fieres, J., Mattes, J., and Eils, R. (2003). Informatics united: Exemplary studies combining medical informatics, neuroinformatics, and bioinformatics. *Methods Inf. Med.* **42**(2):126–133.
- Yates, F. E. (1992). Order and complexity in dynamical systems: Homeodynamics as a generalized mechanics for biology. *Math and Computer Modeling* **19**:49–74.

Databases for Systems Biology

Jürgen Eils, Christian Lawerenz, Kathy Astrahantseff*, Martin Ginkel, and Roland Eils**

Division of Theoretical Bioinformatics, German Cancer Research Center, Heidelberg, Germany

** Division of Hematology/Oncology and Endocrinology, University Children's Hospital, Essen, Germany*

*** Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany*

Chapter 2

ABSTRACT

The ultimate goal of researchers in the interdisciplinary field of systems biology is to solve biological problems at the level of an entire system. Achieving this goal requires supporting the efforts of experimental biologists and computational modelers. Optimally, the phases of planning, actual experimentation, and data analysis (as well as model development, testing, and validation) would all be supported by one database solution. There is currently no integrative source for all information required in a computer-generated model of a biological system, and no system capable of providing support for all three phases of a systems biology endeavor. We present the concept of an integrative database for systems biology that functions as a data warehouse system and supports all three phases of a systems biology project.

This database system consists of three modules with different data models supporting the particular requirements of utilizing the three general types of data required: experimental data, *components*, and reactions of biological systems and mathematical models. The *model* and *experiment modules* are linked through the *component/reaction module*, eliminating the need to store complete information about any one entity more than once in the database. Complete functional models and simulations of particular interest are stored as SBML (Systems Biology Markup Language) files and linked to all necessary information within the database. This combination of modules tailored for dealing with the different data types and the interaction of these modules via links will meet the needs of researchers in the area of systems biology.

I. INTRODUCTION

A. Supporting systems biology

Systems biology attempts to integrate information about the responses of all elements in a biological system to genetic or environmental perturbations. The ultimate goal of researchers in this interdisciplinary field is to solve biological problems at the level of an entire system. To achieve this goal, computational models of the biological system are created that allow *in silico* simulation, and the application of mathematical methods from systems theory. The very nature of this systems approach requires the integration of highly divergent types of data and the efforts of experts in the areas of experimental biology, systems sciences, and applied computer science.

Wet-lab technologies are abundant and require expert knowledge to conduct the experiments, and in some cases to understand and interpret the results. Theoretical research on the development of mathematical models of a biological system also requires expert knowledge (as well as an entirely different vocabulary) to be able to uniquely explain the elements of a system. The experts from both of these areas, however, require information that is stored in a multitude of repositories all geared toward serving a specific clientele. As illustrated in Figure 2.1, although these areas overlap and synergize at several levels they are not well supported in an integrative manner.

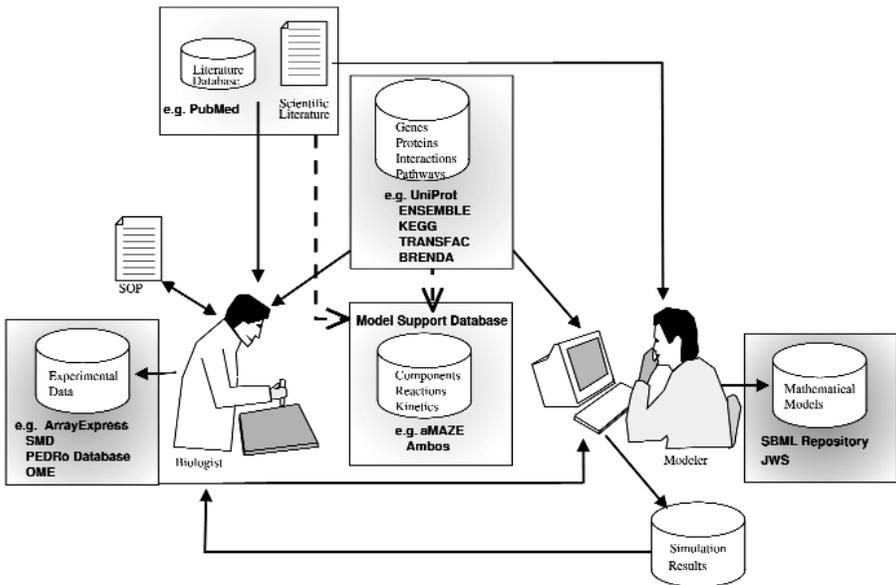


Figure 2.1. Current systems biology research workflow. Interactions among information sources, the experimental biologist, and the mathematical modeler.

B. Databases for biological data

1. Databases for elements

Currently, each of multitudes of databases covers a specific area of expertise. The majority of these are databases containing well-annotated descriptions of a class of biologically important elements, such as genes, proteins, classes, or active sites in proteins or entire genomes. An annotated collection of all publicly available gene sequences is maintained in GenBank at the National Center for Biotechnology Information (NCBI, most recently described in Benson et al. 2004). Ensembl was developed as a cooperative project by the Sanger Institute and European Bioinformatics Institute (EBI), and is a database that produces and maintains automatic annotation of metazoan genomes (Hubbard et al. 2002; Birney et al. 2004).

The most complete repository for protein information (including structure, function, classification, and experimental history) is the Universal Protein Resource known as UniProt (Apweiler et al. 2004). The Braunschweig Enzyme Database (BRENDA) is being developed into a metabolic network information system linking the enzyme description to information about expression and regulation (Schomburg et al. 2004). The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a suite of databases and associated software attempting to integrate knowledge about molecular interaction networks with information describing genes, proteins, chemical compounds, and reactions (Kanehisa 1997).

Several databases exist that contain information pertaining to transcriptional regulation. These include MAPPER for putative transcription factor binding sites; TRANSFAC for transcription factors, their genomic binding sites, and DNA-binding profiles; Cytomer for gene expression sources; S/MARt (Scaffold/Matrix Attached Regions database) for chromatin organizing regions; PathoDB for pathogenic forms of transcription factors; TRANSCompel for composite regulatory elements that are synergistically regulated by two factors binding to two closely positioned sites; and TRANSPATH for signal transduction pathways leading to transcriptional changes (Wingender et al. 2001; Marinescu et al. 2005).

A collaborative effort among the Cold Spring Harbor Laboratory, EBI, and the Gene Ontology Consortium has produced Reactome, a curated resource of core pathways and reactions in human biology (Joshi-Tope et al. 2005). Their data model allows the presentation of many diverse processes in the human system, as well as support for custom data entry, annotation, visualization, and exploration of the final data set.

2. Information resources databases

NCBI (National Center for Biotechnology Information) revolutionized searching and availability of information published in technical and scientific literature with the development of the PubMed and OMIM (Online Mendelian Inheritance in Man) databases (a current review of their resources is found in Wheeler et al. 2003). However, it remains a difficult problem to extract information from the published

sources, and usually requires reading and summarizing of the data by an expert. In addition to the databases with an emphasis on particular biochemical elements are databases that gather several types of information all pertaining to a particular organism, such as FlyBase for *Drosophila melanogaster*, WormBase for *Caenorhabditis elegans*, Xenbase for *Xenopus laevis* and *tropicalis*, Mouse Genome Informatics (MGI), and the TIGR (The Institute for Genomic Research) *Arabidopsis thaliana* database and the yeast virtual library for *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Candida albicans*.

Efforts to completely sequence the genomes of many organisms also provide sources for genomic information, although few of these are complete to date. The Gene Ontology Consortium has made great efforts to provide a gene ontology that can be applied to all organisms even as knowledge of the roles of genes and proteins in cells accumulates and changes (reviewed in Lewis [2005]).

3. Databases for experimental data

Some databases store annotated experimental data from a particular commonly used technical method. One method for examining gene expression at the level of an entire system is analysis using oligonucleotide or cDNA microarrays. The Stanford Microarray Database (SMD) assists in the storage and analysis of two-color microarray data, and has been amended to be able to store, retrieve, display, and analyze various proprietary raw data formats as well as being compliant with accepted standards including recommendations of the Microarray Gene Expression Data group (MGED) (Ball et al. 2002, 2005; Gollub et al. 2003). A fully open-source version of SMD is also available as the Longhorn Array Database (Killion et al. 2003). Both the Gene Expression Omnibus (GEO) and ArrayExpress are public repositories for annotated microarray data that comply with MGED recommendations.

GEO has been expanded to include not only microarray-based experiments but Serial Analysis of Gene Expression (SAGE) and mass spectrometry proteomic technology (Barrett et al. 2005). ArrayExpress can accept data in MAGE-ML (Micro Array Gene Expression Markup Language) format or via the MIAME (Minimum Information About a Microarray Experiment) online submission tool (Brazma et al. 2003). Several web-based solutions—for example, the BioArray Software Environment (BASE) (Saal et al. 2002) and the flexible iCHIP solution (www.dkfz.de/tbi/projects/dataManagement)—exist for the management and analysis of microarray experimental data, which can also be customized. SOURCE is a resource that integrates microarray data with complete gene reports describing alternative names, chromosomal location, functional descriptions, gene ontology annotations, expression data, and links to external databases (Diehn et al. 2003).

To support the area of proteomics research, a Proteomics Experimental Data Repository (PEDRo) makes comprehensive proteomics data sets available for browsing, searching, and downloading (Garwood et al. 2004). The UAB (the University of Alabama at Birmingham) Proteomics Database and SWISS-2DPAGE

provide links between protein spots identified on 2D gels and associated information obtained from mass spectrometric analysis (Hill and Kim 2003; Hoogland et al. 2004). The Expert Protein Analysis System (ExPASy) provides access to a variety of databases and tools useful for research on proteins and proteomics (Gasteiger et al. 2003). Many wet-lab techniques require microscopic images. The Open Microscopy Environment (OME) aims at providing a solution to the storage, analysis, and modeling of optical microscopic image data (Swedlow et al. 2003). In that the number and variety of databases is still expanding rapidly, the Public Catalog of Databases (DBcat) maintained by Infobiogen is a useful source of information (Discala et al. 1999).

4. *Databases supporting modeling*

Early efforts to integrate data necessary for modeling into a common database include Algorithms and Methods for the Development of Biochemical Ontology-based Database Systems (Ambos) (Rojas et al. 2002) and aMAZE (van Helden et al. 2000), both of which support the modeling process. They are based on a data model that allows storage of various types of biological *components* and their interactions. This information is in part imported from other specialized public databases and from scientific literature. The basic information about the players and interactions in the biochemical network is complemented with quantitative information about kinetics and some types of experimental results. Both systems contain tools for complex queries and visualization of biochemical networks, and Ambos can generate SBML (Systems Biology Markup Language) from the content of the database. However, a major common database solution for several data types does not exist in functional form.

Although these databases are good sources for specialized information about a particular topic, there is no integrative source for all of the information required in a computer-generated model of a biological system. To support research in the area of systems biology, a solution to this problem needs to be developed.

II. A DATABASE SOLUTION FOR SYSTEMS BIOLOGY

Computer-assisted modeling and analysis of biological systems requires many types of information. A database supporting research in systems biology must store and integrate complete models in a usable form, detailed descriptions of *elements* in the system, and experimental data and simulations. The data must be in a format that can be used in modeling and simulation, and needs to be preprocessed so that it is free of ambiguities. Details relating experimental context need to be stored with, or linked to, experiments. Finally, it must be possible to access all information required to reconstruct *in silico* experiments and models or to design wet-lab experiments to validate models. The sheer amount and variety of data requires special database solutions to support research in systems biology.

The concept of a database constructed of three modules is presented here as a solution to better integrate the three general areas of data: experimental data, *components* and reactions of biological systems, and mathematical models. Annotation of biomaterials, results from wet-lab experiments, and computer-based simulations of biological systems using existing models are stored in the *experiment module*. The *model module* contains functional models in a standardized format using SBML and the model annotation.

The third module, *component/reaction*, links the *experiment* and *model modules* to provide the complete descriptions of *elements* required by both modules. Information describing biochemical reactions (as well as events such as complex formation and the location of these activities) is critical for model development, and is also stored within this module. Figure 2.2 illustrates the areas that must be integrated to provide support for research in systems biology. Combining expertise in the areas of biology and mathematics/computer science makes the development of an integrative approach to the understanding of complex biological systems possible.

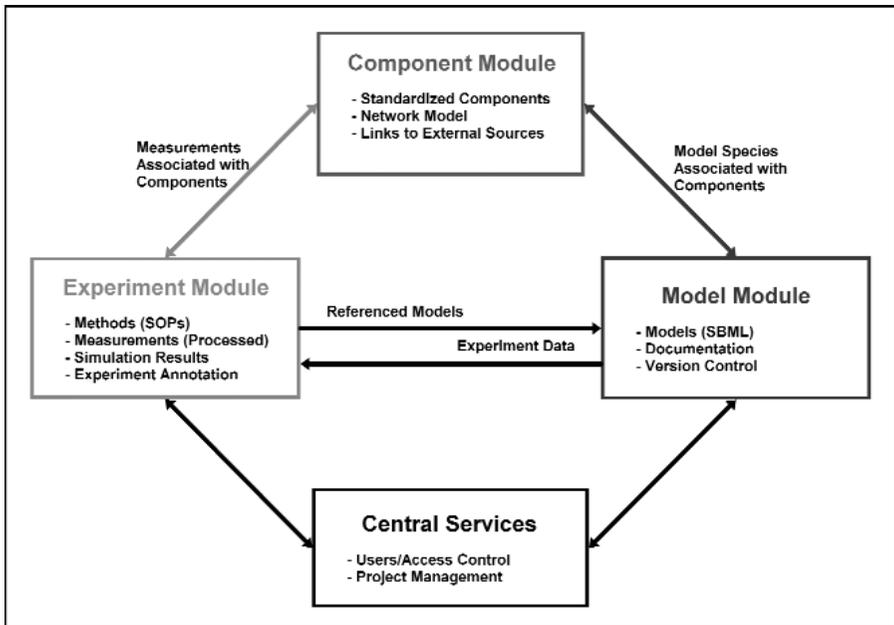


Figure 2.2. Interactive database architecture. *Components* within the *component* module are linked to the experiment block and model repository. Experiments are associated with the models based on them. Models are linked to the experiment results. The central services provide support for user management and features for project management.

A. The component/reaction module

1. *The players*

First, it is necessary to describe the types or categories of information used to analyze a biological topic using computer-assisted modeling and simulation. Common terminology used in the areas of biological and chemical research is often ambiguous, and quite often many terms exist for the same item. For instance, the same names can be used for genes, the enzymes they encode, and the reactions these enzymes catalyze. Biologists derive the correct information from the context, but this is not possible for the computer. Creation of a storage database that uniquely classifies and describes all potential players in models can help to better structure the addition of biological background information into the database, and to make data existing in databases freely exchangeable with other external modeling or simulation tools.

Some databases provide this curation of *elements*, including Ambos (curation of KEGG) and the TRANSFAC and TRANSPATH databases. The application of automated tools to extract data from existing databases and to integrate this data into a common database resource to be used as a supply for modeling and simulation tools would ease this enormous effort. Commercially available tools exist for the automatic extraction and integration of data from public sources, and include BioRS from Biomax, SRS from LION Bioscience, and DiscoveryLink from IBM. The players in a model are grouped according to what roles they play, and these categories are described in the following sections.

Components are the simplest and most central players in models. An experiment deals with *components*, and *components* take part in a reaction or pathway within a specific localization. *Components* are chemical entities, and can be macromolecules such as genes, proteins, or protein complexes—as well as small molecules such as metabolites or ions. A *species* describes a pool of one *component* with contextual information about the localization (to a *compartment*). *Reactions* describe the interactions between *species*, and can be influenced through other *species* (the *modifiers*), which include *elements* such as enzymes, activators, and inhibitors.

The term *reaction* describes, as in SBML terminology, real chemical reactions that transform *components*, as well as the formation of a macromolecular complex from its individual *components* or physical transport of *components* (such as the translocation of a protein through a membrane into another subcellular compartment). *Reactions* contain a description of their stoichiometry, describing the quantitative relationship of the *species* that take part in the reaction. Kinetics is described by the rate of the reaction, typically using a mathematical formula. There are a number of well-known kinetic laws (such as the Michaelis–Menten equation) that can be held as prototypes in the database, but modelers also combine these kinetic laws or define specially tailored ones to completely describe the biological situation. Finally, *reactions* can be combined to describe pathways. By combining these types of information, various cellular processes (such as metabolic networks, cell signaling, and gene regulation) can be modeled and simulations performed.

2. *Integrating components and reactions*

Most modeling systems to date manually curate only the data required to run in silico experiments, making this process time consuming. To reduce the effort put into creating support systems for computer-assisted modeling and simulation, the integrative database stores information describing *components* and *reactions* in one module. The *component/reaction module* contains all unique information about the players in a biological system to be modeled. This module is an integral part of the database, and effectively links the *model* and *experiment modules*. In this way, the complete description of each *component* is stored only once in the database. Other modules reference a *component* by forming a link to the required information. Properties of each reaction—including stoichiometry, kinetics, and whether a reaction is reversible—can also be stored in the *component/reaction module* and referenced by models.

Using references and by building synonym lists, commercially available tools for the automatic extraction and integration of data from public sources are capable of preprocessing data to avoid the ambiguous description of database entities. In addition, these tools are capable of performing periodic updates from public domain databases to ensure incorporation of newly discovered *components* and confirmation of references that ensure data quality. However, the data available in these public sources concerning a particular biological question may not be complete. In these cases, manual curation of the data to include the additional knowledge must also be possible without loss of the reference to the original public source of automatically imported *elements*.

B. The experiment module

1. *Experimental data and biological materials*

Currently, modelers do not typically utilize all information from an experiment, but collect only those pieces of information or values that seem relevant as input for the model. For example, they use measurements of protein concentrations or gene expression over time to reveal the dynamic behavior of their process of interest, and to unveil its basic structure. They may also use these measurements to fit quantitative model parameters so that their model can reproduce observed behavior of the biological system. The ultimate power of a model lies in its capacity for prediction. New perturbations of a biological system can be tested using in silico experiments on the model, and these results are compared with wet-lab experimental results for the same perturbation. In this way, a model can be validated in its ability to reproduce the behavior of the system correctly. This nevertheless requires that the information about the experimental procedures be exhaustive enough so that data from different biological systems and different labs can be integrated to build or to test the models.

To handle experimental data in a structured manner, we suggest a concept for the experimental annotation spanning the most important aspects of the experi-

ment, from the underlying goal through the biological samples, methods, and experimental results (Figure 2.3). The formal description of the experiment enables the experiment to be repeated by another laboratory under similar conditions, and provides more information to be used for modeling or simulation.

Experiments are conducted on various biological materials. These biomaterials are extremely diverse in their nature and description, making the task of obtaining all relevant information for each *biomaterial* complex. In addition, because the types of experiments conducted are also very different from one another the resulting data must be stored with specialized information to completely describe the experiment. The annotation stored about each type of *biomaterial* varies. For instance, to provide a clear picture of the experimental context and interpretation experiments conducted on resected tumor tissue require not only information about the preparation of DNA, RNA, or protein from this tissue for use in the experiment but histological classification of the tumor type and patient information. Whereas experiments conducted on cell lines require information describing culture conditions, passage numbers, cell line derivations and condition of the cells prior to the experiment (including culture confluency at treatment or harvesting, drug treatment time span, time in serum-free medium and so on) to be able to fully interpret the experiment.

The annotation of primary cell lines may also require clinical information about the patient from whom they come. Furthermore, what clinical factors are of importance depends on which disease is being described. Finally, the description of gene knock-down experiments using interfering RNA (RNAi), expression studies using oligonucleotide chip arrays or profiling of proteins in biological fluids using MALDI-MS (Matrix-Assisted Laser Desorption/Ionization Mass Spectroscopy or SELDI-MS (Surface-Enhanced Laser Desorption/Ionization Mass Spectroscopy each require very specific information in order to be comparable to experiments conducted in other laboratories. Storage of an exhaustive description of each *biomaterial* or method in the database is, at least, time intensive and probably utopian. A more realistic approach is the combination of expert-developed SOPs for the usage of classes of *biomaterials* and protocols for commonly used methods, with the storage of some specific and necessary information about the *biomaterials* and methods within the database.

The SOP used could then be referenced in the experiment, and all SOPs stored within the experiment module. Some form of version control is also required for SOPs, so that old versions remain available for analysis of older experiments (Figure 2.3). A flexible architecture is required for storing the necessary annotation for biological materials. The incorporation of an entity attribute model (EAV) is one way of handling the variety of attributes (Nadkarni and Brandt 1998). This data model utilizes an extensible table architecture that stores each attribute as a value (each as a single row) within a table. Attribute access is organized using meta-tables. In this way, all information is accessible, and new attributes can easily be added without changing the database structure.

Experimental data describing complex processes in systems biology stems from varied experimental methods (i.e., expression microarrays, mass spectrometric

analysis, 2D protein gels, RT-PCR (polymerase chain reaction), western blots, electrophoretic mobility shift assays, enzyme activity assays, and microscopic pictures). The raw data are often analyzed by commercially available software packages designed specifically for a certain system (e.g., Affymetrix chip image analysis). Many of these use a proprietary storage format, but most of them can export parts of the data also as Excel tables or as comma-separated values (CSVs). To support model development, it is first necessary to bring these primary data into a format that can be generally used by modeling and simulation systems.

Data required by the model is often stored in databases in the public domain and in local-experiment databases. These databases can be considered repositories of primary data, each containing only one type of data or data concerning one method or biological system. It would be difficult to integrate the wide range of heterogeneous primary data into one source. However, model development and simulation requires access to many types of data. Our goal is not to define the structure for all heterogeneous data types but to suggest a repository as a central platform for relevant experimental data required by the models. To include derived data in a general format in the repository, primary data must be preprocessed with specific software used for the special technology. In this way, these data are directly available to the model without the requirement of further preprocessing or special external software.

Figure 2.3. Universal modeling language (UML) diagram of the object model for the experiment module. Classes representing specific aspects of the experiment have different colors. The class *Contact* is associated to its subclasses, *Person* and *Organization*, and is derived from the *AuditAndSecurity* package of MAGE-OM. Attributes of these classes contain the contact information for the scientist providing *biomaterials* or conducting the experiment. An experiment utilizing biological material is represented in the classes, *Annotation* (experimental goal) and *CellCulture* (cell type and culture conditions). Its *Phase* subclass (ordered steps generating a series of measurements) consists of *Action* (information about experimental steps in a phase). *CellCulture* is linked to *BioMaterial* (description of biological material), whose specifications are dependent on the *biomaterial* (i.e. type, origin, gender, history, cell type, quality, physiological status, bioreactor type or biohazard status). Experimental values are represented by *numCols* of the *Table* class. Standard operating procedures (SOP) are separated into *ActionSOPs* (experimental method protocols) and *ProcessingSOPs* (data preprocessing). *ActionSOPs* describe, for example, the methods used to generate new biological materials from existing materials. *ProcessingSOPs* describe the strategy for deriving data based on primary data with the respective calculation method. Measurement values can be divided into nominal or ordinal discrete values, continual (real) values, stationary values, time-courses or relative values. The *values* attribute of the *Table* class includes a series of columns, each column of which describes one measurement (i.e., protein concentration or gene expression relative to the control). The *valueSemantics* attribute in *Column* describes the type of measurement (i.e., time, concentration, rate, activity, ratio, and so on) and the transformations of the values (i.e., logarithmic). The *Time-column* subclass describes the measuring time points in a time-course experiment. *NominalColumn* consists of nominal values (i.e., disease, gender, and so on), and *nominalSet* defines the set of applicable terms for the nominal value. *NumericColumn* contains all numeric measurements including statistical error estimations. If an entry in an object (represented by *Column*) contains a value of a *component*, the *species* must be referenced. The *Component* subclass of *Species* describes the compartment (i.e., cytoplasm, nucleus, and so on) where the *component* (in *Component*) is localized.

2. Standards

There are ongoing efforts to develop standards for reporting and storing experimental data from particular types of methods. The Minimum Information About a Microarray Experiment (MIAME) standard has been developed to support data export and description of microarray experiments with the goal of unambiguous experiment interpretation by the entire research community (Brazma et al. 2001). The MicroArray Gene Expression Markup Language (MAGE-ML) is based on the MAGE object model (MAGE-OM), and can be used for microarray data exchange (Spellman et al. 2002). The HUPO (Human Proteome Organization) nomenclature to facilitate data comparison, exchange, and verification in the area of proteomics has been developed by the Proteomics Standards Initiative (Orchard et al. 2003; Orchard et al. 2005). Recently, a commission was formed with the aim of setting standards for functional enzyme characterization, called Standards for Reporting Enzymology Data (STRENDA). The standards described in OME for dealing with microscopic data support projects using, for instance, RNAi screening and applications requiring multidimensional image storage and analysis. The XML (Extensible Markup Language) schema, OME XML, has been established to standardize data transfer (Swedlow et al. 2003). Microarray and proteomics standards have been incorporated into the Systems Biology Object Model (SysBio-OM), which supports the representation of microarray and protein expression data as well as data describing protein-to-protein interactions and metabolics (Xirasagar et al. 2004).

C. The model module

1. Models and standards

Models in this contribution are considered formal descriptions of intracellular networks and their environment for use in mathematical analysis. Models utilize information from scientific literature, experimental data, and existing models. In many cases, model descriptions are tailored specifically for a class of mathematical analysis methods. It is necessary to use well-defined terminology that abstracts the biological concepts in the model. However, purely mathematical descriptions containing equations and variables are not adequate because they have no direct connection to biological semantics and the information stored in biological databases.

In the integrative database solution described here, the *model module* acts as a repository of complete and functional models, including the metadata and documentation describing them. A model is typically not developed from scratch each time, but newly generated data or information is integrated into an existing model. Thus, modeling is an iterative process in which parts of existing models are often reused. A model can be extended for application to a new biological question by incorporating information related to the biological problem. This evolution of model development in particular requires a system that supports model version control.

Several tools exist that support the creation and analysis of mathematical models (Mendes 1997; Sauro 2000; Ginkel et al. 2003; Slepchenko et al. 2003; Takahashi et al. 2003). Application of a multitude of tools to one biological problem can be helpful, in that different tools apply different mathematical methods, which have inherent strengths and weaknesses based on the biological question being addressed. Thus, no one tool is capable of solving all problems of modeling and simulation in systems biology. This has been a motivator in the creation of tool-independent model exchange formats, most notably CellML (Cell Markup Language) and SBML (Hucka et al. 2003; Lloyd et al. 2004). Both formats are based on XML notation (Bosak and Bray 1999) and focus on the mathematical description of largely intracellular biochemical processes.

Whereas CellML considers a rather abstract approach of very general and extensible modular mathematical descriptions integrated into a larger family of XML-languages (Anatomical Markup Language (AnatML) and the Field Markup Language (FieldML)), SBML concentrates on the more practical task of developing a common exchange format for models based on descriptions found in various modeling and simulation tools. Tool developers can easily adopt SBML because the descriptive *elements* and attributes used are derived from existing modeling tools. Currently, SBML is supported by approximately 75 software tools, and may be considered the leading standard for model exchange in systems biology. Model descriptions in SBML provide the mathematical information necessary to run simulations or to perform mathematical analyses on the model. To utilize existing models, the SBML file for the model should be linked to background knowledge about the biological system. The integrative database described here stores models as SBML files in the *model module*, and links the model to relevant information stored in the *component/reaction* and *experiment modules*.

2. Model storage in the database

The approaches to storing mathematical model data can be categorized into model repositories and network model databases. Model repositories such as JWS (Java Web Simulation) or the SBML model repository (Schilstra 2002; Olivier and Snoep 2004) store complete and functional models with documentation. The biological context of the model is explained in HTML (HyperText Markup Language) documentation or some RDF (resource description framework) information provided with the model. The JWS approach is more focused on facilities for online simulation and visualization, which is well suited even to occasional and inexperienced users in that it requires no background knowledge or installation of special software. The user of the model repository can directly simulate the fully functional models or download them as a basis for modeling activities. However, these model repositories do not provide up-to-date biological background information for the model *elements*.

Network model databases, such as Ambos or aMAZE, attach the mathematical model information directly to the entities of a network database of *components*

and *reactions*, which is similar to the *component/reaction* module in this contribution. Therefore, the mathematical information becomes an integral part of a biological information system, and can be presented with up-to-date biological background information and links to data from experiments and the literature. These databases are able to ultimately generate an SBML model for an arbitrary part of the biochemical network, selected from the entities in the database. The SBML models generated are well suited to the subsequent generation of new models.

The system presented here combines the advantages of the two approaches and eludes some of their disadvantages. One important requirement for the presented solution is uploading models as SBML files, prepared by external modeling tools. These functional models are made in a specific context with specific assumptions that have an influence on the chosen mathematical description. These models very often contain artificial mathematical *elements*. For instance, one *species* is created in the model that describes all *elements* that can be defined as biomass. Such mathematical *elements* have no direct biological semantics and are completely meaningless outside the specific context of the original model. For this reason, it is difficult to attach mathematical descriptions of the artificial *elements* to entities of a network database.

Other examples of purely mathematical descriptions are equations (rules in SBML terminology) and events, which have no counterpart in the network database. Therefore, it does not make sense to extract this biologically irrelevant information from the SBML file. Moreover, different models can describe the same biological entity with different mathematical expressions or parameter values. For instance, a reaction is described with different kinetic laws in different models. Unlike network model databases, the presented approach avoids inconsistency within the database by leaving such contradictory information in the SBML file.

Long-term storage solutions, such as a database, need to reflect the most recent SBML standard (in addition to older versions) in its internal data structure. To store complete models as SBML files and evade extraction of all utilized *elements* avoids dealing with the rapid evolution of SBML itself. Because no effort must be made to keep the database up to date with current SBML standards, it becomes unnecessary to provide safe transitions for old data into the updated database structures. Uploading and downloading facilities can be easily implemented because the models are available as SBML files and many tools are compliant with the SBML standard.

One highlight of the strategy of including information describing the context of a model is the partial parsing of the SBML file to provide links to further descriptions of its biologically meaningful *elements*, such as *species*, *reactions*, and *compartments*. These *elements* are mapped to entities within the *component/reaction* module of the database, thus assigning globally unique identifiers to all entities. In addition, reaction parameters and start/global values can be parsed from the SBML file and stored within the database. Biologically relevant kinetics included in the SBML file may already exist in the *component/reaction* module (Figure 2.4).

In the case that biologically relevant kinetics included in the SBML file already exist in the *component/reaction* module, reaction parameters can be linked to these kinetics. All other *elements* in the SBML file for the model are left opaque, reducing the effort to implement data structure as well as changes required for new SBML versions. However, even the artificial and purely mathematical *elements* remain stored in the SBML file, allowing the user to export fully functional models. The *species*, *reactions*, and compartments of a model are linked to their full descriptions in the *component/reaction* module via their locally unique identifiers in the model file. This mapping can be constructed semiautomatically as follows. After parsing the model, the database is searched for names of *species*, *reactions*, and compartments used within the SBML model, and the user is asked to choose the final mapping for the *elements*. Because some *elements* may be new, and may not have descriptions stored within the database, the user can choose to create new entities in the database.

Some *elements* that are only meaningful in the context of one model can remain unmapped by classifying them as artificial entities. When an updated model is uploaded, the original model becomes an old version in the model history, but remains in the database to allow for consistent referencing. For a database to be useful on a long-term basis, complete metadata and documentation for each model must be included. Examples of the metadata required to make a model comprehensible include information describing the aim or hypothesis of the model as well as biological information such as cell type and organism. Reference to sources of scientific information used in the model (e.g., initial models, scientific literature used for choosing parameter values, and so on) is also important for model documentation. Finally, a graphical representation of the network described by a model can be stored as an image to allow a quick overview (Figure 2.5).

3. Simulation

Models are used to perform *in silico* experiments on a system using external simulation software. One could argue that with the stored models in the database every user can simply repeat all simulations, rendering it unnecessary to store the simulation results. However, this approach would require that special simulation software be installed on the local computer. In addition, operation of these software tools is often not trivial, requiring some experience with choosing the correct parameters for the model situation being simulated. It is preferable to store simulation results as simulation experiments in the database, especially when they show interesting or unexpected behavior. Reference must be made to the model and version used to create each simulation in order to allow repetition of the *in silico* experiments.

The importance of providing version control for the models becomes apparent because simulations created by different versions of the same model will differ, leading to simulations of a system that may be out of date because not all available information about the system was taken into account. Input values for each

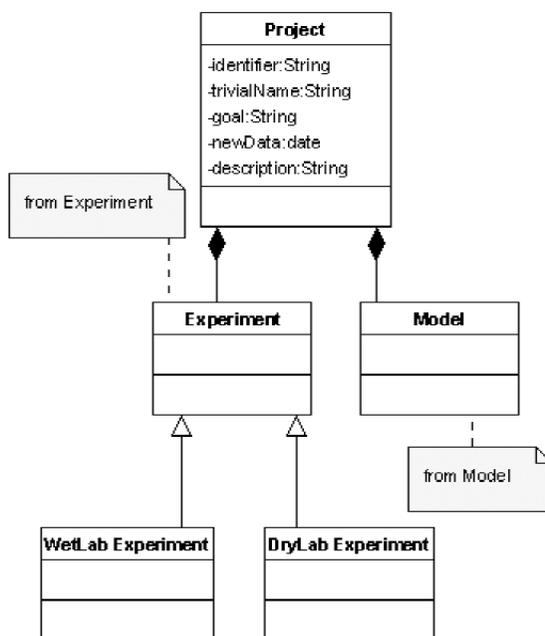


Figure 2.5. UML diagram for project management. The project groups all models and wet-lab and in silico experiments pertaining to the same biological question. By initiation of a project, the goal must be defined. Every experiment and model must be assigned to at least one project in the database.

simulation also need to be stored, including reaction parameters, different values describing the initial situation, and global parameters for the model. These values are considered changes to the default values for the original SBML model. The description of in silico simulation experiments mirrors the description of wet-lab experiments in which the initial situation of the *biomaterial* is described as well as instructions for performing phases of the experiment (simulation runs for the in silico experiment). The majority of stored simulation results describe values selected for the simulation that can be stored in a tool-independent ASCII format. These can then be imported into the database together with a description of the purpose of the experiment and model parameters.

III. PROSPECTIVE APPLICATIONS: USAGE AND WORKFLOW

The interactive database solution presented here provides an infrastructure to better support the workflow of interdisciplinary teams working in the area of systems biology. This interactive database platform is currently being developed for application in the Systems of Life-Systems Biology joint research project

(supported by the German Ministry for Research and Education). We present a general workflow for a generic systems biology project using the described interactive database, and include diagrams to better convey the steps and interactions in the following section (Figure 2.6).

Unanswered biological questions form the basis for a project. Models, experimental data, and simulations concerning one particular biological topic are connected as projects. The available data and references to scientific literature stored in the *component/reaction* module support the wet-lab scientist in investigating which players may be involved and what interactions may occur between or among them, in order to propose hypotheses that further direct the research. The database provides the known players and their interactions directly in a unified and structured manner. It contains directly browsable references to external information sources for scientific literature and online databases. This initial step in the project forms the basis for the following activities in the theoretical and experimental fields (Figure 2.5).

The experimental biologist designs wet-lab experiments to test defined questions based on the hypothesis. To properly design experiments it is necessary to assess the availability of *biomaterials* and their respective characteristics. This information is stored in the experiment module, making it directly accessible to the experimental biologist. To ensure that the experimental results are comparable with existing data and that experiments are repeatable by independent laboratories, the treatment and processing methods used must be available in SOPs stored in the database. Full-text searching and retrieval of SOPs enables the experimental scientist to locate an SOP for use in the planned experiment.

The primary data generated by the wet-lab experiment must first be pre-processed in accordance with accepted standards. These are available as processing SOPs in the experiment module. The resulting secondary data are stored in the experiment module with links to documentation for the methods and standards used. If the measurements refer to specific entities in the cell, the necessary mapping to the *component/reaction* module must be constructed to link the contextual information. Time courses from the wet-lab and *in silico* experiments conducted on the same system can be graphically overlaid for easy comparison of the results.

On the theoretical side, the modeler uses the existing information and newly generated results from the wet-lab experiments to design and implement a hypothesis to answer the central question as a model. This requires an overview of the network in question. This overview guides the decision of which analysis method can best be applied given the available measurements. The iterative modeling process starts after the analysis method has been chosen. Ideally, models already exist in the database that are in part applicable for reuse in the generation of the new model. Experimental results are required to determine the initial and parameter values. The resulting and executable model can be made public in the database.

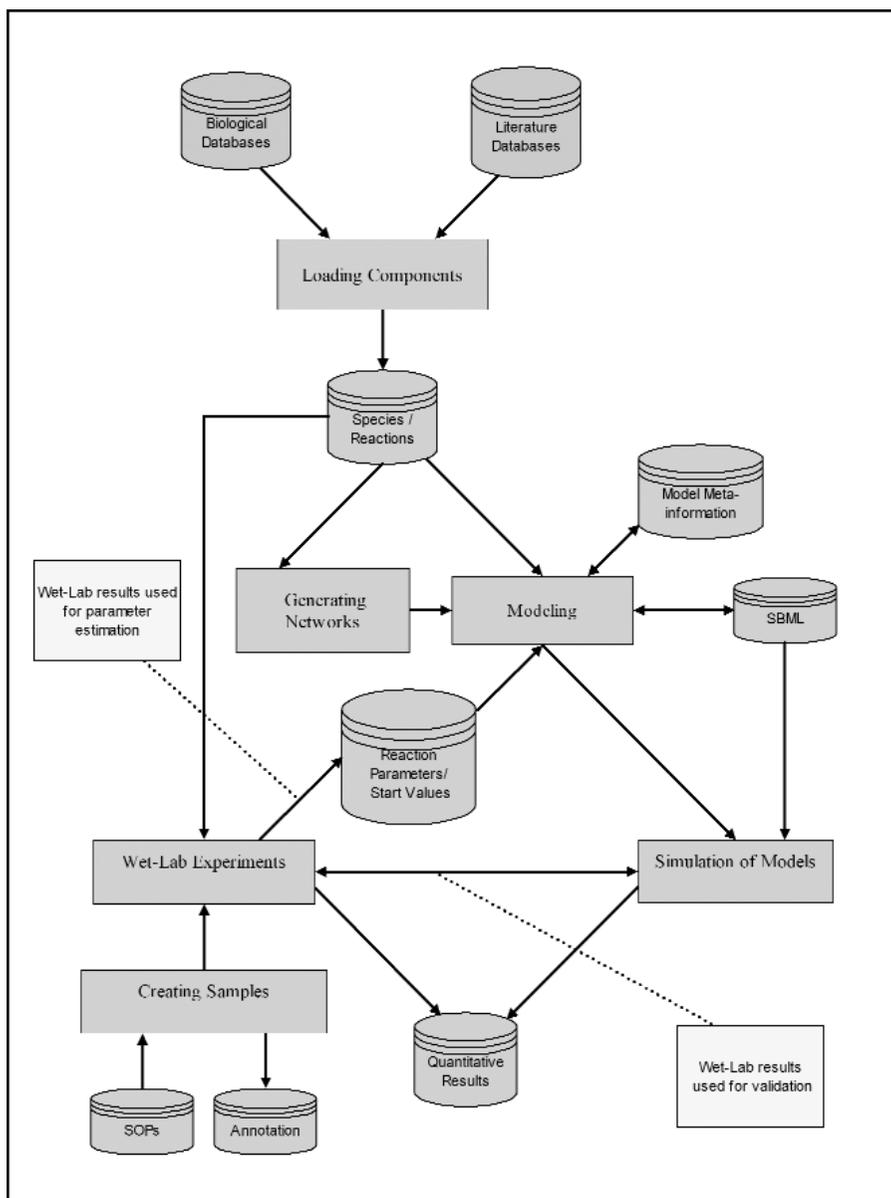


Figure 2.6. Overview of project workflow utilizing the integrative database solution for systems biology. The objects of external sources, such as biological databases and literature databases are shown in red. The objects of processes, whose results are stored in the integrative platform as well as the results, are indicated in blue. The integration tool extracts the *species* and reaction information from the external databases, and stores it in the systems biology database. *Reactions*, *species*, pathways and networks are linked to the respective model. Simulations for the project are carried out using a model, which is stored as an SBML file. Simulation results are used to design new wet-lab experiments, that either validate or disprove parts of the model. Models apply reaction parameters and start values based on wet-lab experiments. *Biomaterial* provided for the experiments is annotated, and the specific handling and treatment is described by SOPs.

The model is then used to conduct *in silico* experiments for various physiological or experimental situations. Start values can be extracted from experiments stored in the experiment module. The simulation results either comply with the experimental results (supporting the hypothesis of the model) or are in conflict with the experimental results (rejecting the hypothesis and leading to redesign or improvement of the model). The model may generate predictions for start values and parameters that have not been used for the model, or proposals for new designs for both wet-lab and *in silico* experiments. If this is the case, the simulation results may be of interest to the experimental biologists as well, and it is beneficial to store the simulation in the *model module*. Because the *species* in the model are mapped to the *component/reaction* module, it is possible to query the database for newly acquired models and experimental results that may be relevant for the model. The model may confirm the estimated start values and parameters initiating the model as well as lead to proposals for new designs of both wet-lab and *in silico* experiments.

If different experiments must be performed or different models developed to answer the biological question, a new project can be created in the database to group the results of these activities. The description of the biological question and the central goal is the primary documentation for a project. Existing models and *in silico* and wet-lab experiments can be assigned to the project.

The concept for an integrative database presented here is designed to better integrate the three general areas of data generated in systems biology (experimental data, *elements* of biological systems, and mathematical models) with derived simulations. Division of the storage of data from each of these areas into separate modules designed to handle their specific needs is a primary advantage of this system. In addition, each module has its own advantages that contribute to the system. The experiment module stores only preprocessed secondary data. This saves the space and effort of including primary data and all tools necessary to work with these primary data. The *component/reaction* module utilizes automated integration tools to incorporate information from publicly available databases, reducing the time and effort required for annotation and data input.

The *model module* stores complete working versions of models and important *in silico* experiments linked to the appropriate experimental data, *elements*, SOPs, and full documentation. The combination of grouping certain types of information into individual modules with the reintegration of only the relevant entities within these modules via mapping and linking will provide an excellent database solution for systems biology that supports both experimental biologists and mathematical modelers.

ACKNOWLEDGMENTS

The concept presented in this paper was developed in cooperation with the Data Management Task Force (Jens Timmer and Titus Sparna, University of Freiburg;

Reinhard Guthke, HKI-Jena; Jan-Michael Heinrich, University of Leipzig; Klaus Mauch, IBVT, University of Stuttgart; Ursula Kummer and Isabel Rojas, EML Heidelberg; and Lothar Terfloth, University of Erlangen) within the Systems of Life-Systems Biology joint research project supported by the German Ministry for Research and Education.

REFERENCES

- Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., and Yeh, L. (2004). UniProt: The universal protein knowledge base. *Nucleic Acids Res.* **32**(Database issue):D115–D119.
- Ball, C., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., Ringwald, M., Spellman, P., Stoeckert, C. Jr., Stewart, J., Taylor, R., Brazma, A., and Quackenbush, J. (2002). Standards for microarray data. *Science* **298**(5593):539.
- Ball, C., Awad, I., Demeter, J., Gollub, J., Hebert, J., Hernandez-Boussard, T., Jin, H., Matese, J., Nitzberg, M., Wymore, F., Zachariah, Z., Brown, P., and Sherlock, G. (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* **33**(Database issue):D580–D582.
- Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: Mining millions of expression profiles—database and tools. *Nucleic Acids Res.* **33**(Database issue):D562–D566.
- Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. (2004). GenBank: Update. *Nucleic Acids Res.* **32**(Database issue):D23–D26.
- Birney, E., Andrews, T., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyra, E., Fernandez-Suarez, X., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004). An overview of *Ensembl*. *Genome Res.* **14**(5):925–928.
- Bosak, J., and Bray, T. (1999). XML and the second-generation web. *Sci. Am.* **280**(5):89–93.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C., Causton, H., Gaasterland, T., Glenisson, P., Holstege, F., Kim, I., Markowitz, V., Matese, J., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**(4):365–371.
- Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., Oezcimen, A., Rocca-Serra, P., and Sansone, S. (2003). ArrayExpress: A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**(1):68–71.
- Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J., Hernandez-Boussard, T., Rees, C., Cherry, J., Botstein, D., Brown, P., and Alizadeh, A. (2003). SOURCE: A unified genomic

- resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31**(1):219–223.
- Discala, C., Ninnin, M., Achard, F., Barillot, E., and Vaysseix, G. (1999). DBcat: A catalog of biological databases. *Nucleic Acids Res.* **27**(1):10–11.
- Garwood, K., McLaughlin, T., Garwood, C., Joens, S., Morrison, N., Taylor, C., Carroll, K., Evans, C., Whetton, A., Hart, S., Stead, D., Yin, Z., Brown, A., Hesketh, A., Chater, K., Hansson, L., Mewissen, M., Ghazal, P., Howard, J., Lilley, K., Gaskell, S., Brass, A., Hubbard, S., Oliver, S. G., and Paton, N. (2004). PEDRo: A database for storing, searching, and disseminating experimental proteomics data. *BMC Genomics* **5**(1):68.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**(13):3784–3788.
- Ginkel, M., Kremling, A., Nutsch, T., Rehner, R., and Gilles, E. (2003). Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics* **19**(9):1169–1176.
- Gollub, J., Ball, C., Binkley, G., Demeter, J., Finkelstein, D., Hebert, J., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J., Schroeder, M., Brown, P., Botstein, D., and Sherlock, G. (2003). The Stanford Microarray Database: Data access and quality assessment tools. *Nucleic Acids Res.* **31**(1):94–96.
- Hill, A., and Kim, H. (2003). The UAB Proteomics Database. *Bioinformatics* **19**(16):2149–2151.
- Hoogland, C., Mostaguir, K., Sanchez, J., Hochstrasser, D., and Appel, R. (2004). SWISS2DPAGE, ten years later. *Proteomics* **4**(8):2352–2356.
- Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pockock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res.* **30**(1):38–41.
- Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., Arkin, A., Bornstein, B., Bray, D., Cornish-Bowden, A., Cuellar, A., Dronov, S., Gilles, E., Ginkel, M., Gor, V., Goryanin, I., Hedley, W., Hodgman, T., Hofmeyr, J., Hunter, P., Juty, N., Kasberger, J., Kremling, A., Kummer, U., Le Novere, N., Loew, L., Lucio, D., Mendes, P., Minch, E., Mjolsness, E., Nakayama, Y., Nelson, M., Nielsen, P., Sakurada, T., Schaff, J., Shapiro, B., Shimizu, T., Spence, H., 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**(4):524–531.
- Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* **33**(Database issue):D428–D432.
- Kanehisa, M. (1997). A database for post-genome analysis. *Trends Genet.* **13**(9), 375–376.
- Killion, P., Sherlock, G., and Iyer, V. (2003). The Longhorn Array Database (LAD): An open-source MIAME-compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics* **4**(1):32.
- Larman, C. (1998). *Applying UML and Pattern*. Upper Saddle River, NJ: Prentice-Hall.
- Lewis, S. (2005). Gene ontology: Looking backwards and forwards. *Genome Biol.* **6**(1):103.
- Lloyd, C., Halstead, M., and Nielsen, P. (2004). CellML: Its future, present, and past. *Prog. Biophys. Mol. Biol.* **85**(2/3):433–450.

- Marinescu, V., Kohane, I., and Riva, A. (2005). The MAPPER database: A multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res.* **33**(Database issue):D91–D97.
- Mendes, P. (1997). Biochemistry by numbers: Simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**(9):361–363.
- Nadkarni, P., and Brandt, C. (1998). Data extraction and ad hoc query of an entity-attribute-value database. *J. Am. Med. Inform. Assoc.* **5**(6):511–527.
- Olivier, B., and Snoep, J. (2004). Web-based kinetic modeling using JWS Online. *Bioinformatics* **20**(13):2143–2144.
- Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics* **3**(7):1374–1376.
- Orchard, S., Hermjakob, H., Binz, P., Hoogland, C., Taylor, C., Zhu, W., Julian, R. Jr., and Apweiler, R. (2005). Further steps towards data standardisation: The Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25–27(th) October, 2004. *Proteomics* **5**(2):337–339.
- Rojas, I., Bernardi, L., Ratsch, E., Kania, R., Wittig, U., and Saric, J. (2002). A database system for the analysis of biochemical pathways. *In Silico Biol.* **2**(2):75–86.
- Saal, L., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., and Peterson, C. (2002). BioArray Software Environment (BASE): A platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3**(8), software 0003.1–0003.6.
- Sauro, H. (2000). Jarnac: An interactive metabolic systems language. In H. Bolouri and R. C. Raymond (eds.). "Computation in cells: Proceedings of an EPSRC Emerging Computing Paradigms Workshop," pp. 11–18, Dept. of Computer Science Technical Report No. 345, University of Hertfortshire, UK.
- Schilstra, M. (2002). The SBML model repository. www.sbml.org/models.
- Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: Updates and major new developments. *Nucleic Acids Res.* **32**(Database issue):D431–D433.
- Slepchenko, B., Schaff, J., Macara, I., and Loew, L. (2003). Quantitative cell biology with the Virtual Cell. *Trends Cell Biol.* **13**(11):570–576.
- Spellman, P., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B., Robinson, A., Bassett, D., Stoeckert, C. Jr., and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**(9), research 0046.1–0046.9.
- Swedlow, J., Goldberg, I., Brauner, E., and Sorger, P. (2003). Informatics and quantitative analysis in biological imaging. *Science* **300**(5616):100–102.
- Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y., and Tomita, M. (2003). E-Cell 2: Multi-platform E-Cell simulation system. *Bioinformatics* **19**(13):1727–1729.
- van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D., and Wodak, S. (2000). Representing and analyzing molecular and cellular function using the computer. *Biol. Chem.* **381**(9/10):921–935.
- Wheeler, D., Church, D., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schriml, L., Sequeira, E., Tatusova, T., and Wagner, L. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**(1):28–33.

- Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29**(1):281–283.
- Xirasagar, S., Gustafson, S., Merrick, B., Tomer, K., Stasiewicz, S., Chan, D., Yost, K. III, Yates, J. III, Sumner, S., Xiao, N., and Waters, M. (2004). CEBS object model for systems biology data, SysBio-OM. *Bioinformatics* **20**(13):2004–2015.

Natural Language Processing and Ontology-enhanced Biomedical Literature Mining for Systems Biology

Xiaohua Hu

College of Information Science and Technology, Drexel University, Philadelphia, Pennsylvania

Chapter 3

ABSTRACT

Despite the rapid electronic dissemination of research results and experimental data in bioinformatics research, most biological knowledge and experimental results still exist only in text formats. Retrieving and processing this information is made difficult due to the large volumes and the lack of formal structure in the natural language narrative in those documents. It is very important to develop efficient and effective technologies that automatically search large collections of biomedical literature, extract and mine the important biological relationships such as protein-protein interaction, functionalities of the genes, etc, so that domain experts can analyze this information to form new hypotheses, conduct new experiments and enable new discoveries in systems biology research. In this chapter, we discuss some of the latest natural language processing and data mining techniques in this area and demonstrate their usefulness in chromatin protein interaction study and microarray data analysis.

I. INTRODUCTION

Most bioinformatics knowledge and experimental results are published only in plain-text documents. These documents, or their abstracts, are collected in biomedical literature databases such as MedLine (www.ncbi.nlm.nih.gov/entrez/query.fcgi), BioMedCentral (www.biomedcentral.com/), and so on. The large number of documents in such databases and the lack of formal structure in the natural language narrative in those documents make the search and processing very difficult to many scientists involved in bioinformatics research. To expedite the

progress of bioinformatics, it is essential to develop efficient and effective natural-language processing and text data mining techniques from this ever-expanding collection of biomedical literature so that genomic and medical experts can analyze this information to form new hypotheses, conduct new experiments, and enable new discoveries.

Natural-language processing and biomedical literature data mining have been applied to a wide range of bioinformatics problems such as extraction of protein interaction (Ono et al. 2001), microarray data interpretation (Kankar et al. 2002), and so on. Biomedical literature databases tend to have large collections of text files, cover a wide range of topics, and grow very fast. In this chapter, we discuss some of the latest natural language processing and data mining techniques in this area and demonstrate their usefulness through application in chromatin protein interaction and microarray data analysis.

A. Text clustering and summarization in biomedical literature

Many data mining methods and algorithms have been adapted to mine biomedical literature (Hirschman et al. 2002; Rzhesky et al. 2004). We review some relevant existing methods and algorithms in the following.

A significant limitation of the current clustering approach in microarray data analysis is that most of these algorithms provide no biological interpretation of the cluster results. Users need to discover and interpret the biological similarities that may underlie the expression pattern by cross-referencing the experimental results in related literature or functional annotations in various genomic databases. Because a gene cluster may include dozens or even hundreds of different genes, it is beyond the limits of biological researchers to detect and organize these data along multiple lines of conceptual similarity by inspecting them manually. Thus, it is essential to develop a system capable of gathering biological information and extracting and summarizing relevant information in a well-organized and coherent manner for the gene cluster. A variety of approaches to provide a biological explanation of gene clusters have been developed. TextQuest (Iliopoulos et al. 2001) is geared toward summarizing documents retrieved in response to keywords-based search on PubMed. It does not retain the association between the genes (keywords) and the retrieved documents.

MedMiner (Tanabe et al. 1999) can provide summarized literature information on genes but is limited to finding relations between two genes only. It also returns a few hundred sentences. Shatkay et al. (2000) suggested a system that attempts to find functional relations among genes on a genome-wide scale, but this requires users to specify a representative document for each gene, which describes the gene very well. Looking for the representative document may require a lot of time, effort, and knowledge on the part of the user. In addition, as genes have multiple biological functions it is very rare to find a document that covers all aspects of genes across various biological domains. GEISHA (Blaschke et al. 2001) is based on a comparison of the frequency of abstracts linked to different gene clusters and contain-

ing a given term. Interpretation by the end user of the biological meaning of the terms is facilitated by embedding them in the corresponding significant sentences and abstracts and by establishing relations with other equally significant terms.

Automatic text summarization has recently become an active research field related to many other research areas, such as IR (information retrieval), IE (information extraction), natural language processing, and machine learning (Goldstein et al. 2000; Hahn and Mani 2000). A variety of approaches exist for determining the salient sentences in the text and then synthesizing them to form a summary report: statistical techniques based on word distribution (Salton et al. 1994), symbolic techniques based on discourse structure (Marcu 1997), and semantic relations between words (Barzilay et al. 2001). Other recently addressed text summarization research topics have been multi-document summarization, multilingual summarization, and hybrid multisource summarization. A knowledge-based text summarization has also been addressed by Hahn and Reimer (1999), emphasizing the potential of the concepts and conceptual relations as a vehicle for terminological knowledge representation.

B. Biomedical ontologies

Biomedical ontologies are critical to the understanding and processing of biomedical literature (Sarkar et al. 2003; Bard and Rhee 2004). In general, bioinformatics has a so-called "communication problem" (Schulze-Kremer 1997). Even the meanings of high-level fundamental concepts are often ambiguous. For example, biology researchers have been suffering from inconsistent descriptions of gene products and ambiguous term definitions from disparate biology databases, which also hamper the semantic computational processing of biomedical literature such as text summarization or document clustering. The use of ontology would be a very promising solution. Currently, there is no ontology that captures the entire range of concepts in the biomedical domain. However, there are several well-designed biomedical ontologies, such as the UMLS (Unified Medical Language System) (www.nlm.nih.gov/research/umls/), the Gene Ontology (GO) (www.geneontology.org/), and the EcoCyc Ontology (Karp 2000).

In our current research, we focus on UMLS and GO because UMLS and GO are very well supported, extensively used, and freely available in (respectively) the medical and bioinformatics fields. UMLS consists of three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST. SPECIALIST is a lexicon that provides a mechanism for integrating all major biomedical vocabularies, including MeSH. GO describes gene products in terms of molecular functions, biological processes, and cellular components, and provides controlled vocabularies for the description of three independent ontologies (molecular function, biological process, and cellular component of gene product). Their combined use covers a wide range of topics in the biomedical domain.

II. ONTOLOGY-ENHANCED BIOMEDICAL LITERATURE MINING

A. NLP for automatic pattern generation and evaluation based on mutual bootstrapping for robust and portable IE

This section examines an NLP-based prototype system named SPIE (Scalable and Portable Information Extraction), shown in Figure 3.1. This system addresses efficient and effective information retrieval and extraction from large biomedical literature databases (Hu et al. 2004b). Our preliminary study on protein-to-protein interaction indicates that SPIE has significant advantages over traditional keyword-based search and IE methods. SPIE retrieved 0.5 million abstracts to obtain 9,000 unique protein-to-protein interactions, while traditional keyword based search methods retrieved about 1.5 million abstracts from MedLine to obtain the same number of unique protein-to-protein interactions (Hu et al. 2004b).

A crucial step in the extraction process is the generation of new patterns, which is accomplished by grouping the occurrences of known patterns in documents that occur in similar contexts. A good pattern should be selective but have high coverage so that it does not generate many false-positive relationships and can identify many new relationships. Another issue is that individual relationships of interest may be found in multiple contexts within collections. In deciding which putative relationships should be extracted, a key problem is how to combine evidence across the multiple occurrences of these relationships. The heart of our approach is a mutual bootstrapping approach that learns extraction patterns from the relation-

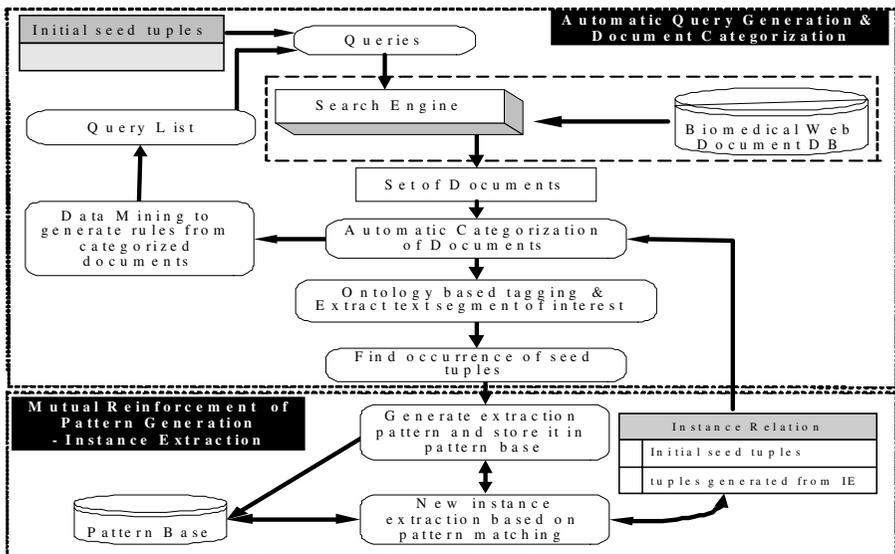


Figure 3.1. The architecture of SPIE.

ships and then exploits the learned extraction patterns to identify more relationships that belong to the relation. Our goal was to automate the construction of both the relationships and extraction patterns using bootstrapping.

As input, our technique requires only unlabeled texts and a handful of seed relationships of the relation. We use mutual bootstrapping techniques to alternatively select the best extraction patterns and bootstrap its extractions into the relation, which is the basis for selecting the next extraction. The mutual bootstrapping algorithms work well, but the performance can rapidly deteriorate when low-quality or spurious relationships enter the relation. To make this approach more robust, we add a second level of bootstrapping (meta-bootstrapping), which combines the multiple evidences of each relationship generated from the multiple matched patterns in the relationship ranking procedure. It will retain only the most reliable ones produced by mutual bootstrapping, and then restart the process with the enhanced patterns. This two-tiered bootstrapping process is less sensitive to noise than a single level of bootstrapping and produces high-quality relations and patterns.

Pattern generation: IE systems are commonly based on pattern matching. Each pattern is applied to each text segment, instantiating appropriate slots in the pattern with entities from the document. For example, a protein-to-protein interaction pattern in our approach is a tuple (or expression) consisting of two protein names that correspond to some conventional way of describing interaction. We can use these patterns to characterize those sentences that capture this knowledge. For every such protein pair p_1, p_2 it finds segments of text in the sentences where p_1 and p_2 occur close to each other and analyze the text that connects p_1 and p_2 to generate patterns.

For example, our approach inspects the context surrounding chromatin protein HP1 and HDAC4 in the sentence "HP1 interacts with HDAC4 in the two-hybrid system . . ." to construct a pattern {" ", Protein, "interacts with", <Protein>, " "}. After generating a number of patterns from the initial seed examples, it scans the available sentences in search of segments of text that match the patterns. As a result of this process, it generates new relationships and evaluates them and uses the most reliable ones as the new "seed." It then starts the process again by searching for these new relationships in the documents to identify new promising patterns.

SPIE pattern representation uses Eliza-like patterns (Weizenbaum 1966) that can make use of limited syntactic and semantic information. It represents the context around the related entities in the patterns in a flexible way that produces patterns that are selective and yet have high coverage. As a result, minor syntactic variations (such as an extra comma or a determiner) will not stop us from matching contexts that are otherwise close to our pattern. More specifically, SPIE represents the left, middle, and right "context" associated with a pattern, just like the vector-space model of information retrieval represents documents and queries. A pattern is a 5-tuple (*left*, tag_1 , *middle*, tag_2 , *right*), where tag_1 and tag_2 are named-entity tags, *left* are arbitrary strings of nonspace characters before tag_1 , *middle* are those nonspace

characters between tag_1 and tag_2 , and $right$ are those non-space characters after tag_2 in the text segment. $left$, $middle$, or $right$ may be empty in the pattern.

To match text portions with the 5-tuple representation of patterns, SPIE also associates an equivalent 5-tuple with each document portion that contains two named entities with the correct tag. After extracting the 5-tuple representation of string S , SPIE matches it against the 5-tuple patterns by taking the inner product of the corresponding $left$, $middle$, and $right$ segments. Our approach is based on an extended version of Harris's distributional hypothesis (Harris 1985), which states that words that occur in the same contexts tend to be similar. Instead of using this hypothesis on words, we apply it to sentences in the document. To learn patterns from these sentences, we use a sentence alignment method to group similar patterns and then learn each group separately for the generalized patterns. In our method, by aligning sentences similar parts in sentences could be extracted as patterns.

The similarity score $Match(T_i, T_j)$ between two 5-tuples $T_i = \langle l_i, tag_{i1}, m_i, tag_{i2}, r_i \rangle$ and $T_j = \langle l_j, tag_{j1}, m_j, tag_{j2}, r_j \rangle$ is defined as $Match(T_i, T_j) = W_{left} * S(l_i, l_j) + W_{middle} * S(m_i, m_j) + W_{right} * S(r_i, r_j)$, where W_{left} , W_{middle} , and W_{right} are the weight for (respectively) the left, middle, and right segments. We develop a new sentence alignment function to evaluate the similarity of two sentence segments such as l_i and l_j , which are ordered lists of words, numbers, punctuation marks, and so on. The advantage of using sentence alignment for similarity measurement is that it is flexible and can be implemented efficiently based on dynamic programming. The same idea is also used in comparing the similarity between protein or DNA sequences (Gusfield 1997). Given two sentence segments $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, the similarity score $S(i, j)$ is defined as the score of the optimal alignment between the initial segment from x_1 to x_i of X and the initial segment from y_1 to y_j of Y [$'_'$ denotes a white space, $S(i, 0) = 0$, $S(0, j) = 0$].

$$S(i, j) = \max \left\{ \begin{array}{l} 0, \\ S(i-1, j-1) + f(x_i, y_j) \\ S(i-1, j) + f(x_i, '_') \\ S(i, j-1) + f(' ', y_j) \end{array} \right\} \quad (3.1)$$

$$f(x_i, y_j) = \log \frac{p(x_i, y_j)}{p(x_i) * p(y_j)} \quad (3.2)$$

Here, $p(x_i)$ denotes the appearance probability of word x_i and $p(x_i, y_j)$ denotes the probability that x_i and y_j appear at the same position in two text segments. Following the same method proposed in Li et al. (2004), probabilities $p(x_i)$, $p(x_i, y_j)$ can be estimated by Equations 3.3a and 3.3b with prealigned training data.

$$p(x_i) = \frac{[C(x_i) + 1]}{\sum_{all\ x_i} [C(x_i) + 1]} \quad (3.3a)$$

$$p(x_i, y_j) = \frac{[C(x_i, y_j) + 1]}{\sum_{\text{all pairs}(x, y)} [C(x, y) + 1]} \quad (3.3b)$$

Here, $C(x_i)$ denotes the count of word x_i appearing in the training corpus, and $C(x_i, y_j)$ denotes the number of aligned pairs (x_i, y_j) being observed in the training set. For sentence segments X with a length of m and Y with a length of n , totally $(m + 1) * (n + 1)$ scores will be calculated by applying Equation 3.1 recursively. The scores are stored in a matrix as $M = M(x_i, y_j)$. Through back-tracing in M , the optimal local alignment can be searched.

Evaluation of patterns and relationships: Estimating the reliability of the pattern to ignore patterns that tend to generate bogus relationships is one of the problems we need to address. We can weigh the patterns based on their selectivity, and trust the relationships they generate accordingly. Thus, a pattern that is not selective will have a low weight. The relationships generated by such a pattern will be discarded, unless they are supported by other selective patterns. The case for relationships is analogous. The reliability of the relationships is a function of the selectivity and the number of patterns that generate it. At each iteration, the relationships in the relation are constantly growing and the extraction patterns need to be rescored. SPIE evaluates the quality of these patterns and relationships, and retains only the most reliable ones for the next iteration. A relationship may be produced by multiple patterns simultaneously. The scoring heuristic is based on how many different relationships a pattern extracts and how many relationships are extracted by this pattern only. We adapt a metric originally proposed by Riloff (1996) to evaluate extraction patterns generated by the IE system, and define the confidence of pattern P_i as

$$Conf(P_i) = (F_i/N_i) * \log(F_i) \quad (3.4)$$

where F_i is the number of unique relationships among the extractions produced by P_i and N_i is the total number of unique relationships that P_i extracted. One interesting, yet largely unexplored, aspect of IE is that potential relationships usually occur redundantly in text files. In deciding which putative relationships should be extracted, a key problem is how to combine evidences across the multiple occurrences of these relationships. We present a statistical method for addressing it. For each relationship T_j , we store the set of patterns $P = \{P_i\}$ ($i = 1, \dots, m$) that produce it, together with the measure of similarity match $Match(T_j, P_i)$ between the context in which the relationship occurred and the matching pattern P_i . The confidence of a candidate relationship T_j is defined as

$$Conf(T_j) = 1 - \prod_{i=1}^m (1 - (Conf(P_i) * Match(T_j, P_i))) \quad (3.5)$$

Thus, in the previous formulas $Conf(T_j)$ is not simply the count of the relevant relationships but their cumulative relevance. Formulas 4 and 5 capture the mutual dependency of patterns and relationships. After determining the confidence of the

candidate relationships using the previous definition, our method discards all relationships with low confidence because these low-quality relationships could add noise to the pattern generation process, which would in turn introduce more invalid relationships, degrading the performance of the system.

B. Ontology-enhanced text clustering and summarization for microarray data interpretation

A big limitation of the current clustering approach in microarray analysis is that most of these algorithms provide no biological interpretation of the cluster results. We developed (Hu et al. 2004) a robust system, GE-Miner (Gene Expression Miner), to integrate cluster ensemble and text mining to overcome this limitation. Here we further enhance GE-Miner by integrating biomedical ontology. Our objective is not to address all research issues raised in biomedical text clustering and summarization. As a complement to a number of relevant techniques—such as text mining, automatic text classification and others—we focus on exploiting our capability to automatically extract and analyze knowledge from text documents through text clustering and summarization. We develop a hybrid approach for integrating ontology-based text clustering and summarization, as shown in Figure 3.2.

First, we preprocess the texts, enriching their representations by background knowledge provided by the biomedical ontology. Then, we cluster the documents by a frequent term set-based partitioning clustering method, which partitions the large number of documents into a relatively small number of clusters. The document clusters are then analyzed by a multi-document summarization method. Within each cluster, saliency scores for key terms and sentences are generated based on the mutual reinforcement principle. Then, the key terms and sentences are ranked according to their saliency scores and selected for inclusion in the top key terms list and summaries of the documents.

In our method, enriching the term vectors with concepts from ontology has three benefits. First, it resolves synonyms; second, it introduces more general concepts, which help identify related topics; and third, the higher-level concepts from ontology subsume the lower-level concepts/primitive words in the term vector (thus reducing the dimensions and in turn improving the clustering accuracy and effi-

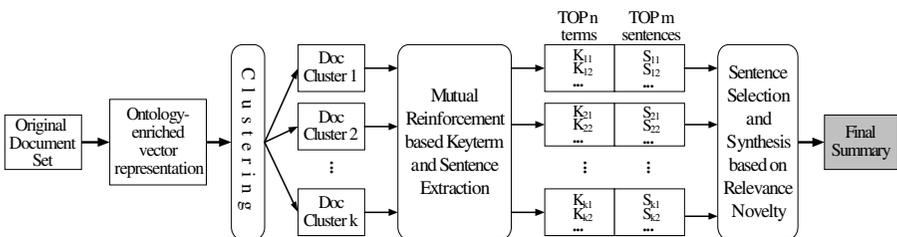


Figure 3.2. Data flow of text clustering and text summarization.

ciency). Ontology-based summaries can actually capture the essential information of textual documents at different levels of granularity. This issue is closely related to the compression rate and coverage/diversity of summaries (Hahn and Mani 2000).

1. *Frequent term set-based concept clustering*

The standard clustering algorithms can be categorized into partitioning algorithms such as k-means or k-medoid and hierarchical algorithms such as single-link or average-link (Han and Kamber 2001). A recent study (Steinbach et al. 2000) has clearly indicated that k-means methods have outperformed hierarchical clustering algorithms on a broad variety of text databases. These methods of text clustering, however, do not really address the special problems of text clustering: very high dimensionality of the data, very large size of the databases, and understandability of the cluster description (Beil et al. 2002).

Existing text clustering solutions use all words except the *Stop* words from the documents in their term vectors, thus generating a very high dimensional vector. They also only relate documents that use identical terminology, whereas they ignore conceptual similarity of terms and relationships between words such as synonyms, hyponyms, and hypernyms defined in terminological resources in ontology. Thus, semantically identical but differently spelled terms (e.g., *cancer*, *malignant tumor*) are treated as completely different words in traditional document clustering approaches. Such words hamper document similarity measurement (Hotho et al. 2003).

It is possible to reduce the dimensionality by selecting only frequent words from each document, or to use some other method to extract the salient features of each document. However, the number of features collected using these methods still tends to be very large, and due to the loss of some of the relevant features the quality of clusters tends not to be good. Other, more general, methods—such as principal component analysis (PCA) and latent semantic indexing (LSI)—have also been proposed for dimensionality reduction, which attempt to transform the data space into a smaller space in which relationships among data items is preserved. An inherent problem with dimensionality reduction is that in the presence of noise in the data it may result in the degradation of the clustering results (Steinbach et al. 2000).

This has motivated the development of new special text clustering methods that are not based on the vector space model. Suffix-tree clustering (Zamir and Etzioni 1998) is the first method following this approach. The drawback of suffix-tree clustering is that although two directly neighboring basic clusters in the graph must be similar two distance nodes (basic clusters) within a connected component do not have to be similar at all. In our system, we use frequent terms (items) for text clustering.

A frequent-term-based concept clustering is promising because it provides a natural way of reducing the large dimensionality of the document vector space. A

well-selected subset of the set of all frequent term sets can be considered a clustering. Strictly speaking, a frequent term set is not a cluster (candidate) but only the description of a cluster. The corresponding cluster itself consists of the set of documents containing all terms of the frequent terms. Unlike the case of classification, there are no class labels to guide the selection of such a subset from the set of all frequent term sets. Instead, we propose to use the mutual overlap of the frequent term sets with respect to their sets of supporting documents (the clusters) to determine a clustering.

The rationale behind this approach is that a small overlap of the clusters will result in a small classification error when the clustering is later used for classifying new documents. Our clustering algorithms are designed to efficiently handle very high dimensional spaces, without the need for dimensionality reduction. In contrast to traditional clustering methods, our proposed methods are linearly scalable, an advantage that makes them particularly suitable for use in regard to large collections of biomedical literature.

A frequent-term-based approach of clustering method first finds sets of terms that occur frequently together in documents using association rule discovery methods. These frequent term sets correspond to a set of documents that have a sufficiently large number of features (words or terms) in common, and are mapped into hyperedges in a hypergraph (Han et al. 1998; Beil et al. 2002). The similarity among documents is captured implicitly by the frequent term sets. The hypergraph representation can then be used to cluster relatively large groups of related terms by partitioning them into highly connected partitions.

One way of achieving this is to use a hypergraph partitioning algorithm that partitions the hypergraph into two parts recursively such that the weight of the hypergraphs that are cut by the partitioning is minimized (Han et al. 1998). Depending on the support threshold, documents that do not meet support (i.e., documents that do not share large enough subsets of terms with other documents) will be pruned. This feature is particularly useful for clustering large biomedical document sets.

2. *Text summarization*

Our algorithm is designed to produce summaries that emphasize “relevant novelty.” Relevant novelty is a metric for minimizing redundancy and maximizing both relevance and diversity. Complementary to work by Jing and McKeown (1999), whose emphasis is on summary fluency, our approach focuses on ensuring summary informativeness. Our system dynamically determines the foci of the documents through textual clustering, which in turn determines the specific information that will be extracted. The summary is formed by first extracting sentences from the clusters that contain the desired information, and later synthesizing them.

Our method extracts key phrases and sentences from the documents based on the mutual reinforcement principle. Similar ideas have been used to find the hub and authority web pages in link graphs in search engines (Kleinberg 1999). The heart

of our mutual reinforcement principle is that a *key phrase should have a high saliency score if it appears in many sentences with high saliency scores, whereas a sentence should have a high saliency score if it contains many key phrases with a high saliency score*. We explicitly model key phrases and sentences that contain them using undirected and weighted bipartite graphs and generate sentence extraction from textual documents on the fly without extensive training.

Bipartite graph of key phrases and sentences: For all documents in the same cluster, we generate two sets of objects: one is the set of key phrases $K = \{k_1, \dots, k_m\}$ from the cluster, and the other is the set of sentences $S = \{S_1, \dots, S_n\}$. We build a weighted bipartite graph from K to S in the following way: if the phrase k_i appears in sentence S_j , we then create an edge between k_i and S_j . We can also specify non-negative weights on the weighted bipartite graph with w_{ij} , indicating the weight on the edge (k_i, S_j) . For example, we can choose w_{ij} to be the number of times k_i appears in S_j . More sophisticated weighting schemes such as normalized frequency value for comparison study will be investigated in our experiments. We denote the weighted bipartite graph by $G(K, S, W)$, where $W = |w_{ij}|$ is the m -by- n weight matrix containing all pairwise edge weights.

Mutual reinforcement principle to calculate the salient scores of key phrases and sentences: In essence, the principle dictates that the saliency score of a phrase is determined by the saliency score of sentences it appears in, and the saliency score of a sentence is determined by the saliency scores of the phrases it contains. Mathematically, this statement is rendered as

$$Ss'(S_j) = \sum_{k_i \in N(S_j)} w_{ij} Sk(k_i), Sk'(k_i) = \sum_{s_j \in N(k_i)} w_{ij} Ss(S_j) \quad (3.6)$$

$$Ss(S_j) = Ss'(S_j) / \|Ss'\|, Sk(k_i) = Sk'(k_i) / \|Sk'\| \quad (3.7)$$

Final scores of the key phrases and sentences are obtained by iteratively solving the previous equations, where $N(k_i)$ is the neighbor of term k_i and $N(S_j)$ is the neighbor of S_j . The summations are over the neighbors of the vertices in question (i.e., when computing a term score), the summarization is over all sentences that contain the phrase and when computing a sentence score, the summation is over all phrases in the sentence. The corresponding component values of Sk and Ss give key phrases and sentence saliency scores, respectively. There are many numerical computation methods developed to calculate the scores of terms and sentences efficiently. See Hu (2004) for detailed discussions.

Sentence selection and synthesis: The sentence extraction part of our system is similar to the domain-independent multi-document summarization of Carbonell and Goldstein (1998) and Goldstein et al. (2000) in the way it clusters sentences across documents to help determine which sentences are central to the collection, as well as to reduce redundancy among sentences (as it does not make use of comparisons to the centroids of the multi-document set). We integrate the ideas from maximum marginal relevance (MMR) measure and cross-sentence information subsumption (CSIS) (Radev et al. 2000) to minimize redundancy and maximize both

relevance and diversity for extracted sentences. To achieve this goal, an important component is a good measure to evaluate the similarity of two sentences.

For two sentence $S_i = \{k_{i1}, k_{i2}, \dots, k_{ip}\}$ and $S_j = \{k_{j1}, k_{j2}, \dots, k_{jq}\}$ to measure the similarities between two sentences, every pair of terms in S_i and S_j are compared. If they are exactly the same, the similarity score is 1. If two terms are different but are related in the ontology, the similarity score is the semantic similarity in the ontology. There are many proposals to use the distance between two concepts in an ontology as the basis for their similarity (Resnik 1995). For example, assuming the commonality between terms k_{iu} and k_{jv} in the ontology is K_p , where K_p is the most specific class that subsumes both k_{iu} and k_{jv} , we define the semantic similarity $d(k_{iu}, k_{jv}) = 2 * \log P(K_p) / (\log P(k_{iu}) + \log P(k_{jv}))$, where $P(K_x)$ represents the probability that a randomly selected concept belongs to the K_x in the ontology. The similarity measure of S_i and S_j is defined as

$$Sim(S_i, S_j) = \frac{\sum_{u=1}^m \sum_{v=1}^n w_{iu,jv}}{m+n}, \text{ where } w_{iu,jv} = \left. \begin{array}{l} 1, \text{ if } k_{iu} = k_{jv} \\ d(k_{iu}, k_{jv}), \text{ if } k_{iu} \text{ is related to } k_{jv} \text{ in} \\ \text{the ontology} \\ 0, \text{ if } k_{iu} \text{ \& } k_{jv} \text{ are different literally} \\ \text{\& semantically} \end{array} \right\}$$

III. EXPERIMENT RESULTS

Biomedical literature data mining is essential for many bioinformatics problems and will ultimately enhance many related biomedical projects. Our experiments in biomedical literature data mining focus on, in particular, those that use large-scale genome-wide gene expression analysis as well as chromatin protein-to-protein interaction networks.

A. Extracting and mining the chromatin protein-to-protein interaction network

To test the scalability of SPIE, we conducted two experiments, as outlined in Tables 3.1 and 3.2. Table 3.1 is to stimulate the biologist to manually create a set of keyword filters to select the documents relevant to protein interactions, and then run the IE procedure on these documents. This manual approach is used by most users of MedLine. However, information retrieval in such databases becomes very time consuming because searchers who are likely to identify much relevant information also find many irrelevant documents at the same time. For example, a text query for "protein interaction" of the MedLine database retrieves 176,559 documents (as of December of 2004). In this study, we use 1,600 human chromatin protein names. When we used synonyms derived from LocusLink and nucleotide databases maintained by NCBI, the total number of protein names was about 7,000.

Table 3.1. Number of MedLine abstracts used in keyword-based searching.

Keywords	No. of Abstracts	No. of PPI	No. of Distinct PPI
Protein associate	8,025	2,526	760
Protein interact	33,835	8,457	2,158
Protein bind	69,981	12,034	2,664
Protein association	82,767	9,440	2,093
Protein binding	83,397	13,854	3,184
Protein interaction	145,857	19,344	3,795
Protein complex	185,157	24,938	4,300
Protein acetylate	172	434	116
Protein acetylation	5,027	5,622	827
Protein conjugate	18,770	225	92
Protein destabilize	879	100	31
Protein destabilization	2,233	231	62
Protein inhibit	124,178	7,690	1,602
Protein modulate	41,727	2,984	945
Protein modulation	71,159	2,843	913
Protein phosphorylate	3,991	1,186	315
Protein phosphorylation	90,475	15,106	2,249
Protein regulate	58,586	7,991	2,121
Protein regulation	289,940	32,669	5,915
Protein stabilization	27,349	1,630	340
Protein stabilize	5,714	775	221
Protein suppress	20,069	2,005	633
Protein target	74,714	10,735	2,433
Total	1,444,002	183,119	37,769
Total (elimination of redundant)	1,006,699	37,769	9,980

Table 3.2. Experimental results (SPIE).

No. of Abstracts	No. of PPI	No. of Distinct PPI
50 k	2,224	1,749
100 k	4,412	3,100
150 k	8,348	4,400
200 k	10,527	5,300
250 k	12,461	6,040
300 k	15,152	6,500
350 k	16,612	7,200
400 k	18,202	8,420
450 k	19,070	8,900
All	19,461	9,483

The result is shown in Table 3.2. In our second experiment, we started with 10 pairs of protein-to-protein interaction (PPI) pairs as seed instances. We then used SPIE to automatically construct queries and used the learned queries to retrieve documents from MedLine. We set the maximum document size to 10 k for each iteration, starting with 50,000 documents, and stopped at 500,000 documents when the new tuples added were very few. We repeated the experiments five times with different seed pairs and took the average number of documents. The results are summarized in Table 3.2.

Whereas a keyword-based approach examined 1.4 million abstracts from MedLine to extract 9,980 distinct chromatin protein-to-protein interactions, SPIE examined only 500,000 abstracts from MedLine to extract 9,483 distinct chromatin protein-to-protein interactions. It is very obvious that SPIE has a significant performance advantage over the keyword-based approach.

B. Text mining for enrichment of microarray data analysis

To explain the underlying biological mechanisms and to assign “biological meaning” to clusters of genes obtained by analytical methods, it is necessary to cross-reference genes with external information sources. Our method provides a much-needed framework domain experts can use to take full advantage of existing knowledge about transcription factors, regulatory elements, sequences, structural information, and assigned gene functions. It provides the ability to obtain an overview of the entire landscape of thousands of genes and their related literature. This is useful in the analysis of microarray data at the genomic scale, producing very insightful information such as which genes are functionally related to each other, what their shared functionality is, and which documents discuss this functionality.

We conduct some experimental study on yeast gene data sets (<http://rana.lbl.gov/EisenData.htm>), as outlined in Table 3.3. The reason we use the yeast DNA microarray is because the validity of our methods is best assessed by comparison of the results with existing summaries of biological information. The *Saccharomyces* Genome Databases (Cherry et al. 1998) and the Yeast Proteome Database (Costanzo et al. 2000), as well as the functional analysis given by Spellman et al. (1998), are critical for objective evaluation of our results. There are 6,221 genes in the data sets, but not every gene is associated with a particular functional family. The yeast gene has a lot of functional families. In our experiment, we considered the genes in one functional family as one cluster. The top-rated significant terms and best sentences from the articles related to each cluster are outlined in Table 3.4.

IV. CONCLUSIONS

In this chapter, we presented some NLP and text data mining techniques for extracting important biological relationships from huge amounts of clustering and

Table 3.3. Yeast gene functional families.

Gene Cluster ID	Functional Family	No. of Genes	No. of Documents	No. of Document Clusters for Each Gene Cluster
1	ATP synthesis	19	94	2
2	Mitosis	19	468	6
3	Vacuolar protein targeting	19	227	3
4	Silencing	20	425	5
5	Fatty acid metabolism	20	151	2
6	Meiosis	21	319	6
7	Phospholipid metabolism	21	209	3
8	TCA cycle	22	168	3
9	Chromatin structure	42	533	6
10	DNA replication	42	1,473	15

Table 3.4. Top-rated significant terms and the best sentence for each cluster.

1	Acid, alpha, <i>atp synthase</i> , beta subunit, mutant, mitochondriaorf, <i>oscp</i> . A fusion between the N-terminal 15 amino acid residues of beta-subunit and the mouse cytosolic protein dihydrofolate reductase (DHFR) was transcribed and translated <i>in vitro</i> and found to be transported into isolated yeast mitochondria.
2	Anaphase, <i>apc</i> , centromere, chromosome, kinetochore, <i>mitotic</i> , spindle, ubiquitin. Mutant cells also showed increased levels of mitotic chromosome loss, supersensitivity to the microtubule destabilizing drug MBC, and morphologically aberrant spindles. <i>mif2</i> mutant spindles, arrested development immediately before anaphase spindle elongation, and then frequently broke apart into two disconnected short half spindles with misoriented spindle pole bodies.
3	Clas, endosome, golgi, <i>ptdin</i> , syntax, transport, <i>vacuolar</i> , vacuole, vesicle. Protein transport in eukaryotic cells requires the selective docking and fusion of transport intermediates with the appropriate target membrane. t-SNARE molecules that are associated with distinct intracellular compartments may serve as receptors for transport vesicle docking and membrane fusion through interactions with specific v-SNARE molecules on vesicle membranes, providing the inherent specificity of these reactions.
...
10	Cell cycle, <i>DNA replication</i> , kinase, mitosi, mutant, phosphorylation, replication, topoisomerase. Our data link a potent inhibitor of Cdc2 kinase to a key protein required for the initiation of DNA replication and strongly suggest that inhibition of Cdc18 by cyclin-dependent kinases has an important role in ensuring that the genome is duplicated precisely once each cell cycle.

summarization biomedical literature for gene cluster functional interpretation from microarray data analysis. There are many other useful applications of NLP and text data mining for systems biology. These include recognition of biomedical terms (Fukuda 1998), improving homology search (Chang et al. 2001), discovering protein functional regions (Eskin and Agichtein 2004), identifying cellular locations

(Skounakis et al. 2003), and hypothesis generation (Swanson 1986; Hu 2005). The promise of NLP and biomedical literature data mining goes well beyond the discovery of biological relationships. When coupled with experimental validation, data mining of the literature provides a promising direction in assisting the system biologist in conducting original researches and in designing novel experiments and new treatments.

ACKNOWLEDGMENTS

This work was supported in part by research grants from NSF Career IIS 0448023, NSF CCF 0514679 and PA Dept. of Health Research Grant for the Center of Public Health Readiness and Communication. The author thanks Dr. Lechner for providing the chromatin protein name list, Illhoi Yoo for the partial implementation and experiments, and Il-Yeol Song and Min Song for suggestions and discussion related to information extraction.

REFERENCES

- Bader, G. D., Donaldson, I., Wolting, C., Quellet, B. F., Pawson, T., and Hogue, C. W. (2001). BIND: The biomolecular interaction network database. *Nucleic Acids Research* **29**(1):242–245.
- Bard, J. B., and Rhee, S. Y. (2004). Ontologies in biology: Design, applications, and future challenges. *Nature* **5**:213–222.
- Barzilay, R., Elhadad, N., and McKeown, K. R. (2001). Sentence ordering in multidocument summarization. In *Proceedings of Human Language Technology Conference (HLT-2001)*, pp. 149–156, San Diego, CA.
- Beil, F., Ester, M., and Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD) 2002*.
- Blaschke, C., Hoffmann, R., Oliveros, J. C., and Valencia, A. (2001). Extracting information automatically from the biological literature. *Comp. Funct. Genom.* **2**:310–313.
- Carbonell, J., and Goldstein, J. (1998). The use of MMR: Diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in IR*, pp. 335–336, Melbourne, Australia.
- Chang, J. T., Raychaudhuri, S., and Altman, R. B. (2001). Including biological literature improves homology search. In *Pacific Symposium on Biocomputing*, pp. 374–383.
- Cherry J. M. (1998). SGD: *Saccharomyces Genome Database*. *Nucleic Acids Res.* **26**:73–79.
- Costanzo, M. C. (2000). The Yeast Proteome Database (YPD) and *Caenorhabditis Elegans* Proteome Database (Wordpd): Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28**:73–76.
- Eskin, E., and Agichtein, E. (2004). Combining text mining and sequence analysis to discover protein functional regions. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 288–299.

- Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 707–718.
- Goldstein, J. Mittal, V., Carbonell J., and Callan, J. (2000). Creating and Evaluating Multi-Document Sentence Extract Summaries. In *Proceedings of the Ninth International Conference on Information Knowledge Management*, pp. 165–172, McLean, VA.
- Gusfield, D. (1997). *Algorithm on Strings, Trees, and Sequences: Computer Science and Computational Biology*. New York: Cambridge University Press.
- Hahn, U., and Mani, I. (2000). The challenges of automatic summarization. *IEEE Computer* 33(11):29–36.
- Hahn, U., and Reimer, U. (1999). Knowledge-based text summarization: Saliency and generalization operators for knowledge-based abstraction. In *Advances in Automatic Text Summarization*, (I. Mani and M. Maybury eds.). pp. 215–232, Cambridge, MA: MIT Press.
- Han, E., Karypis, G., and Kumar, V. (1998). Clustering in a high-dimensional space using hypergraph models. Technical Report, Department of Computer Science, University of Minnesota.
- Han, J., and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.
- Harris, Z. (1985). Distributional structure. In *The Philosophy of Linguistics*, (J. J. Katz ed.), pp. 26–47, New York: Oxford University Press.
- Hirschman, L., Park, J. C., Tsujil, J., Wong, L., and Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18(12):1553–1561.
- Hotho, A, Staab, A. S., and Stumme, G. (2003). Ontologies improve text document clustering. In *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 541–544.
- Hu, X. (2004). Integration of cluster ensemble and text summarization for gene expression analysis. In *Proceedings of the IEEE 2004 Symposium on Bioinformatics and Bioengineering*, pp. 251–259.
- Hu, X. (2005). Mining novel connections from large online digital library using biomedical ontologies. *Library Management Journal* (in press).
- Hu, X., Lin T. Y., Song, I-Y., Lin, X., Yoo, I., Lechner, M., and Song, M. (2004). Ontology-based scalable and portable information extraction system to extract biological knowledge from huge collections of biomedical web documents. In *Proceedings of the 2004 IEEE/ACM Web Intelligence Conference*, pp. 77–83.
- Iliopoulos, I., Enright, A. J., and Ouzounis, C. A. (2001). TextQuest: Document clustering of MEDLINE abstract for concept discovery. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 384–395.
- Jing, H., and McKeown, K. (1999). The decomposition of human-written summary sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 129–136.
- Kankar, P., Adak, S., Sarkar, A., Murari K., and Sharma, G. (2002). Medmesh summarizer: Text mining for gene clusters. In *Proceedings of the Second SIAM International Conference on Data Mining*.
- Karp, P., D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics* 16(3):269–285.
- Kleinberg J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms*.

- Li, M., Ma, B., Kisman, D., and Tromp, J. (2004). PatternHunter II: Highly sensitive and fast homology search. *Journal of Bioinformatics and Computational Biology* 2(3): 417–439.
- Marcu, D. (1997). From discourse structures to text summaries. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82–88.
- Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein: Protein interactions from the biological literature. *Bioinformatics* 17(2): 155–161.
- Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, pp. 21–29.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 448–453.
- Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 1044–1049.
- Rzhetsky A. (2004). Geneways: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics* 37(2004):43–53.
- Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science* 264:1421–1426. 37.
- Sarkar, I. N., Cantor, M. N., Gelman, R., Hartel, F., and Lussier, Y. A. (2003). Linking biomedical language information and knowledge resources: GO and UMLS. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 439–450.
- Schulze-Kremer, S. (1997). Adding semantics to genome databases: Towards an ontology for molecular biology. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pp. 272–275.
- Shatkey, H., Edwards, S., Wilbur, W. J., and Boguski, M. (2000). Genes, themes, and microarrays: Using information retrieval for large-scale gene analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems*, pp. 317–328.
- Skounakis, M., Craven, M., and Ray, S. (2003). Hierarchical hidden Markov models for information extraction. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 427–433.
- Spellman, P. T. (1998). Comprehensive identification of cell cycle-regulation genes of the yeast *Saccharomyces cerevisiae* by micorarray hybridization. *Molecular Biology of the Cell* 9:3273–3297.
- Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. KDD 2000 Workshop on Text Mining.
- Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30:7–18.
- Tanabe L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., and Weinstein, J. N. (1999). Med-Miner: An Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 27(6):1210–1217.
- Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communications between men and machine. *Communications of the Association for Computing Machinery* 9:36–45.
- Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. (2001). DIP: The database of interacting proteins: 2001 update. *Nucleic Acids Reg.* 29:239–241.
- Zamir, O., and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. *SIGIR*, pp. 46–54.

Integrated Imaging Informatics

**Bahram Parvin, Qing Yang, Gerald Fontenay,
and Mary Helen Barcellos-Hoff**

*Lawrence Berkeley National Laboratory, Berkeley,
California, USA*

Chapter 4

ABSTRACT

Organisms express their genomes in a cell-specific manner, resulting in a variety of cellular phenotypes or phenomes. Mapping cell phenomes under various experimental factors is necessary in order to understand the responses of organisms to stimuli or environmental conditions. Biological heterogeneity requires collection of large sets of data. These data sets require an integrated view of experimental and computational components, which is facilitated through the BioSig Imaging Bioinformatics framework. BioSig enables cataloging of protein localization and subcellular responses as a function of experimental factors (e.g., antigen, molecular inhibitor) for cell culture assays as well as for fixed tissue samples.

The underlying data model leverages emerging new standards in microscopy, assay development, and experimental design process. The presentation layer is web-based and utilizes a graphical interface to navigate through the annotation, data, and quantitative representation of data. Quantitative representations are derived through novel computational components that enable multiscale representation of images in terms of average behavior (per image) and subcellular responses (for each object in the image). These computational components are model based, leveraging geometric properties of the objects of interest as a means of delineating them from the background.

I. INTRODUCTION

Systems biology may be viewed either as a purely informatics-driven or database-driven mathematical modeling problem. In either case, the intent is to bring

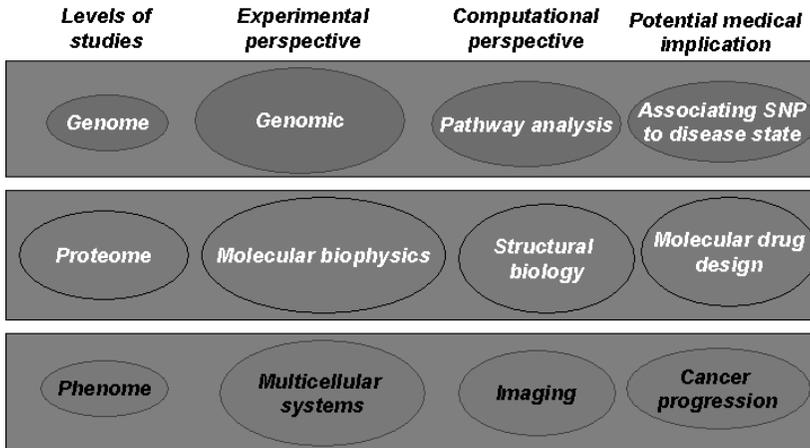


Figure 4.1. Levels of facets of systems biology in context.

together a computed representation of each facet of cell and molecular biology so as to gain new insights into the dynamics of cell and protein signaling. These facets vary in scale, experimental and computation perspective, and potential medical applications, as shown in Figure 4.1. For example, if the level of study is the phenome, the experimental perspective is multicellular systems, the computational perspective is image analysis and the corresponding informatics systems, and the potential medical application is cancer progression. Whether the end point of systems biology is informatics or multiscale mathematical modeling, quantitative representation of the experimental data facilitates correlation of molecular signatures and serves as input for estimating parameters of models.

Within this framework, imaging is an integral component of systems biology, revealing protein localization, tissue architecture, and cellular morphology under a variety of experimental factors and for different biological materials. Imaging enables insights into the dynamics of phenotype generation and maintenance, where a phenotype is the result of selective expression of the genome. It is an expression of the history of the cell and its response to the extracellular environment. To define cell phenomes, one would track the kinetics and quantities of multiple constituent proteins, cellular context, and morphological features in large populations. Such studies should also include responses to stimuli so that functional models can be generated and tested. Furthermore, signaling between cells and their extracellular microenvironment has a profound impact on cell phenotype (Roskelley et al. 1995).

These interactions are the fundamental prerequisites for control of cell cycle, DNA replication, transcription, metabolism, and signal transduction. The ultimate decision of a cell to proliferate, differentiate, or die is the response to integrated signals from the extracellular matrix, cell membrane, growth factors, and hormones.

From a systems perspective, each study endpoint can be addressed with a different model system. For example, an endpoint can first be tested on a monolayer cell culture system (ideal for high-throughput screening) and then on 3D cell culture assays (Schmeichel and Bissell 2003) that are closer to *in vivo* models, and finally on *in vivo* models. At present, the system is used to examine the impact of ionizing radiation on DNA damage and repair for 2D and 3D cell culture models and *in vivo* mouse models.

Whereas the 2D cell culture model reveals issues such as the intracellular kinetics of the repair mechanism, the 3D model enables studies of intercell communication and cell-ECM interactions that lead to better understanding of tissue architecture. For example, recent studies have shown that certain intracellular signaling pathways are linked via the cell adhesion system (Wang et al. 1998). Cell adhesion is how a cell attaches itself via integral membrane receptors to the extracellular matrix. Experimentally manipulating extracellular matrix receptors affects cell shape, alters the response of cells to new stimuli, and modifies multicellular organization as a function of time (Maniotis et al. 1997; Giancotti and Ruoslahti 1999).

A significant aspect of a phenotypic study is that changes in shape, response, and organization are heterogeneous and cell specific in tissue sections. Given the need for a large sample size (number of images) and complex hierarchical representation, it is necessary to maintain a detailed data model for managing data and information. The data model can then be used as a guided workflow for user-based annotation and browsing of the database. It can also be used to construct a visual interface for querying multiple targets, including positional references and morphological features.

The end results can then be visualized in terms of plots and a collage of images with sensitivity measures. Our research has three novel components: (1) development of a novel set of algorithms for capturing cellular morphology, protein expression, and cellular organization in tissue, (2) development of a data model that couples immunofluorescence with images, instrument configuration, and multilayered quantitative representation, and (3) development of a distributed imaging bioinformatics system that couples the data model with a web-based visual interface.

The organization of this chapter is as follows. Section II provides a brief overview of the system architecture and database interaction. Section III outlines various components of the informatics system. Section IV provides several classes of image analysis techniques used in understanding biological images. Section V outlines the details of specific phenotypic studies. Section VI concludes the chapter.

II. ARCHITECTURE

The architecture, shown in Figure 4.2, is a standard enterprise multilayer system consisting of a data layer, an object layer, a web service layer, and a presentation layer. The data layer is a PostgreSQL relational database, and the object layer

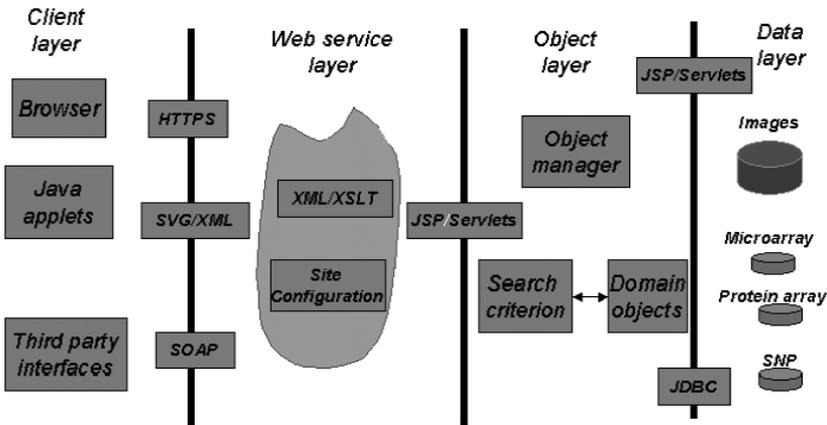


Figure 4.2. Imaging bioinformatics architecture is layered, and it uses a graphical interface for improved functionality.

provides conceptual mapping to this relational database. BioSig contains a flat-file mechanism for storing raw image data. However, compressed thumbnail representations of these data are maintained in the database itself. Retrieving a view of the database object requires sending the unique identifier of the database object via a URL request. This request is directed to a Java servlet that uses an object manager to retrieve a handle to the target object. Once this handle is retrieved, the servlet can interact with the database objects directly through JDBC.

The system supports five classes of operations in order to construct the object hierarchies and provide access to the database. These include creation and validation of content, transformation, communication, security, and storage. These operational classes, with the partial exceptions of security and storage, are implemented through a component-based architecture in which processing and communication tasks are generally divided into the smallest partitions of server resources, e.g., servlets. Servlets can coexist on a single computing platform or on disparate ones. The servlet platforms maintain computing resources such that they allow scaling for an increased load when communicating with distant web browsers for interoperable networked applications. The servlets are intentionally small to allow for extensibility. Several servlets allow for creation of database hierarchies through the Web. These servlets leverage modern markup techniques and provide validation against the schema that constrains both the structure of the data hierarchy and the content of each element.

III. INFORMATICS

To understand complexities associated with the informatics system, consider the following. A typical *in vivo* study includes a number of genetically similar mice at

different stages of their development: virgin, pregnant, lactate, and involution. In each category, mice are partitioned for treatment types (e.g., radiation and dosage) they will receive. Within each treatment population, mice are sacrificed at different time points. Tissues are then collected and sectioned, and coverslips are prepared for antibody treatment and subsequent imaging. The same experiment is then repeated using genetically altered mice for comparative analysis. It is clear that even such a simple study can generate a large quantity of annotated data that requires an underlying data model for subsequent query and analysis.

The data model has evolved from its previous version (Parvin et al. 2003) to a new design that leverages emerging standards in microscopy and experimental design. In this context, the data model is influenced by the Open Microscopy Environment (OME) and the MAGE model for managing microarray data. OME provides syntax for instrument configuration (complete description of the optical light path) and the type of analysis that has been performed on data. OME is an extensible data model providing a “semantic data type” structure with four levels of granularity: Global, Dataset, Image, and Features.

Within this framework, a subset of MAGE (experimental designs and specifics of the assay development) is embedded into the Global semantic type. The MAGE model provides a concise definition of the experimental factors (e.g., cell line, radiation, dosage, and other treatments) and protocol associated with assay development (e.g., plating, incubation time with a reagent at a specific concentration, number of washout, and fixation). The coarse representation of coupling between different entities is shown in Figure 4.3.

In this model, a “physical bioassay” has an “imageable target” (e.g., protein, mRNA, DNA), an “imageable probe” (e.g., labeled antibody, labeled synthetic oligonucleotide) and a set of “experimental factors” (e.g., radiation type and dosage), which corresponds to a collection of images. These are the high-level views of the data model. Specifics corresponding to the assay development are captured in the “treatment” and steps associated with it. Furthermore, the use of controlled vocabularies from the NCICB database, the MAGE model, and in-house-specific terminologies facilitates uniform annotation of database content.

Whereas modeling and structure of the database server side may evolve and become increasingly sophisticated, extension and maintenance of effective user interfaces remains difficult to manage. Functionality of these interfaces is especially important with regard to biological data entry, which is often time consuming, error prone, and repetitive. Toward improving interface functionality and usability, we have developed a framework, which facilitates the development of customized semantic views. From a functional perspective, one is not interested in the complete data model on the server side but in browsing a series of easily sortable high-level views, efficient data entry, and query-by-example for representing common types of queries.

Implementing these functions with the typical web browser scripting code is difficult and becomes especially error prone because of differences in browser implementation. To facilitate this functionality, the design incorporates Java applets from

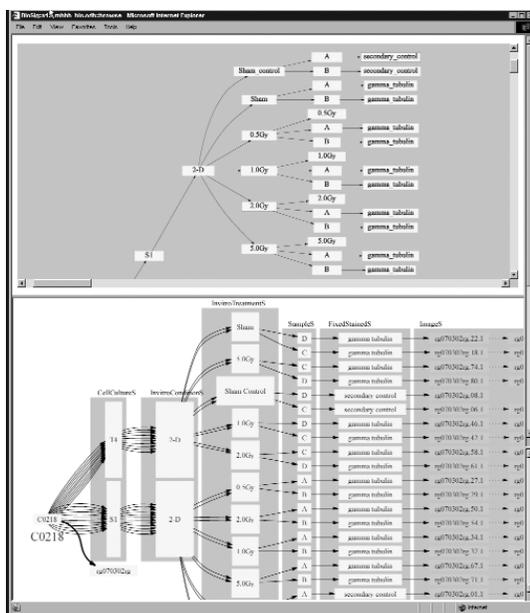


Figure 4.4. Guided workflow annotation and exploration of the database content.

Browsing of the database is enhanced through use of the Web and scalable vector graphics (SVG), which is a W3C standard for describing 2D graphics. The browser view of the data is represented as a directed graph, and its layout is enhanced through GraphViz, which is an open-source ATT software project. SVG is an extensible XML-based format for interactive presentation that incorporates images, text, shapes, and video and that allows for their precise layout and animation through declarative methods. SVG greatly facilitates rich presentation of data-driven graphics, and its rendering is accomplished through viewers that work as web browser plug-ins or as standalone applications. In addition, the presentation manager enables visualization of a query function as plots or as a collage of images. For example, the plot may be a dose-response or scatter diagram for computed features as a function of dependent variables. Examples of the use of presentation manager are shown in Figures 4.4 and 4.5.

IV. QUANTITATIVE ANALYSIS

Microscopy is often multichannel (multispectral), with one channel providing the required context for subsequent measurements. Segmentation of this context enables quantitative representation of protein localization as a function of micro-environment or genetic alterations. For example, if the level of study is the intra-

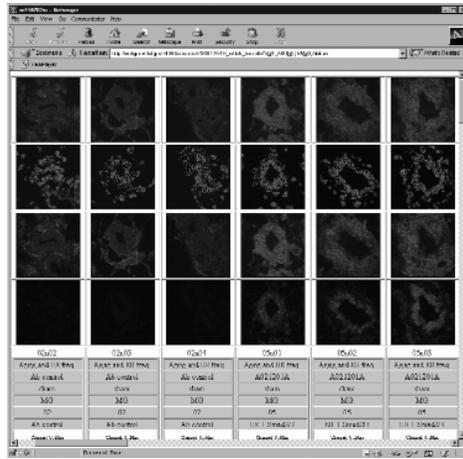


Figure 4.5. Query results for a collage of images and their annotations for protein co-localization studies. Composite images are automatically generated and scaled.

nuclear localization pattern, nuclear segmentation provides the context for subsequent analysis. Segmentation of the nuclear region is often hampered by the fact that staining may not be uniform, many substructures may be present, and nuclear regions often overlap to form perceptual boundaries. The latter problem is more severe in tissue sections than in cell culture assays.

Delineation of regions of interest is often model based, wherein the model can be either geometrical or statistical. Within the geometric framework, we offer two approaches based on variational and voting technique (Parvin et al. 2003, 2004; Yang and Parvin 2004). In each case, geometric constraints are specified through a Hamiltonian or a discrete model, and the solution is evolved into a fixed point. Within the statistical framework, a training set corresponding to features of interest is constructed to provide a probabilistic representation. The training set may be subsequently expressed with either declarative or generative basis functions that project the feature set into a higher dimensional space to improve the performance and robustness of the classifier. Present computational components, within the BioSig framework, are based on geometric principles.

One rationale for this design principle is that nuclei and cells of interest (epithelial cells) are radially symmetric or locally quadratic. This is a typical high-level constraint, sometimes based on human vision perception, imposed to derive the solution into a fixed point. Two examples of segmentation through geometric modeling follow. One is based on a variational technique, wherein each type of anomaly is modeled and ambiguities are resolved through constrained regularization. The second method is based on voting.

A. Variational approach to segmentation

In this approach, four computational steps are defined for removing a specific type of artifact (noise, touching nuclei) (Yang and Parvin 2003). The model assumes that a 3D nucleus imaged at a given focal depth is locally quadratic. Structures within the image are initially removed, and the image is interpolated to reveal a smooth surface. Each component of this binarized surface is then partitioned into several nuclei through a process called regularized centroid transform (RCT). These computational steps are shown in Figure 4.6a. The centroid transform essentially projects each point along the contour into a localized center of mass, as shown in Figure 4.6b. The solution is regularized to eliminate noise and other artifacts along the contour. This is shown in Figure 4.7. In the remainder of this section, each step of the process is described in detail.

Step 1. Elliptic regions: Let $I_0(x,y)$ be the original image. In the linear (Gaussian) scale space, its representation at scale σ is given by $I(x,y;\sigma) = G * I$, where G is a 2D Gaussian. The vector field of gradient $\nabla I = (I_x, I_y)^T$ can be classified by its Jacobian or by the Hessian matrix:

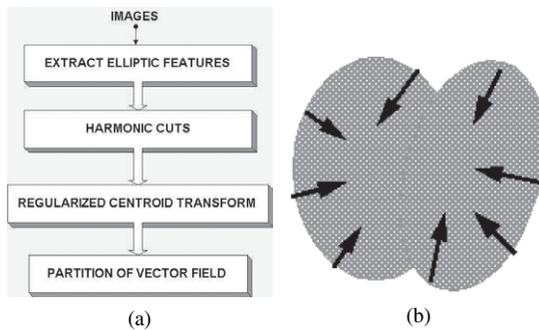


Figure 4.6. Segmentation process: (a) protocol for extracting delineating touching nuclei and (b) evolution of centroid transform between two adjacent nuclei.

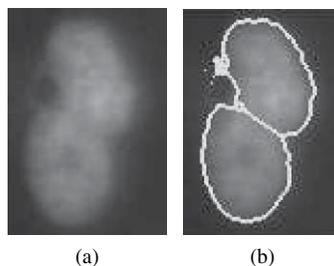


Figure 4.7. Segmentation of two touching nuclei.

$$H(x, y) = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix}$$

Bright elliptic regions can then be defined as the set of points satisfying the following conditions:

$$\begin{cases} I_{xx} < 0 \\ I_{yy} < 0 \\ I_{xx}I_{yy} - I_{xy}^2 > 0 \end{cases} \quad (4.1)$$

This means that both eigenvalues of the Hessian matrix are negative; that is, $H(x, y)$ is negative definite. Similarly, a dark elliptic region can be identified by the following conditions:

$$\begin{cases} I_{xx} < 0 \\ I_{yy} < 0 \\ I_{xx}I_{yy} - I_{xy}^2 > 0 \end{cases} \quad (4.2)$$

This classification is deduced directly from the classic method for flow pattern classification (Rao and Jain 1992). In scale-space theory (Lindeberg 1994), $I_{xx}I_{yy} - I_{xy}^2$ is referred to as the elliptic feature. Other properties of this feature are discussed in Section IV.A.3.

Step 2. Harmonic cuts: The next step of the computational process is to remove small elliptical regions from the cell and interpolate their region. This is essentially a noise removal step. However, our data set has both random noise (CCD noise) and speckle noise (internal structures within the cell). Previous efforts in noise removal have been limited to filtering random noise (Perona and Malik 1990). However, structural details behave much like speckle noise and more advanced techniques need to be developed.

To motivate our solution, let's first consider the 1D interpolation problem. A 1D function $l(x)$ with the region in the interval (a, b) can be interpolated with the average of the two endpoints, $\frac{l(a)+l(b)}{2}$. However, this approach breaks continuity of interpolation. A better approach is to weight the interpolation, at each point x , as a function of its distance to the boundary condition. That is, let $l^{new}(x) = (b-x)l(a)/(b-a) + (x-a)l(b)/(b-a)$. It can be shown that this representation is equivalent to minimizing

$$\frac{1}{2} \int_a^b l_x^2 dx \quad (4.3)$$

subject to the boundary conditions

$$\begin{cases} l(a) = l_a \\ l(b) = l_b \end{cases} \quad (4.4)$$

The 2D case is more complex because the boundary is often noisy and irregular, and it is not clear whether propagating intensity based on the distance transform will have desirable properties. We suggest that one way to ensure continuity is to regularize the solution by extending the 1D solution to 2D. That is, this is achieved by minimizing the following functional.

$$\frac{1}{2} \iint_D I_x^2 + I_y^2 dx dy \quad (4.5)$$

The Euler solution to this functional is the Laplace equation:

$$\nabla^2 I = I_{xx} + I_{yy} = 0 \quad (4.6)$$

Equation 4.6 is a 2D harmonic function defined on D , and thus we call this method a "harmonic cut." Harmonic functionals satisfy the Laplace equation and have many important properties (Alfors 1966). The Laplace equation is a special case of the Poisson equation, which has been studied extensively.

Step 3. Regularized centroid transform: At this stage of the computational process, each cell is represented with a smooth surface corresponding to each of its subcompartments. The next step of the process is to separate nuclei that are grouped together into a clump (i.e., touching one another). This is achieved using the RCT.

Figure 4.6b shows the basic idea for the RCT technique. The intent is to map vectors originating from the boundary of an ellipse to its centroid. If these vectors can be computed, the entire boundary can be grouped. This is true for both boundaries and their *interior points* (i.e., grouping utilizes not only the edges but the regional information). The main issue is that centroids are unknown and there are many centroids in the image. This is resolved by first computing a vector field that can then be used to partition touching objects. Let $I(x, y)$ be the original intensity image. At each point (x_0, y_0) , its equal-height contour is defined by

$$I(x, y) = I(x_0, y_0) \quad (4.7)$$

Expanding and truncating Equation 4.7 using Taylor's series, we have the estimation

$$I_x u + I_y v + \frac{1}{2} [I_{xx} u^2 + 2I_{xy} uv + I_{yy} v^2] = 0 \quad (4.8)$$

where $u = x - x_0$ and $v = y - y_0$, or, in the standard form,

$$\frac{1}{2} w^T H w + b^T w = 0 \quad (4.9)$$

where $H(x, y) = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{pmatrix}_{(x_0, y_0)}$ is the Hessian matrix, $H = \begin{pmatrix} I_x \\ I_y \end{pmatrix}_{(x_0, y_0)}$ is the gradient of intensity, and $w = (u, v)^T$ is the centroid in the local coordinate system. Recall that

the centroid of the quadratic curve defined by Equation 4.9 satisfies the following linear constraint.

$$Hw + b = 0 \quad (4.10)$$

If H is non-singular, the centroid can be determined directly, as follows.

$$w = -H^{-1}b \quad (4.11)$$

However, this is not always true, and in general the zero set defined by

$$\begin{vmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{vmatrix} = I_{xx}I_{yy} - I_{xy}^2 = 0 \quad (4.12)$$

is nontrivial and can be further classified into the following two categories.

- Uniform regions that correspond to a zero-intensity gradient of the image, with the result that there is no information to estimate the centroid
- Elliptic features that occur in nonuniform regions

The major limitation is that the centroids at singular points of the Hessian are not well defined. Because the basic formulation of centroid transform is ill posed (Tikhonov 1963), a regularized formulation is implemented. Let the centroid at (x, y) be denoted by $(u(x,y), v(x,y))^T$. The regularized model can then be expressed as

$$\min E(u, v) = \frac{1}{2} \iint \|H \cdot (u, v)^T + b\|^2 + \alpha (\|\nabla u\|^2 + \|\nabla v\|^2) dx dy \quad (4.13)$$

or

$$\min E(u, v) = \frac{1}{2} \iint (I_{xx}u + I_{xy}v + I_x)^2 + (I_{xy}u + I_{yy}v + I_y)^2 + \alpha (u_x^2 + u_y^2 + v_x^2 + v_y^2) dx dy \quad (4.14)$$

where the first and second terms are the error of estimation, the third term is the smoothness constraint, and $\alpha (>0)$ is the weight factor. The discrete Euler–Lagrange equations of the variational problem of Equation 4.14 can then be expressed as

$$\begin{cases} I_{xx}(I_{xx}u + I_{xy}v + I_x) + I_{xy}(I_{xy}u + I_{yy}v + I_y) - \alpha(u_{xx} + u_{yy}) = 0 \\ I_{xy}(I_{xx}u + I_{xy}v + I_x) + I_{yy}(I_{xy}u + I_{yy}v + I_y) - \alpha(v_{xx} + v_{yy}) = 0 \end{cases} \quad (4.15)$$

Step 4. Partitioning vector field: The final step of segmentation is to compute the partition of a vector field corresponding to the RCT. Consider an autonomous system of differential equations:

$$\begin{cases} \frac{dx}{dt} = u(x, y) \\ \frac{dy}{dt} = v(x, y) \end{cases} \quad (4.16)$$

The computed vector field can be partitioned simply by migrating each point to its local centroid, as shown in Figure 4.6b. There is a strong similarity between RCT

and watershed methods. However, because RCT is regularized and model based, it does not lead to excessive fragmentation and has a far better delineation profile. An example of segmentation results for two overlapping nuclei is shown in Figure 4.7.

Step 5. Representation and classification: Following segmentation, nuclear structure is represented with either an ellipse or hyperquadrics. The corresponding protein localization is then represented with a feature vector at each spectral channel. The ellipse fit is based on estimating the parameters of polynomial $F(a,x) = ax^2 + bxy + cy^2 + dx + ey + f$ subject to the constraint that $4ac - b^2 = 1$ (Fitzgibbon et al. 1996). A 2D hyperquadric (Hanson 1988; Kumar et al. 1995) is a closed curve defined by

$$\sum_{i=1}^N |A_i x + B_i y + C_i|^\gamma = 1 \quad (4.17)$$

Because $\gamma > 0$, Equation 4.17 implies that

$$|A_i x + B_i y + C_i| \leq 1 \quad \forall i = 1, 2, \dots, N \quad (4.18)$$

which corresponds to a pair of parallel line segments for each i . These line segments define a convex polytope (for large γ) within which the hyperquadric is constrained to lie. This representation is valid across a broad range of shapes that need not be symmetrical. The parameters A_i and B_i determine the slopes of the bounding lines and, along with C_i , the distance between them. γ determines the “squareness” of the shape.

The fitting problem is as follows. Assume that m data points $p_j = (x_j, y_j)$, $j = 1, 2, \dots, m$ from n segments ($m = \sum_{i=1}^n m_i$) are given. The cost function is defined as

$$\epsilon^2 = \sum_{j=1}^m \frac{1}{\|\nabla F_j(p_j)\|^2} (1 - F_j(p_j))^2 + \lambda \sum_{i=1}^N Q_i \quad (4.19)$$

where $F_j(p_j) = \sum_{i=1}^N |A_i x_j + B_i y_j + C_i|^\gamma$, ∇ is the gradient operator, λ is the regularization parameter, and Q_i is the constraint term (Kumar et al. 1995). The parameters A_i , B_i , C_i , and γ are calculated by minimizing using the Levenberg–Marquart non-linear optimization method (Press et al. 1992) from a suitable initial guess (Kumar et al. 1995). Classification of each cell in tissue is performed by representing cellular organization with an attributed graph, as shown in Figure 4.8. The nodes and edges in this graph correspond to cells and their relationships, respectively. The attributed graph provides the macro information about the micro anatomy where lumen can be localized and cell lines can be labeled with respect to their positions relative to lumen.

B. Voting-based techniques

Nuclear regions and certain classes of cell lines demonstrate radial symmetry. It is well known that radial symmetry—and in general, symmetry—is a pre-attentive

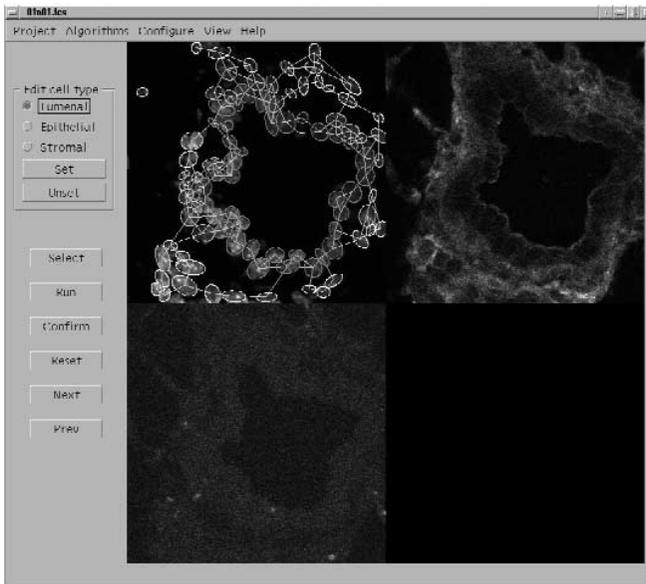


Figure 4.8. Segmentation is followed by the graph representation of the tissue where protein co-localization in specific cell lines can be registered in the spectral stack (see color plate 1).

human vision process (Attneave 1955) that improves recognition. Radial symmetry persists in biological images at multiple scales. At the lowest level, radial symmetry can be used to localize punctate protein expression. At a higher level, radial symmetry can detect and localize nuclear regions and assist in the quantitative analysis of proliferation assays. The critical features for detection of radial symmetries are noise immunity, invariance to deviation in shape and scale, and delineation of adjacent symmetries. However, the notion of radial symmetry is used in a weak sense, in that the basic geometry can deviate from convexity and strict symmetry for the purpose of approximating the center of mass.

Similarly, continuity and closure from incomplete boundaries and subjective surfaces are another form of saliency. The method introduced here allows inference of saliency through voting and perceptual grouping, and is implemented through the refinement of specifically tuned voting kernels (Yang and Parvin 2004). Spatial voting has been studied for at least four decades. Hough introduced the notion of parametric clustering in terms of well-defined geometry, which was later extended to the generalized Hough transform. In general, voting operates on the notion of continuity and proximity, which can occur at multiple scales (e.g., points, lines, lines of symmetry, or generalized cylinders). The novelty of our approach is in defining a series of kernels that vote iteratively along the radial or tangential directions.

Voting along the radial direction leads to localization of the center of mass, whereas voting along the tangential direction enforces continuity. At each iteration, the kernel orientation is refined until it converges to a single focal response. Several different variations of these kernels have been designed and tested. They are cone shaped, have a specific but variable scale and spread, and target geometric features of approximately known dimensions. In the case of radial symmetry, the voting kernels are initially applied along the gradient direction, then at each consecutive iteration and at each grid location voting orientation is aligned along the maximum spatial response. In the case of continuous boundary inference, the voting kernels are initially applied along the normal to the gradient. The shape of the kernel is also refined and focused as the iterative process continues. The method is applicable to perceptual shape features, has excellent noise immunity, is tolerant to variations in target shape scale, and is applicable to a large class of application domains.

Voting algorithm: Let $I(x, y)$ be the original image, where the domain points (x, y) are 2D image coordinates. Let $\alpha(x, y)$ be the voting direction at each image point, where $\alpha(x, y) := (\cos(\theta(x, y)), \sin(\theta(x, y)))$ for some angle $\theta(x, y)$ that varies with the image location. Let $\{r_{\min}, r_{\max}\}$ be the radial range and Δ be the angular range. Let $V(x, y; r_{\min}, r_{\max}, \Delta)$ be the vote image, dependent on the radial and angular ranges and having the same dimensions as the original image. Let $A(x, y; r_{\min}, r_{\max}, \Delta)$ be the local voting area, defined at each image point (x, y) and dependent on the radial and angular ranges, defined by

$$A(x, y; r_{\min}, r_{\max}, \Delta) := \{(x \pm r \cos \phi, y \pm r \sin \phi) | r_{\min} \leq r \leq r_{\max} \text{ and } \theta(x, y) - \Delta \leq \phi \leq \theta(x, y) + \Delta\} \quad (4.20)$$

Finally, let $K(x, y; \sigma, \alpha, A)$ be a 2D Gaussian kernel with variance, masked by the local voting area $A(x, y; r_{\min}, r_{\max}, \Delta)$ and oriented in the voting direction $\alpha(x, y)$. Figure 4.9 shows a subset of voting kernels that vary in topography, scale, and orientation. The iterative voting algorithm for detection of radial symmetry is outlined in the following.

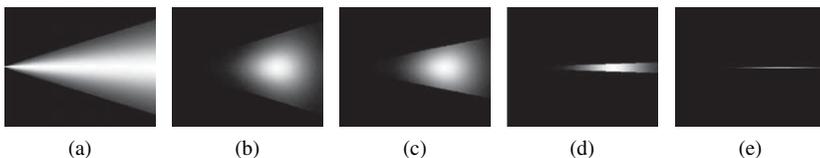


Figure 4.9. Kernel topography: (a–e) Evolving kernel for the detection of radial symmetries (shown at a fixed orientation) has a trapezoidal active area with Gaussian distribution along both axes.

Iterative Voting

Step 1. Initialize the parameters: Initialize r_{\min} , r_{\max} , Δ_{\max} , and a sequence $\Delta_{\max} = \Delta_N < \Delta_{N-1} < \dots < \Delta_0 = 0$. Set $n := N$, where N is the number of iterations, and let $\Delta_N = \Delta_{\max}$. Also fix a low gradient threshold, Γ_g and a kernel variance, depending on the expected scale of salient features.

Step 2. Initialize the saliency feature image: Define the feature image $F(x, y)$ to be the local external force at each pixel of the original image. The external force is often set to the gradient magnitude or maximum curvature, depending on the type of saliency grouping and the presence of local feature boundaries.

Step 3. Initialize the voting direction and magnitude: Compute the image gradient, $\nabla I(x, y)$, and its magnitude, $\|\nabla I(x, y)\|$. Define a pixel subset $S := \{(x, y) \mid \|\nabla I(x, y)\| > \Gamma_g\}$. For each grid point $(x, y) * S$, define the voting direction to be

$$\alpha(x, y) := -\frac{\nabla I(x, y)}{\|\nabla I(x, y)\|}$$

Step 4. Compute the votes: Reset the vote image $V(x, y; r_{\min}, r_{\max}, \Delta_n) = 0$ for all points (x, y) . For each pixel $(x, y) * S$, update the vote image as follows.

$$V(x, y; r_{\min}, r_{\max}, \Delta_n) := V(x, y; r_{\min}, r_{\max}, \Delta_n) + \sum_{(u,v) \in A(x,y,r_{\min},r_{\max},\Delta_n)} F\left(x - \frac{w}{2} + u, y - \frac{h}{2} + v\right) K(u, v; \sigma, \alpha, A),$$

Here, $w = \max(u)$ and $h = \max(v)$ are the maximum dimensions of the voting area.

Step 5. Update the voting direction: For each grid point $(x, y) * S$, revise the voting direction. Let

$$(u^*, v^*) = \arg \max_{(u,v) \in A(x,y,r_{\min},r_{\max},\Delta_n)} V(u, v; r_{\max}, \Delta_n)$$

Let $d_x = u^* - x$, $d_y = v^* - y$, and

$$\alpha(x, y) = \frac{(d_x, d_y)}{\sqrt{d_x^2 + d_y^2}}$$

Step 6. Refine the angular range: Let $n := n - 1$, and repeat steps 4 through 6 until $n = 0$.

Step 7. Determine the points of saliency: Define the centers of mass or completed boundaries by thresholding the vote image as follows.

$$C = \{(x, y) \mid V(x, y; r_{\min}, r_{\max}, \Delta_0) > \Gamma_v\}$$

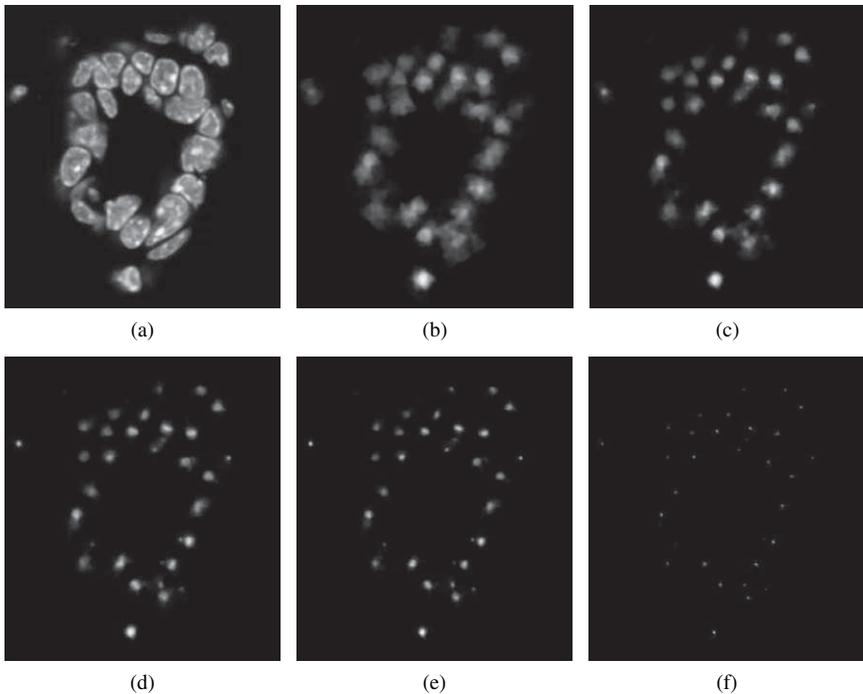


Figure 4.10. Evolution of the voting landscape for localization of nuclei in a mouse mammary tissue section indicate separation of touching nuclei: (a) original image, (b–e) refinement of the voting map, and (f) final localization of radial symmetries.

Experimental results: Two examples demonstrating the utility of radial voting for detection and localization studies are shown in Figures 4.10 and 4.11. The technique is tolerant to variations in scale, has excellent noise immunity, and can detect overlapping objects with incomplete boundaries.

V. APPLICATIONS

Two applications demonstrate the use cases of BioSig. The first corresponds to foci formation; that is, a potential representation of a DNA double-strand break (DSB) as a function of ionizing radiation. The second corresponds to cell culture studies involving cell-to-cell communication and adhesion within a 3D cell culture model.

A. 2D cell culture models

It is well known that ionizing radiation (IR) is a carcinogen in both humans and animals (Upton 1986). Using the standard monolayer cell culture provides a basic

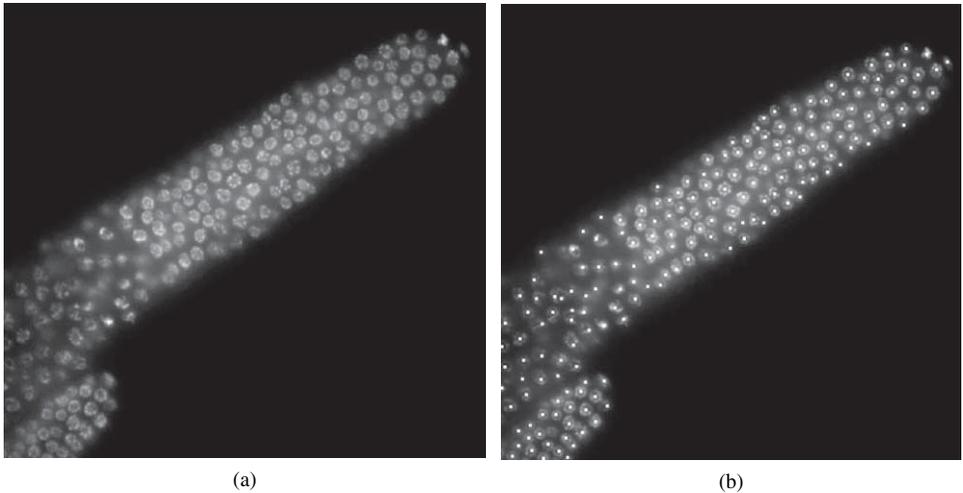


Figure 4.11. Proliferation assay: (a) original image corresponding to a sample of *C. elegans* observed through fluorescence microscopy and (b) detected nuclei.

understanding of how cells respond to a variety of factors that influence the degree, response type, and its kinetics. IR has the potential to induce DNA damage resulting in punctate protein co-localization (foci) within nuclei after induction of DSB. An example is shown in Figure 4.12, where a single cell has been extracted from the image for ease of demonstration. Punctate events corresponding to protein localization are extracted using the radial voting algorithm. Quantitative data are then imported into the database for subsequent analysis and visualization.

B. 3D cell culture models

To determine whether low-dose radiation promotes aberrant extracellular matrix (ECM) interactions, we have utilized BioSig to examine integrin and E-cadherin localization in preneoplastic human cells surviving radiation. Integrins are a family of epithelial receptors for the ECM, whereas E-cadherin maintains normal cell-to-cell interactions and architecture. We used the HMT-3522 (S1) human breast cell line cultured within a reconstituted ECM (Briand et al. 1987). These cells are genomically unstable but phenotypically normal in that they recapitulate normal mammary architecture in the form of a multicellular 3D acinus (Weaver et al. 1996). These clusters express integrins in a polarized fashion and develop an organized ECM over the course of 7 to 10 days in culture. The intent is to examine the consequences of exposing these cells to ionizing radiation and a protein modifier known as EGF, as shown in Figure 4.13.

Antibodies to E-cadherin, beta 1 integrin, or alpha 6 integrin were detected using a green fluorescent label, and nuclei were counterstained with a red fluorescent

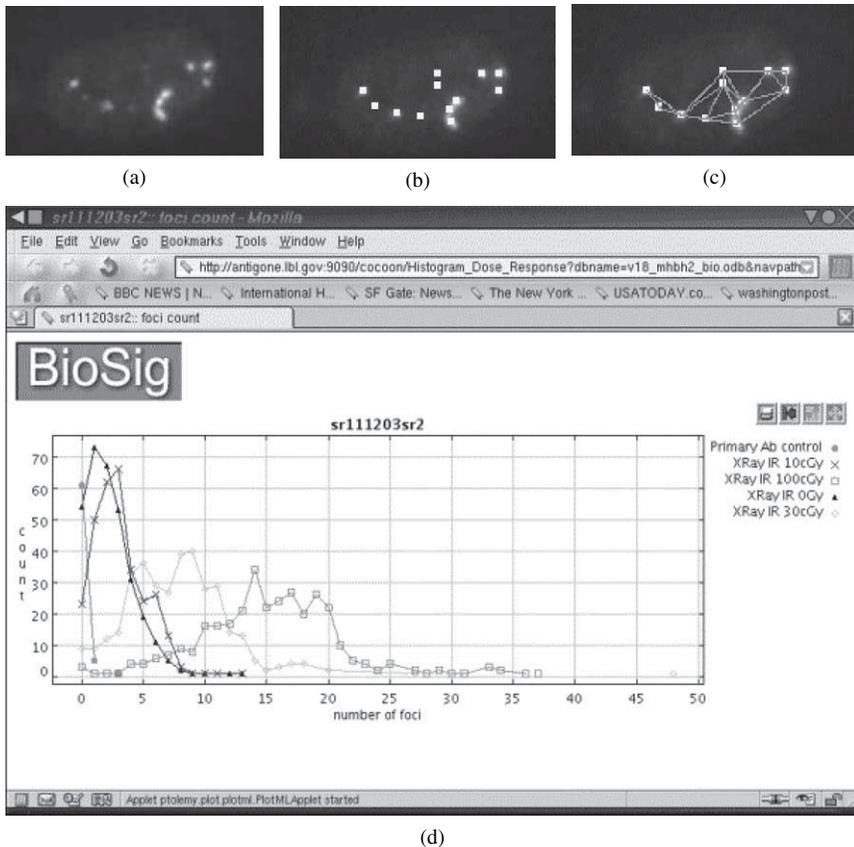


Figure 4.12. Detection and representation of punctate events: (a) original, (b) detection of foci, (c) graph-based representation of detected foci, and (d) informatics interface to the resulting population studies.

DNA dye. These were imaged using confocal fluorescence microscopy and were recorded using a 12-bit CCD camera. Cells that survived either 2 Gy or EGF-(400 pg/ml) showed decreased beta 1 or alpha 6 integrin localization, respectively. However, when cells were exposed to both radiation and EGF additional perturbations were noted. The clusters were disorganized, did not polarize the integrins at the cell surface, and failed to express E-cadherin, indicative of a lack of structural organization. An example of the untreated cells is shown in Figure 4.14a, which is stained for beta 1 integrin (green) with red nuclei.

Comparing this sample to Figure 4.14b, which is a colony of cells that were irradiated and treated with EGF-, shows that the localization of beta 1 integrin is perturbed, as is the organization of the colony. The previously cited characteristics, along with the organization of each colony, were computed and stored in the

Experimental Protocol

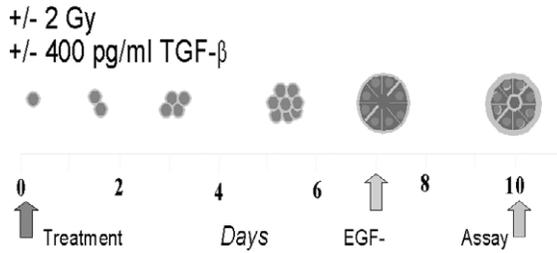


Figure 4.13. Experimental protocol for *in vitro* treatment of a colony.

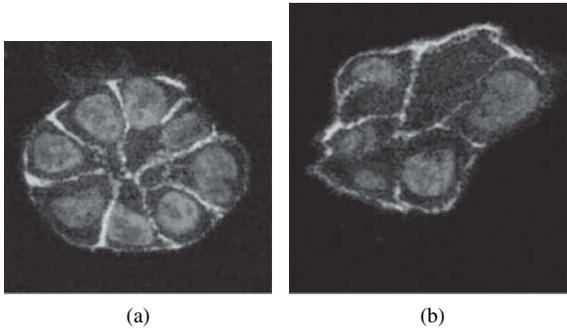


Figure 4.14. Organization of a colony as a result of radiation and EGF treatment: (a) an untreated sample maintained its symmetry along the lumen and (b) a treated sample lost its symmetric organization (see color plate 2).

database using the techniques described in Section IV. A pair of segmented images from untreated and treated samples, their segmentation, and organization are shown in Figure 4.15. These images correspond to a feature-based representation of the “organized” and “disorganized” state of the colony in the database.

VI. CONCLUSIONS

Imaging assays are only one endpoint of the computational and experimental perspective in systems biology. In this context, cellular responses and patterns of protein localizations are quantified as a function of stress, environmental conditions, therapeutic agents, or molecular inhibitors. The nature of data and its annotations has grown in complexity. This fact is coupled with the need to exchange information about protocol and experimental factors in a uniform way. From an informatics perspective, these needs are addressed through leveraging emerging standards in ontologies and controlled vocabulary.

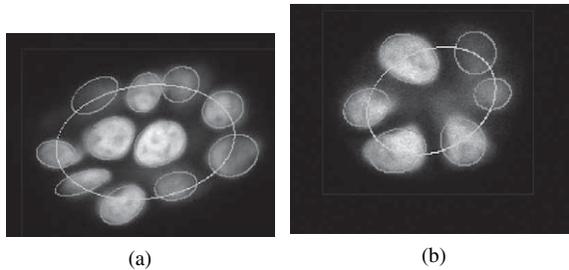


Figure 4.15. Organization of a colony as a result of low-dose radiation and EGF treatment indicates lack of symmetry around the lumen. Nuclei are segmented, represented with hyperquadrics, and a measure of symmetry by fitting an ellipse is measured: (a) an untreated sample maintains symmetry along the lumen and (b) a treated sample loses its symmetric organization (see color plate 3).

From the computational perspective, novel techniques have been developed to characterize cellular morphology and patterns of protein localization. This class of quantitative data needs to be coupled with other data types, such as expression profile, to reveal patterns of protein expression as a function of genetic signature. Such an ensemble of data will eventually contribute to a more precise representation of genetic pathways within each compartment of molecular machinery.

ACKNOWLEDGMENTS

Research was funded by the Low Dose Radiation Program of the Life Sciences Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098 with the University of California. The publication number is LBNL-57027. E-mail: parvin@media.lbl.gov. Web site is at <http://vision.lbl.gov>.

REFERENCES

- Alfors, L.V. (1966). *Complex Analysis*. New York: McGraw-Hill.
- Attneave, F. (1955). Symmetry information and memory for patterns. *American Journal of Psychology* **68**:209–222.
- Briand, P., Petersen, O., and Van Deurs, B. (1987). A new diploid nontumorigenic human breast epithelial cell line isolated and propagated in chemically defined medium. *In Vitro Cell Development Biology* **23**:181–188.
- Fitzgibbon, A., Pilu, M., and Fisher, R. (1996). Direct least square fitting of ellipses. *In Proceedings of the International Conference on Pattern Recognition*, IEEE Computer Society, Vol. 1, pp. 253–257.
- Giancotti, F. G., and Ruoslahti, E. (1999). Integrin signaling. *Science* **285**:1028–1032.
- Hanson, A. (1988). Hyperquadrics: Smoothly deformable shapes with convex polyhedral bounds. *Computer Vision, Graphics, and Image Processing* **44**:191–210.

- Kumar, S., Han, S., Goldgof, D., and Boeyer, K. (1995). On recovering hyperquadrics from range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(11):1079–1083.
- Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics* **21**(2):225–270.
- Maniotis, A. J., Chen, C. S., and Ingber, D. E. (1997). Demonstration of mechanical connections between integrins, cytoskeletal filaments, and nucleoplasm that stabilize nuclear structure. *Proceedings of the National Academy of Sciences of United States of America* **94**:849–854.
- Parvin, B., Yang, Q., and Barcellos-Hoff, M. H. (2004). Localization of saliency through iterative voting. *Proceedings of International Conference on Pattern Recognition*, **1**:63–66.
- Perona, P., and Malik, J. (1990). Scale space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**:629–640.
- Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. New York: Cambridge University Press.
- Rao, A. R., and Jain, R. C. Computerized flow field analysis: Oriented texture fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**:693–709.
- Roskelley, C. D., Srebrow, A., and Bissell, M. J. (1995). A hierarchy of ECM-mediated signaling regulates tissue-specific gene expression. *Current Opinion in Cell Biology* **7**(5):736–747.
- Schmeichel, K. L., and Bissell, M. J. (2003). Modeling of tissue-specific signaling and organ function in 3D. *Journal of Cell Science* **116**:2377–2388.
- Tikhonov, A. N. (1963). The regularization of ill-posed problems. *Dokl. Akad. Nauk. SSR* **153**(1):49–52.
- Upton, A. C. (1986). *Radiation Carcinogenesis*. New York: Elsevier.
- Wang, F., Weaver, V. M., Petersen, O. W., Larabell, C. A., Dedhar, S., Briand, P., Lupu, R., and Bissell, M. J. (1998). Reciprocal interactions between beta 1-integrin and epidermal growth factor receptor in three-dimensional basement membrane breast cultures: A different perspective in epithelial biology. *Proceedings of the National Academy of Sciences of United States of America* **95**(25):14821–14826.
- Weaver, V. M., Fischer, A. H., Petersen, O. W., and Bissell, M. J. (1996). The importance of the microenvironment in breast cancer progression: Recapitulation of mammary tumorigenesis using a unique human mammary epithelial cell model and a three-dimensional culture assay. *Biochemical Cell Biology* **74**(12):833–851.
- Yang, Q., and Parvin, B. (2003). Harmonic cut and regularized centroid transform for localization of subcellular structures. *IEEE Transactions on Biomedical Engineering* **50**(4):469–476.
- Yang, Q., and Parvin, B. (2004). Perceptual organization of radial symmetries. *Computer Vision and Pattern Recognition*, **1**:320–325.
- Yang, Q., Parvin, B., and Barcellos-Hoff, M.H. (2004). An imaging bioinformatics system for phenotypic studies. *IEEE Transactions on Systems, Man and Cybernetics* **33**(B):814–824.

Simpathica: A Computational Systems Biology Tool Within the Valis Bioinformatics Environment

Bud Mishra, Marco Antoniotti, Salvatore Paxia, and Nadia Ugel

NYU/Courant Bioinformatics Group, Courant Institute, New York University, New York, New York, USA

Chapter 5 Simpat

In memory of a friend, and mentor, Dr. Isidore "Izzy" Edelman, 1920–2004

ABSTRACT

Biology thrives on complexity, and yet our approaches to deciphering complex biological systems have been simple, observational, reductionist, and qualitative. The observational nature of biology may even seem self-evident, as expressed more than three centuries ago by Robert Hooke, whose work *Micrographia* of 1665 contained his microscopical investigations that included the first identification of biological cells: "The truth is, the science of Nature has already been too long made only a work of the brain and the fancy. It is now high time that it should return to the plainness and soundness of observations on material and obvious things."

As we begin to observe, infer, and list the fundamental "parts" out of which biology is created, we cannot stop marveling at how these same components and their variants and homologues interconnect, intertwine, and interact via universal principles that still remain to be fully deciphered. To unravel this biological complexity, of which we only have a hint so far, it has become necessary to develop novel tools and approaches that augment and rigorously formalize those human reasoning processes—tools that until now could be used for only tiny toy-like subsystems in biology.

To this end, the anticipated computational systems biology tools aim to draw upon constructive mathematical approaches developed in the context of dynamical systems, kinetic analysis, computational theory, and logic. The resulting toolkit aspires to build powerful simulation, analysis, and reasoning facilities that can be used by working biologists for multiple purposes: in making sense of existing data, in devising new experiments, and ultimately in

understanding functional properties of genomes, proteomes, cells, organs, and organisms. If this ambitious program is to ultimately succeed, there are certain *critical components* that require special attention of computer scientists and applied mathematicians. This chapter studies the nature of these components, software architecture for integrating them, and illustrative examples of how such an integrated system may function in practice.

I. INTRODUCTION

Computational systems biology faces many opportunities, obstacles, and challenges:

- There is a critical need for powerful computational environments, where novice users can build prototyping tools quickly. An example of such a tool is the multiscripting Valis environment, which provides rapid prototyping facilities in the same way Matlab and Mathematica do for other disciplines (Paxia et al. 2002).
- There is a critical need for research and pedagogic modeling tools that allow a novice user to understand—and reason and ponder about—large, complex, and detailed biochemical systems effectively, efficiently, and still effortlessly. Our effort in this direction is exemplified by the modular and hierarchical modeling, simulation, and reasoning tool called Simpathica, which can extract nontrivial temporal properties of diverse classes of biochemical networks, be they regulatory, metabolic, or signaling. Simpathica is constructed using the Valis environment (Mishra 2002b; Antoniotti et al. 2003a, 2003c; Mishra et al. 2003).
- There is a critical need for further and rapid development of new biotechnological approaches to provide measurements at single-molecule scales with high throughput and enhanced accuracy. We believe that significant improvements will emerge from the confluence of ideas from nanomechanical sensing devices, single-molecule biochemistries, better photochemistry, photonics and microscopy, and clever experiment and algorithmic designs, integrating these complex multicomponent devices (Anantharaman et al. 1997, 2005; Aston et al. 1999; Mishra 2002a, 2003).
- Finally, there is a critical need for a catalog of illustrating examples, where the aforementioned methodologies prove their power unambiguously. Given the infancy of this emerging field, these pioneering experiments will face many unpredictable hurdles, but the experience gained will most likely revolutionize our collective scientific viewpoint. Primary among these grand challenges could be the one related to various processes involved in cancer: cell cycle regulation, angiogenesis, DNA repair, apoptosis, cellular senescence, tissue space modeling enzymes, and so on. We note that presently there is no clear way to determine if the current body of biological facts—in this instance, those related to cancer—is sufficient to explain phenomenology. In these particular cases, rigorous mathematical models with automated tools for reasoning, simulation, and computation can be of enormous help to uncover cognitive flaws, qualitative simplification, or overly generalized assumptions.

This chapter is organized as follows. We first describe the structure of the computational systems biology toolkit (the Valis environment with related software and database system), in which Simpathica is embedded (Section II). This discussion is followed by a description of Simpathica software architecture and implementation within Valis (Section III) and an illustrative example (Wnt signaling in Section IV). We conclude (in Section V) with a list of grand challenges. Sections II and IV should be of interest to systems biologists interested in applying these tools to other examples. Section III should interest bioinformaticists engaged in building ever-more powerful computational tools for new rapidly arriving biological problems, protocols, and technologies. Section V should interest systems engineers, mathematicians, and computer scientists excited by the new challenges biology has created for many of our classical fields.

II. VALIS AND SIMPATHICA SYSTEMS

The toolkit combining the Valis software environment and the Simpathica systems biology reasoning tool is the product of over three years of research and development. Although these systems were designed for researchers in the life science community, the basic elements of their design are rather flexible and the tools can be adapted easily for other areas (e.g., medical informatics or computational finance). Currently, the NYU computational systems biology toolkit consists of the following three core components.

- *Valis*: An environment for rapidly integrating bioinformatics research performed by many different groups
- *NYU Microarray Database*: A database for collecting, sharing, distributing, and analyzing microarray abundance data
- *Simpathica*: An advanced systems biology reasoning tool for simulating and reasoning about biological processes

All of the tools are built with an open architecture, allowing modular enhancements to be developed easily and integrated rapidly. Because Valis allows rapid prototyping, and Simpathica can model biological domain knowledge, these tools allow scientists to quickly develop new hypotheses based on earlier experiments and available literature, and a platform to explore the steps needed to deepen their understanding.

A. Valis

The bioinformatics environment, Valis, includes tools for visualization of biological information, design, and simulation of *in silico* experiments and storage and communication of biological information. Valis sets itself apart from other environments through two key features.

- *Language-independent architecture*: The Valis advanced scripting engine can integrate research from multiple groups into a single environment. Researchers

using the Valis framework can share both the data and the algorithms for the analysis of that data. Valis's language-independent architecture allows research groups to leverage programs written in different languages. Valis currently supports scripting in R, Perl, Python, JavaScript, SETL, and Common Lisp, among others. This effectively allows Valis users to seamlessly integrate the major open-source computational biology platforms Bioconductor, BioPerl, and BioPython. Native libraries can be integrated in the system and used by all supported languages.

- *Whole genome analysis and systems biology analysis libraries:* Valis is versatile. Custom-built data structures and algorithms make it possible to perform whole genome analysis as well as simulation and reasoning of large biochemical networks on commodity hardware. As the throughput of sequencing efforts increases, Valis opens up new avenues for comparative genomics studies through computationally efficient large-scale whole genome analysis tools.

For instance, Valis has been used in conjunction with single-molecule physical mapping technology and microarray CGH technology to develop a set of comparative and functional genomic methods that can validate and find errors in genome sequence data, search for copy number variations in cancer cell lines, and create models of genome evolution to understand large segmental duplication and functional evolution of genes through duplication or splicing variants. The ability to create new algorithmic approaches rapidly within Valis is hoped to have an immediate and direct impact on the biological community: creating algorithms for understanding and extracting information from genomic and transcriptomic data in a coordinated manner; building, modifying, and correcting existing models to understand biological processes; and creating a common and unified language for biologists to communicate, exchange data, design, and disseminate experimental protocols.

B. NYUMAD

Currently, a significant portion of experimental biological measurement is focused on gene expression or genomic polymorphisms, and is obtained with microarrays. The wealth of microarray data being generated by biological researchers necessitates a system that can manage, analyze, persist, and distribute this information efficiently to other researchers. Such a system faces numerous challenges, including the sheer quantity and complexity of such data, lack of interoperability among systems, and the often proprietary methodologies used by the research laboratories generating the data.

Significant improvement has been accomplished through standardization. For instance, over the last couple of years MAGE-ML (MicroArray Gene Expression Markup Language) has emerged as the accepted standard for microarray data (www.mged.org), allowing for the transmission of XML documents describing this data. A Java object model (known as the MAGE-OM) derived directly from this specification also exists, thereby allowing MAGE-ML documents to be converted

into their corresponding runtime Java objects, and vice versa. This standard has grown widely in its adoption, and has made specification in one of its subsets (MIAME) required for most publication in archived journals. As the only current standard for microarray data, MAGE-ML continues to grow in popularity.

We have developed in our toolkit a system to maintain and analyze biological abundance data (for example, microarray expression levels or proteomic data), along with associated experimental conditions and protocols. The prototypic system is called the NYU Microarray Database (NYUMAD), which has been expanded to deal with many other related experiments. It uses a relational database management system for the storage of data and has a flexible database schema designed to store any type of abundance data along with general research data such as experimental conditions and protocols.

NYUMAD is a secure repository for both public and private data. Users can control the visibility of their data. Initially, the data might be private, but after the publication of the results the data can be made visible to the larger research community. Data analysis tools are supplemented with visualization tools. The goal is to not only provide a set of existing techniques but to incorporate ever more sophisticated and mathematically robust methods in the data analysis and to provide links and integration with other NYU tools such as the Valis system.

- Strict adherence to the MAGE-ML standard for microarray data to provide a foundation for interoperability with other data systems
- Modularization of software services to allow easy reuse and deployment of system subcomponents based on a specific laboratory's research needs
- Extensibility to allow developers to quickly create powerful data-editing GUI clients specific to their laboratory needs

The software system (under development) is a three-tier system whereby client applications used to edit/manipulate microarray data (GUI applications, analysis tools) exchange data with Java servlets via XML documents.

A different but related database, NYUSIM, is used to store *in silico* time-course data obtained through various methods of simulation. NYUSIM and NYUMAD share many features in common, and NYUSIM can be used interchangeably when the microarray data is obtained *in vivo* or *in vitro* by a series of experiments or sampling over time. The traces obtained from this database can be analyzed in many different ways, such as by time-frequency analysis with NYU BioWave or temporal logic analysis with Simpathica, and GOALIE (Go Algorithmic Logic for Information Extraction).

C. Simpathica

The Simpathica system occupies a central role in our systems biology toolkit. It allows biologists to construct and simulate models of metabolic, regulatory, and signaling networks and then to analyze their behavior. Biochemical pathways can be drawn on the screen through a visual programming environment or in a

specialized XML format (SBML format, see [SBML 2002]), a language originally designed to promote information exchange between multiple systems and programs. The system allows a biologist to combine simple building blocks representing well-known objects: biochemical reactions and modulations of their effects.

The system then simulates the pathways thus entered. Coupled with a natural language system, the Simpathica tool allows a user to ask questions, in plain English, about the temporal evolution of the pathways previously entered. In general, using modeling tools such as Simpathica to simulate biological processes *in silico* a biologist can model and study the behavior of complex systems—exploring many different scenarios rapidly without relying solely on experimentation.

D. Theoretical basis for Simpathica

As noted previously, Simpathica has a modular and hierarchical design that allows a user to effortlessly construct and rigorously analyze models of biochemical pathways composed of a set of basic reactions. Each reaction is thought of as a module and belongs to one of many types: reversible and irreversible reactions, synthesis, degradation, and reactions modulated by enzymes and co-enzymes or other reactions satisfying certain stoichiometric constraints. If the stochastic nature of these reactions is ignored (i.e., mass-action models), each of them can be described by a first-order algebraic differential equation whose coefficients and degrees are determined by a set of thermodynamic parameters.

As an example, a reaction modulated by an enzyme leads to the classical Michaelis–Menten’s formulation of reaction speed as essentially differential equations for the rate of change of the product of an enzymatic reaction. The parameters of such an equation are the constants K_m (Michaelis–Menten constant) and V_{max} (maximum velocity of a reaction). In a simple formulation, such as in S-system (Voit 1991, 2000), this approach provides a convenient way of describing a biochemical pathway as a composition of several primitive reaction modules (which can be automatically translated into a set of ODEs with additional algebraic constraints). Simpathica and XS-system (an extension of the basic S-System) (Mishra 2002b; Antoniotto et al. 2003a, 2003c; Mishra et al. 2003) retain this modular structure while allowing for a far richer set of modules and constraints.

The Simpathica architecture consists of two main modules and several ancillary modules. The main module is a graphical front end used to construct and simulate the networks of ODEs (ordinary differential equations) that are part of the model being analyzed. Simpathica uses, among others, the SBML format (SBML 2002) for exchange. The second module, XSSYS, is an analysis module based on a branching-time temporal logic that can be used to formulate questions about the behavior of a system, represented as a set of traces (time-course data) obtained from wet-lab experiments or computer simulations. The simplest forms of such queries are about the system steady-states, as there is very little interesting temporal structure to such queries.

These queries are of the form “Is it true that starting at a particular initial state the system can eventually get to a state and remain there without any variation in the states?” Other queries can be about the system robustness (system eventually returns to a state retaining certain properties under various forms of perturbation), reachability analysis (all states the system can eventually get to or all states from which the system can enter a state with some desirable or undesirable property), frequently visited states, and so on. The class of queries in such a branching-time temporal logic is rather rich, but yet amenable to efficient computational manipulation. Thus, starting with a state-trace of a biochemical pathway (i.e., a time-indexed sequence of state vectors representing a numerical simulation of the pathway) as input, Simpathica performs the following operations.

- Simpathica answers complex questions involving several variables about the behavior of the system. This is rather different from visually examining intertwined sets of simulation traces of a large complex system.
- Simpathica stores traces in an ancillary database module, NYUSIM, and allows easy search and manipulation of traces in this format. The analysis tools allow these traces to be further examined to extract interesting properties of the biochemical pathway.
- Simpathica classifies several traces (either from a single experiment or from different ones) according to features discernible in their time and frequency domains. Multiresolution time-frequency techniques can be used to group several traces according to their features: steps, decreases, increases, and even more complex features such as memory.
- Simpathica can automatically generate interesting properties that distinguish one model from a variant in the same family. For instance, by examining cell-cycle models of wild types, mutants, and double-mutants Simpathica can generate a story about how they subtly differ in their temporal behaviors.

With these tools, Simpathica provides an environment to suggest plausible hypotheses and then refute or validate these hypotheses with experimental analysis of time-course evolution. It also allows investigating conditions or perturbations under which a biochemical pathway may modify its behavior to produce a desired effect (an instance of a control engineering problem).

The XSSYS, a Simpathica back end, implements a specialized model checking (Browne et al. 1986; Clarke et al. 1999) algorithm that given a “model trace” and a temporal logic formula expressed in an extended CTL form can state whether the formula is true or false, while providing a counterexample in the latter case (i.e., the system gives an indication at which point in time the formula becomes false).

A full description of the syntax and semantics of the temporal logic language manipulated by Simpathica/XSSYS is beyond the scope of this chapter and is hence omitted. For the purpose of the present discussion, it suffices to assume that all standard CTL operators are available (e.g., modal operators such as *always*, *even-*

tually, globally, in future, until and the standard Boolean operations such as *and, or, implies, and not*). For instance, robustness of a “purine metabolism pathway model” is succinctly expressed by a statement such as “Always (PRPP > 50 * PRPP1 implies (steady_state()) and Eventually (IMP > IMP1) and Eventually (HX < HX1) and Eventually(Always(IMP = IMP1)) and Eventually(Always(HX = HX1))”. This statement captures a very complex notion of biological robustness: An (instantaneous) increase in the level of PRPP will not make the system stray from the predicted steady state, even if temporary variations of IMP and HX are allowed.

Thus, the main operators in XSSYS (and CTL) are used to denote possibility and necessity of propositions over time. In our case, such propositions involve statements about the value of the variables representing concentrations of molecular species. For instance, to express the query asking whether a certain protein level (ρ) will eventually grow above a certain threshold value (K), we write “eventually ($\rho > K$).” We also augment the standard CTL language with a set of domain-dependent queries. Such queries may be implemented in a more efficient way and express typical questions asked by biologists in their daily data analysis tasks.

As an example, we can formulate complex queries such as “Always (Globally (X in [L, H]) and eventually (X = L) and eventually (X = H) and globally (X = L implies next (X in [L, H] until X = H)) and globally (X = H implies next (X in [L, H] until X = L)))”. The query expresses the fact that the value of the X variable “oscillates” between the two values of L and H. Note that our temporal logic deals with time in a topological sense and hence lacks the expressive power to assert that the time period between L and H is constant.

On the other hand, this same topological nature of time helps us express natural ordering among important biological events, independent of whether the events are controlled by processes operating in fast or slow time scales. Thus, in spite of few obvious shortcomings CTL is still powerful enough to describe many properties of the system, such as liveness and safety. Furthermore, for those temporal properties expressible in the logic the analysis tool efficiently constructs counterexamples when input query fails to hold true or restricts the conditions under which the query can be satisfied. A more thorough introduction to XSSYS and its capabilities can be found in Antoniotti et al. (2002, 2003c) and Mishra (2002b).

III. SIMPATHICA WITHIN VALIS

In this section we examine how the possibility of using multiple scripting languages within Valis has proven very useful in rapid construction of tools for bioinformatics and computational biology. To this end, we consider here the Simpathica system described earlier and developed as part of the DARPA BioCOMP project.

The Simpathica/XSSYS system is logically divided into a front end and a simulation system (i.e., Simpathica proper and its analysis back end XSSYS). The two components work together to construct, simulate, and analyze the behavior of

metabolic and regulatory networks. The biochemical pathways are entered into the system either via the main Simpathica user interface or in an XML format. The system then simulates the pathways entered and produces trace objects. The XSSYS back end, written in Common Lisp, manipulates these traces (or traces produced by other simulation software or experiments) and evaluates queries about the temporal evolution of the pathways in an appropriate temporal logic language. In summary, the following are the key steps.

1. The Simpathica front end takes as input descriptions of metabolic and regulatory pathways constructed from a set of standard building blocks, which describe a repertoire of biochemical reactions, and can display these pathways in a graphical representation.
2. Simpathica then transforms this graph into an internal XML representation that can also be used for data exchange purposes. This internal representation consists of a set of ODEs along with initial conditions. These ODEs are then translated into Python code, which performs the actual simulation by integrating the set of equations. The result of such a simulation is the trace object to be input into the XSSYS trace analysis system.
3. The output of the Simpathica front end consists of an XML model and a trace object produced indirectly by the chosen ODE integrator (Python in this specific case).
4. Once these are available, the XSSYS system takes the trace object and a temporal logic query and evaluates the truth value of the query using a model-checking algorithm. If the query turns out to be false over the trace, XSSYS will also return a counterexample (in the form of a time index indicating a point where the trace falsifies the query).

The modules produced for the BIOCAMP project initially used the OAA Object Agent Architecture to facilitate integration between modules written in different languages and produced by different groups. However, we found that the OAA architecture initially selected to speed up prototyping of the BioCOMP system—Bio-SPICE—has a few shortcomings which we wanted to circumvent.

- In this architecture, each agent must register with a “facilitator” (written in Prolog), which centralizes most exchanges.
- The facilitator serves to solve queries written in an interagent communication language (ICL) that must be built by the clients. The ICL uses most of the power of the unification-based semantics of Prolog. However, this approach requires agent writers to actually know and write in Prolog, which is further compounded by the problem that requests in ICL must be laboriously constructed using an abstract syntax tree library in Java and/or C.
- Performance issues arise for in-process calls. Limits may be imposed on message sizes.

Valis sidesteps these problems by integrating several subsystems in a much tighter way. Once having assembled all of the underlying building blocks needed (e.g., the

XML parsers, graph viewers, ODE integrators, and XSSYS subsystem), it is possible to prototype in Valis a system such as Simpathica/XSSYS in a matter of a couple of weeks.

A basic graphical user interface can be put together in a Valis form in a few hours, in that most of the widgets needed are standard controls of the form manager. The interface can be organized using multiple “Tab” container widgets and using different tabs for I/O, the model editing widgets, the simulation pane, the graphical results of the simulation, and the interface with the XSSYS subsystem. Figure 5.1 shows the tabs and the “model editing” pane. The code that handles events from the forms and customizes the interface can readily be written in JavaScript.

The only graphical element needed that is a bit unusual is a viewer for showing a graphical representation of the pathways. For this widget, we use the Adobe SVG viewer. This is a freely available control that can render models written in the SVG language with zooming capabilities. Because most of the internal data structures with which Simpathica/XSSYS works are based on XML, it is appropriate to use the versatile XML parser from Microsoft to handle them. In Valis this can be made available using just one code line:

```
xmlparser=CreateObject("Msxml2.DOMDocument.4.0");
```

A model of a pathway can be easily stored into XML files and retrieved using functionalities provided by the XML parser object. Once loaded and parsed, this model is used to update the internal data structures (namely, the “compounds” and “reactions” lists) and the corresponding graphical widgets.

We construct a graphical representation of the model from the internal XML representation and feed it to the SVG widget. We use the DOT language (a general graph description language) as an intermediate language for this graphical representation. The DOT code is produced by applying a style sheet to the XML model. For example, a subset of the Wnt Signaling Model (discussed in detail later in the chapter) will yield the following DOT code.

```
digraph G {
  X0 (label="W", style=filled);
  X1 (label="Dshi");
  X2 (label="Dsha");
  X1 ->"Yv1" (label="v1", arrowhead=none);
  X0 ->"Yv1" (style=dotted);
  "Yv1" ->X2;
  "Yv1" (shape=point);
  X2 ->X1 (label="v2");
}
```

In this representation, *X0* through *X2* and *Yv1* are nodes (each with certain properties, such as label, style, and so on). The DOT code shows a reversible reaction between *Dshi* and *Dsha* modulated by *Wnt*.

The Graphviz system can produce a variety of other graphical representations (among them SVG) once provided with models described in the DOT language.

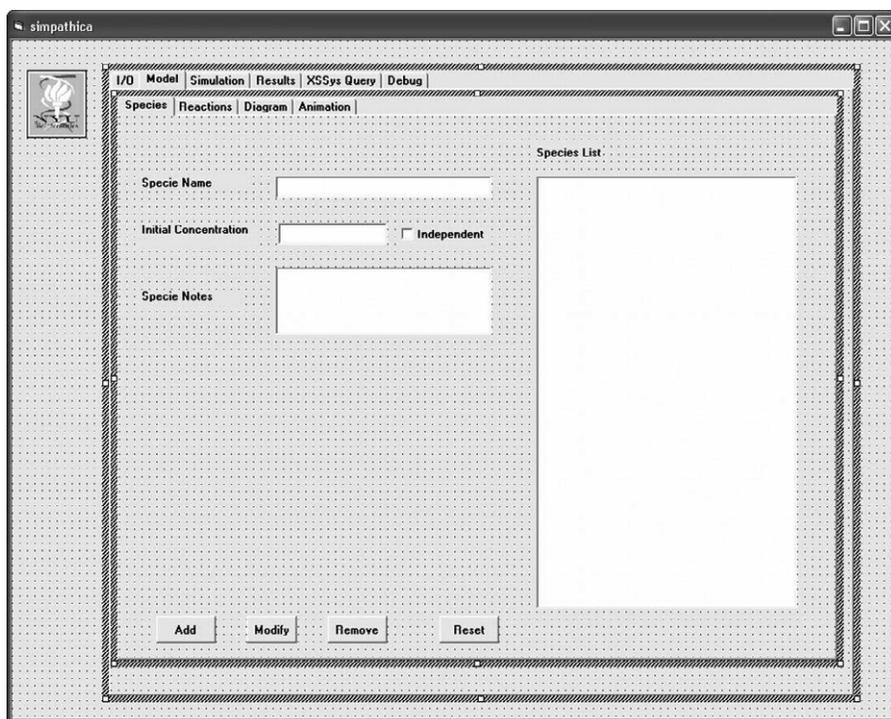


Figure 5.1. Simpathica GUI design.

We reworked this system into a standalone control, which is then made available to Valis.

```
// this function reloads the SVG from the dot string
// dotStr is the DOT description of the model
function updateSVG(dotStr) {
  var f, svgStr;
  // use the graphviz control to obtain SVG code
  svgStr=graphviz.DotToSvg(dotStr);
  // save the svg string to file for efficiency purposes
  f=fso.CreateTextFile(pathname+"\\diagram.svg", true,
    false);
  f.write(svgStr);
  f.close();
  // visualize the svg diagram
  activeSvgCtl.SRC=pathname + "\\diagram.svg";
}
```

This program fragment yields a graph that summarizes the reaction pictorially, as shown in Figure 5.2. Furthermore, the system allows the user to navigate through

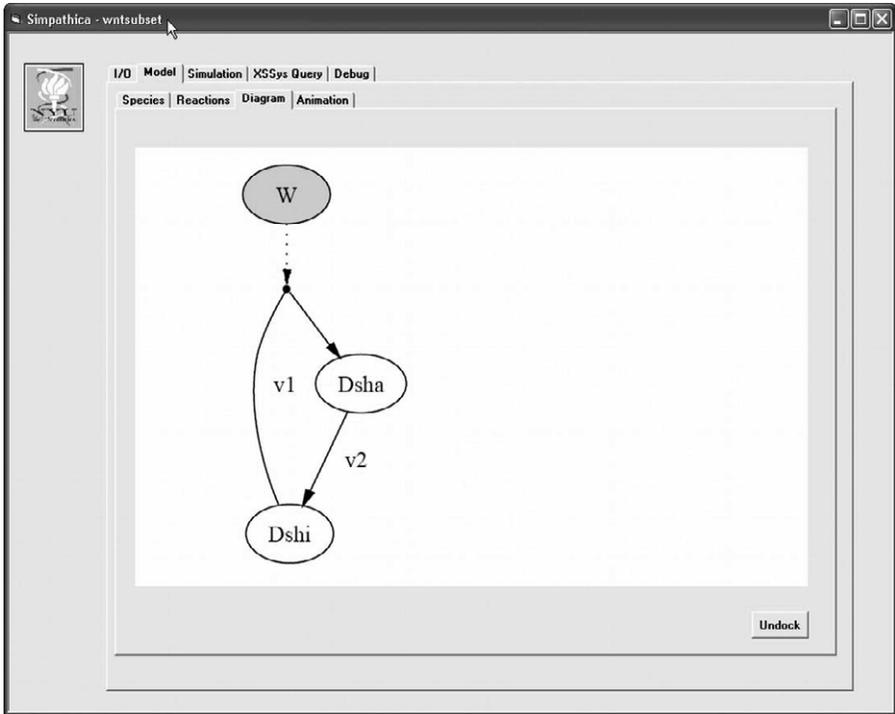


Figure 5.2. The SVG viewer embedded in a Valis form.

this graph using the SVG viewer. Note that the internal model used to produce the graph representation can be transformed into an intermediate representation suitable for the generation of a set of ODEs. This intermediate representation is obtained with the application of another XML style sheet, as follows.

```
function generateScript4Map() {
  var xmlmap=null;
  //generate the xml map from the gui
  xmlmap=downloadMap();
  //transform the map (xmlmap) to the graph internal //rep
  representation (xmlgraph) using the style sheet
  (xslmap2graph)
  xmlmap.transformNodeToObject(xslmap2graph, xmlgraph);
  writeDebugInfo("Graph", xmlgraph.xml);
  //generate the python script for the ODE
  return xml2py(xmlgraph);
}
```

Without much difficulty, we can then dynamically produce some Python code (in the `xml2py` function shown previously) with the step function for the integrator.

```
class __simplathica:
    def WntPathway_subset(self, X, t):
        xdot=()
        xdot.append(0)
        xdot.append(+1*(+0.182*pow(X(1),1)*pow(X(0),1))
                    ++1*(+1.82e-2*pow(X(2),1)))
        xdot.append(++1*(+0.182*pow(X(1),1)*pow(X(0),1))
                    +- 1*(+1.82e-2*pow(X(2),1)))
        return xdot

initial = (1,100,0)
compoundsNames = ("W", "Dshi", "Dsha")
functionName="__simplathica().WntPathway_subset"
```

A Python ODE integrator (based on Numeric Python) will integrate the ODEs generated.

```
from Numeric import *
from scipy import *
from scipy.integrate import *
from scipy import gplt
def executeSimulation(script, fT, tT, st):
    exec script
    global fromTime, toTime, steps, precision, time, Y
    fromTime = fT
    toTime = tT
    steps = st
    precision = (toTime - fromTime) / float(steps)
    time = arange (fromTime, toTime, precision)
    Y = odeint(eval(functionName), initial, time)
    gplt.plot(time, Y)
```

This Python function is called directly from the `Simplathica` event handlers (written in JavaScript) once the simulation is started.

```
// Call the Python integrator. Pass the equations and the
// simulation parameters
executeSimulation(generateScript4Map(), from, to, steps);
```

The `executeSimulation` Python function provides also for a default visualization of the traces of the simulation. It is very easy to customize the current plotting program used by the visualizer, or even to choose another plotting control (e.g., Microsoft's *Chart* control). (See Figure 5.3.)

The `XSSYS` query event (generated by the `Run XSSys` button in the `XSSys Query` pane, shown in Figure 5.4) can be handled by some JavaScript.

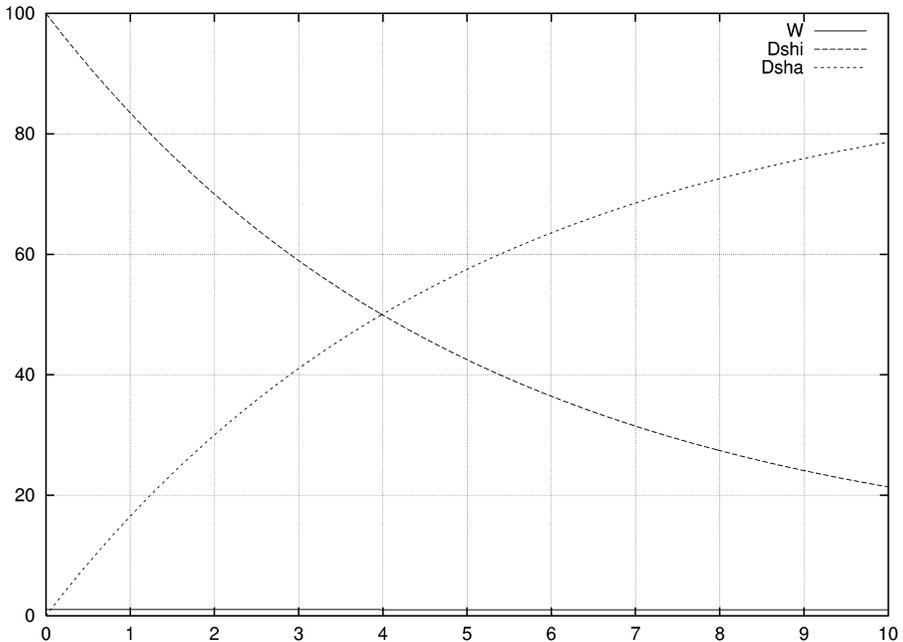


Figure 5.3. Simulation of the Wnt subset.

```
function Form1_LoadTraceCommandButton::Click() {
    i = Load_Trace(filename);
    Select_Trace(filename);
    Form1_LoadedTracesListBox.AddItem(filename, i);
}
function Form1_RunXSSysButton::Click() {
    Form1_TLResultTextArea.text = "";
    Form1_TLResultTextArea.text=
        Analyze_This(Form1_TLQueryTextArea.text);
}
```

The JavaScript Query-Handler, in turn, calls (the front end to) the XSSYS system in Common Lisp. The XSSys query pane is shown in Figure 5.4. This pane indicates how the user may enter the queries and get a response. All of this is integrated in the code in Common Lisp as follows. The Common Lisp code is a simple wrapper around the XSSYS package. This wrapper implements the core of the Temporal Logic analysis facility (with the identifiers prefixed by `xssys`). The Common Lisp integration within Valis and the ActiveX Scripting Engine is as tightly coupled as VisualBasic, and much more so than that in Perl or Python.

A function defined within Common Lisp appears directly within the ActiveX Scripting Engine name spaces, and any function or procedure defined (for example)

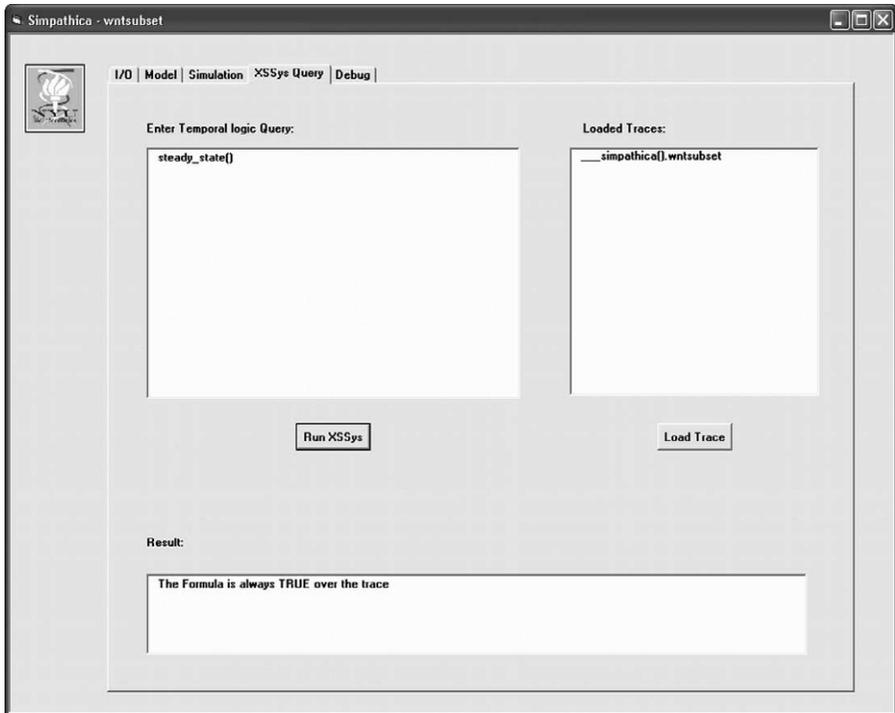


Figure 5.4. The XSSys query pane.

in Perl or JavaScript appears as a regular function in a Common Lisp “script.” Of course, Common Lisp is compiled natively, thus enhancing the performance over other “scripting languages.” The functions `|Load_Trace|` and `|Analyze_This|` in the code following thus become visible in the ActiveX Scripting Engine name spaces and can be referenced by (for example) a VisualBasic user interface. No special registration code is necessary.

```
(defun |Load_Trace| (filename)
  (unless (probe-file filename)
    (return-from |Load_Trace| -1))
  (setf xssys:*the-current-trace*
        (xssys:load-trace (pathname filename) :btd))
  (or (position (xssys:trace-system-name xssys:*the-current-trace*)
              (xssys:list-all-traces)
              :test `string=
              :key `xssys:trace-system-name)
      -1))
```

```
(defun |Analyze_This| (query)
  (multiple-value-bind (result
                      satisfying-state-groups
                      counter-example)
    (xssys:analyze-this trace-data form)
    (when counter-example
      (setf counter-example-index (second counter-example))))
  ...
  several variables in this example are introduced
  ;; elsewhere.
  (format *standard-output*
    "~&::: Query ~S prop ~S prop-ag ~S result ~S counter
     ~S~2%"
    query
    propositionalp
    propositional-always-p
    result
    counter-example-index)
  ...
)
```

IV. Wnt SIGNALING EXAMPLE

There has been considerable interest in signaling pathways involving Wnt proteins, which form a family of highly conserved secreted signaling molecules. These proteins regulate cell-to-cell interactions during embryogenesis. Furthermore, Wnt genes and Wnt signaling are also implicated in cancer. (See Figure 5.5.)

While at a qualitative level, scientists now have significant insights into the mechanisms of Wnt action, and data from better experiments through genetics in *Drosophila* and *Caenorhabditis elegans* (and gene expression in *Xenopus* embryos) we still only have a rudimentary understanding of how the complete pathway operates under various situations.

In a widely accepted model of the Wnt pathway, Wnt proteins bind to their receptors on the cell surface and transduce the signal (through several cytoplasmic relay components) to beta-catenin, which then enters the nucleus and forms a complex with TCF to activate transcription of Wnt target genes. A clear description of this model and an earlier numerical analysis can be found in the paper by Lee et al. (2003). The same analysis could be repeated in Simpathica within about a week (as described in the following), involving few steps.

Step 1: First, we took each reactant and each reaction and entered them into Simpathica. All we needed to do was to input the reactants' names and concentrations, and for each reaction list the reactants, products, and rate constants. We obtained almost all of the data from the article by Lee et al. (2003) with one exception. Instead

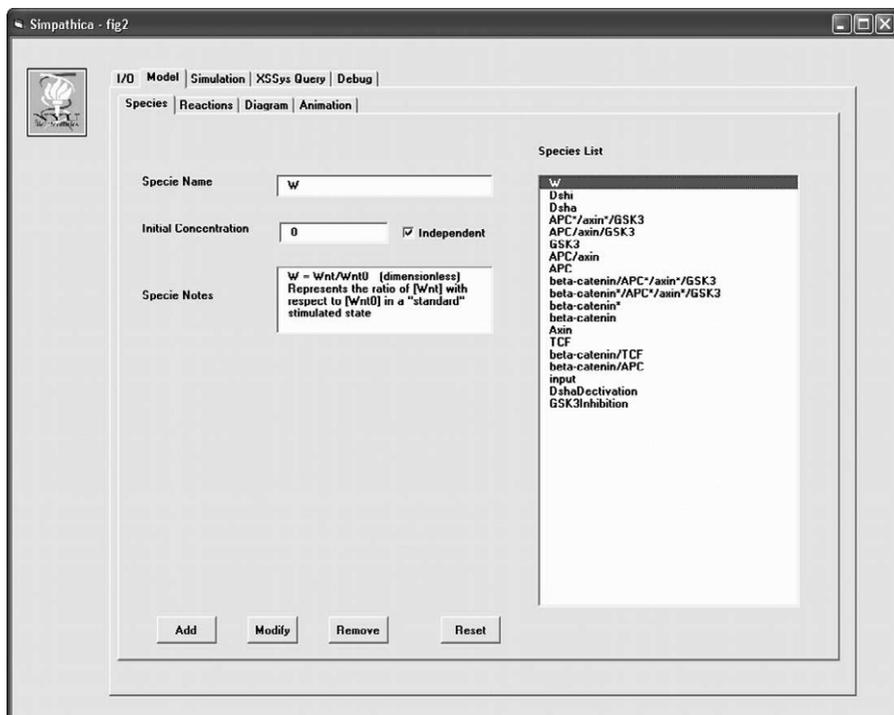


Figure 5.6. List of reactants in Wnt pathway entered in Simpathica.

of using a rapid equilibrium approximation as in Lee et al. (2003), we made educated guesses for the forward and backward rate constants that would be consistent with fast enzymatic reactions reaching equilibrium quickly. These differences may explain some discrepancies in the scale of the results. Simpathica automatically generates the entire pathway graphically and computes a system of differential equations to simulate the system evolution over time. (See Figures 5.6 and 5.7.)

Step 2: Next we checked that the system had different steady states under the two different conditions corresponding to the presence or absence of Wnt. These can be tested by queries: $W = 0$ implies eventually steady_state() and $W = 1$ implies eventually steady_state(). We can now compare the steady-state concentrations generated by our simulation to the experimental data. (See Figure 5.8.)

Step 3: Further validation of the model is obtained studying the degradation rate of beta-catenin under different conditions. We can reproduce different experimental settings simply by parameterizing initial concentrations or rate constants through Python scripts. (See Figure 5.9.)

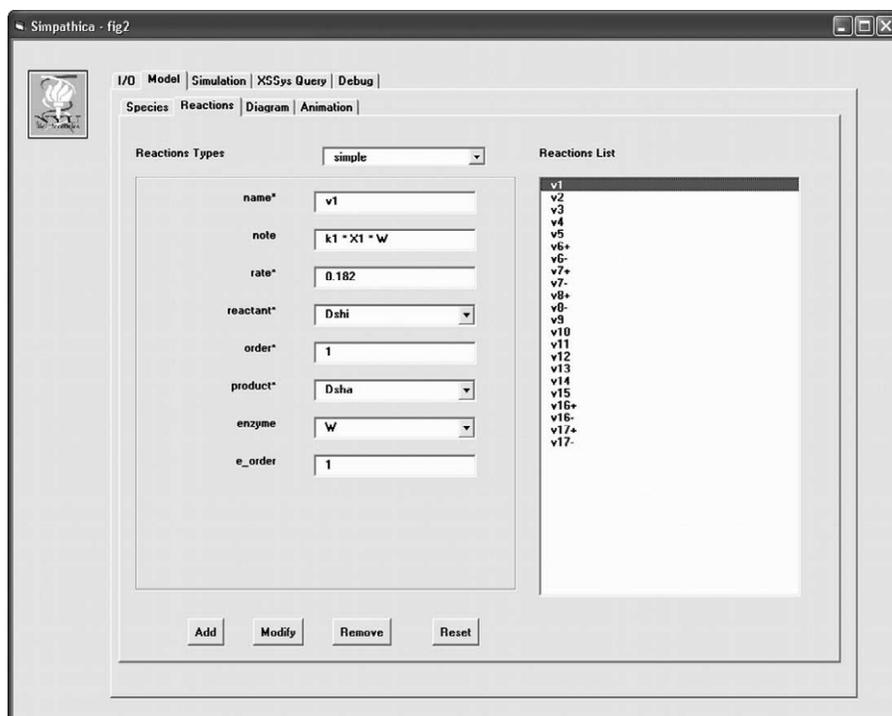


Figure 5.7. List of reactions in Wnt pathway entered in Simpathica.

Step 4: Finally, we can model the transient Wnt stimulation, where Wnt is present at the beginning of the simulation but then decays exponentially. (See Figures 5.10 and 5.11.)

Following the analysis (presented in the Lee et al. (2003) paper), we also noticed that beta-catenin's increase is only temporary, whereas axin remains down-regulated. Moreover, the response by axin precedes that of beta-catenin.

V. CONCLUSIONS

Many scientists and engineers have articulated that the biology of the new millennium needs a "regime change" and that the formal tools from systems sciences, with their rigor and depth, are desperately needed. And yet in spite of such noble goals systems biologists still wait patiently to be greeted as liberators by the vast majority of biologists. Perhaps in this lies the grandest of all challenges for systems biologists.

The most important grand challenge concerns better measurements and experiment design, as well as making data available in an electronic public forum. The

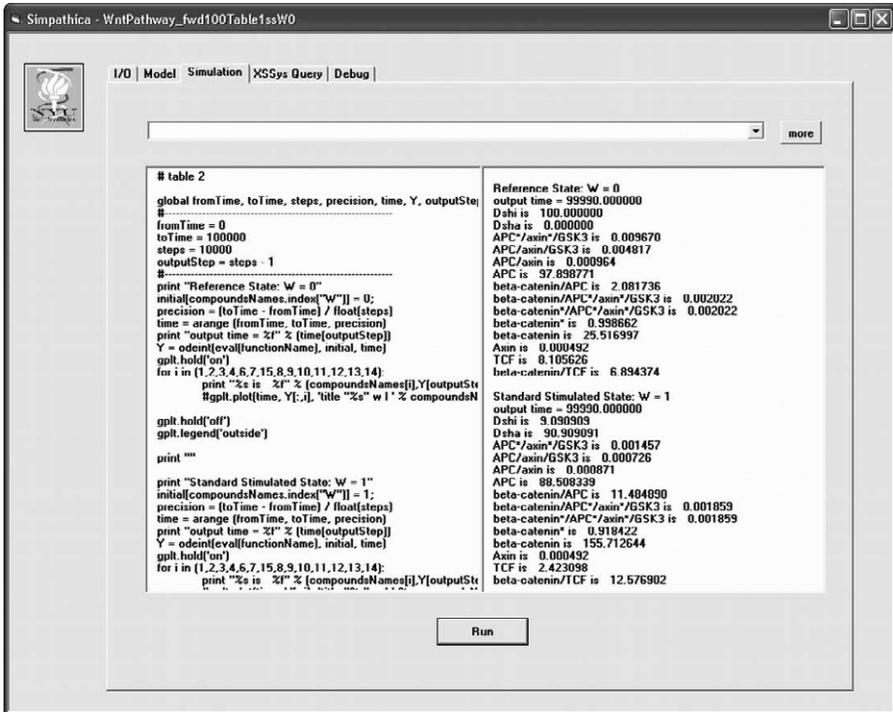


Figure 5.8. Steady-state analysis for Wnt pathway for different values of Wnt.

solution should comprise steps to intervene and measure at the single-molecule and single-cell levels, publication of the experimental data using a clear and unambiguous lexicon, and the ability to conduct experiments inexpensively with facilities that can be shared by the entire community. A community of biologists working within a social framework, where each scientist contributes from his or her own accumulated knowledge and experience, can create the needed lexicon and ontology. Software to ease the communication among scientists is not difficult, but does not exist at this point.

There should be a public database of biological models at various spatio-temporal resolutions and with as much of the *in vitro* and *in vivo* kinetic parameters as is possible to compile. Experiments at single-cell and population levels using wild-type cells, mutants, cells perturbed by different conditions, or RNA interference should be cataloged with precise time-course measurements. Along these directions, it will be worthwhile to focus on a complete map of pathways for one organism, say *C. elegans*. This digital worm, which can be dubbed *C⁺elegans*, could provide an enhanced environment for *in silico* experiments. Other pathways of interest might be cell cycles, proliferation, degradation, and apoptosis. Ultimately, a focus on models of aging and diseases will be of considerable human interest.

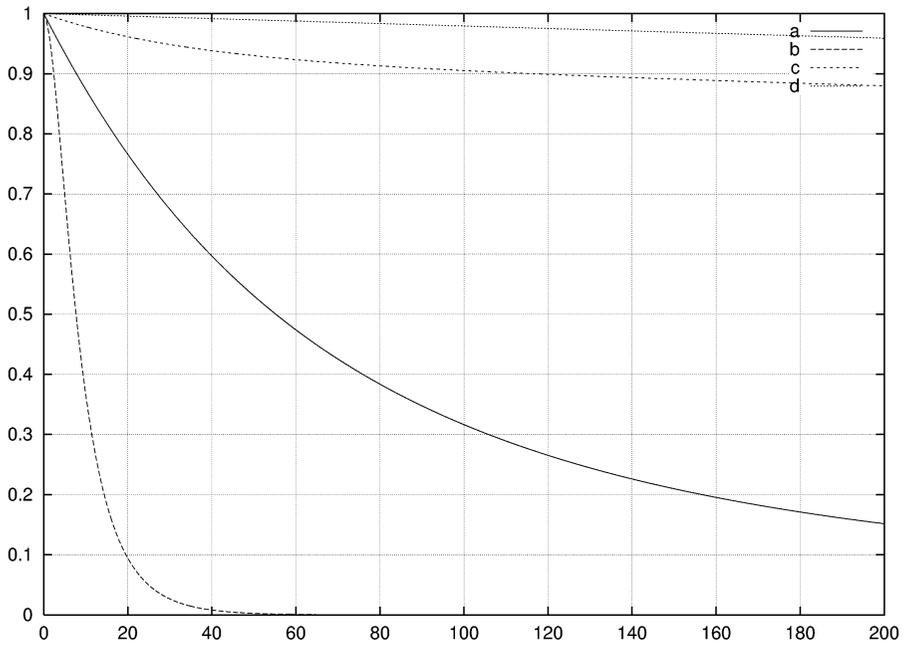


Figure 5.9. Kinetics of beta-catenin degradation.

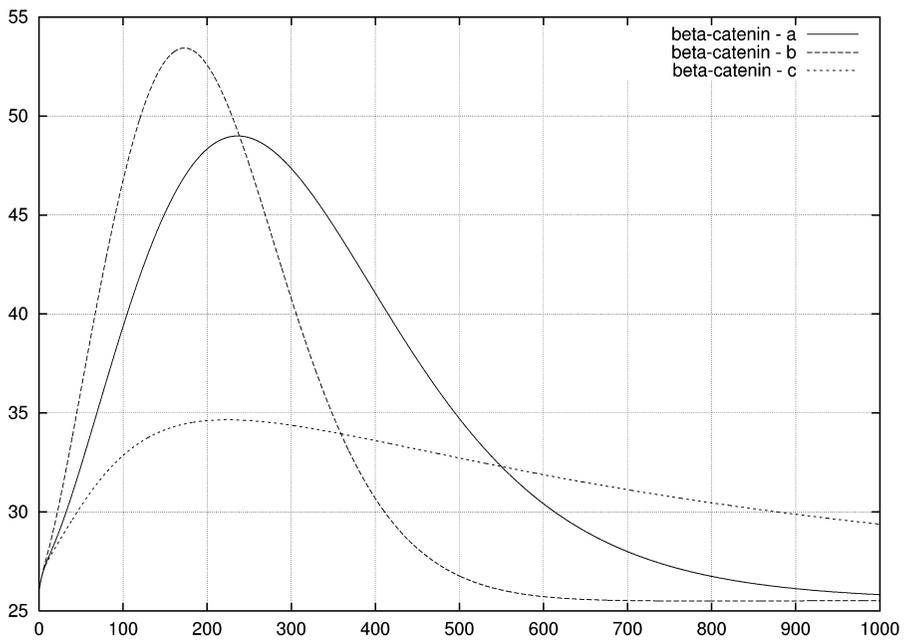


Figure 5.10. Beta-catenin response.

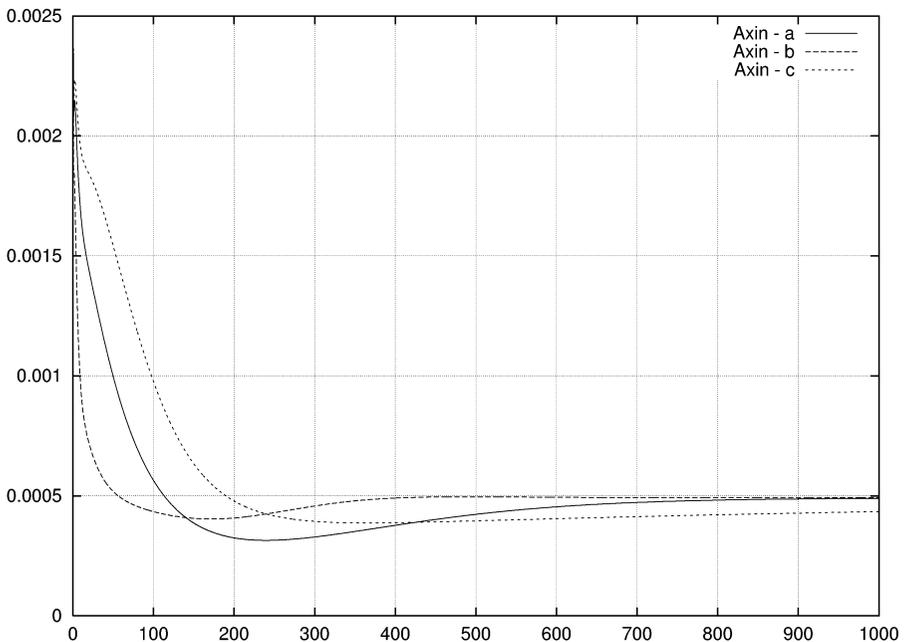


Figure 5.11. Axin response.

Thus, the purely technical grand challenges for this field will be experimental and computational, and will stay with us for a considerably long time. Most of these computational problems deal with accuracy and uncertainty in the model, model complexity, and computational complexity.

- *Reactions models*: Instead of just ODE models using DAEs, one must generalize our tools to PDEs (incorporating spatial properties), SDEs (small population size for interacting molecules), and hybrid models (part continuous, part discrete, but also spatial and probabilistic) in one general framework.
- *State space (product space)*: A number of interacting cells can be modeled by product automata. In addition to the classical "state-explosion problem," we need to pay attention to the variable structure due to (a) cell division, (b) apoptosis, and (c) differentiation.
- *Communication*: We need to model communication among cells mediated by interactions between extracellular factors and external receptors, efficiently and accurately.

We believe that the solution to such computational grand challenges is in reduction of complexity by *hierarchical modeling* and *symbolic modeling*. As we go to more and more complex cellular processes, a clear understanding can be obtained only through modularized hierarchical models. For this process to succeed, we

will need to derive simple I/O models of low-level modules by projection (elimination of state variables) or by reduction (state collapsing), while retaining bisimulation properties. The system dynamics should have a succinct symbolic representation that can be manipulated algebraically (without explicit and exhausting simulation).

For instance, in the case of a hybrid automaton model one may be able to represent flow, invariant, jump, and reset conditions, with a subset of the kinetic parameters left as unknown variables (e.g., k_1, k_2, \dots, k_n). By algebraically manipulating the equations (and inequations and inequalities), one can elicit many biological properties of the system in terms of constraints on the unknown and unmeasured variables and parameters. Interestingly enough, because of a similar development of symbolic (and to a less significant degree, hierarchical) model checking procedures in the discrete asynchronous setting we have been able to tame the computational complexity of computer-aided verification of complex and large engineered systems such as VLSI circuits (Browne et al. 1986; Clarke et al. 1999).

ACKNOWLEDGMENTS

All correspondence should be addressed to *mishra@nyu.edu*. The work reported in this chapter was supported by grants from the NSF's ITR programs, Defense Advanced Research Projects Agency (DARPA)'s BioCOMP program, and New York State Office of Science, Technology & Academic Research (NYSTAR).

REFERENCES

- Anantharaman, T. S., Mishra, B., and Schwartz, D. C. (1997). Genomics via Optical Mapping II: Ordered restriction maps. *Journal of Computational Biology* 4(2):91–118.
- Anantharaman, T. S., Mysore, V., and Mishra, B. (2005). Fast and cheap genome-wide haplotype construction via optical mapping. In (R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, eds.), *Proceedings of the Pacific Symposium on Biocomputing*. Singapore: World Scientific.
- Antoniotti, M., Policriti, A., Ugel, N., and Mishra, B. (2002). XS-systems: Extended S-systems and algebraic differential automata for modeling cellular behaviour. In (S. Sahni, V. K. Prasanna, and U. Shukla, eds.), *Proceedings of HiPC 2002*, pp. 431–442. New York: Springer-Verlag.
- Antoniotti, M., Park, F. C., Policriti, A., Ugel, N., and Mishra, B. (2003a). Foundations of a query and simulation system for the modeling of biochemical and biological processes. In (R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, eds.), *Proceedings of the Pacific Symposium of Biocomputing*, pp. 116–127. Singapore: World Scientific.
- Antoniotti, M., Piazza, C., Policriti, A., Simeoni, M., and Mishra, B. (2003b). Modeling cellular behavior with hybrid automata: Bisimulation and collapsing. *Computational Methods in Systems Biology*, (C. Priami, ed.), *Lecture Notes in Computer Science*: 2602, pp. 57–74. New York: Springer-Verlag.

- Antoniotti, M., Policriti, A., Ugel, N., and Mishra, B. (2003c). Model building and model checking for biological processes. *Cell Biochemistry and Biophysics* **38**:271–286.
- Aston, C., Schwartz, D. C., and Mishra, B. (1999). Optical mapping and its potential for large-scale sequencing projects. *Trends in Biotechnology* **17**:297–302.
- Browne, M. C., Clarke, E. M., Dill, D., and Mishra, B. (1986). Automatic verification of sequential circuits using temporal logic. *IEEE Trans. Computers* **35**(12):1035–1044.
- Clarke, E. M., Grumberg, O., and Peled, D. (1999). *Model Checking*. Cambridge, MA: MIT Press.
- Lee, E., Salic, A., Krüger, R., Heinrich, R., and Kirschner, M. W. (2003). The roles of APC and axin derived from experimental and theoretical analysis of the Wnt pathway. *Biology* **1**:116–132.
- Mishra, B. (2002a). Comparing genomes special issue on biocomputation. *Computing in Science and Engineering* **4**(1):42–49.
- Mishra, B. (2002b). A symbolic approach to modeling cellular behavior. In (S. Sahni, V. K. Prasanna, and U. Shukla, eds.), *Proceedings of HiPC 2002*, pp. 725–732. New York: Springer-Verlag.
- Mishra, B. (2003). *Optical Mapping Encyclopedia of the Human Genome*, pp. 448–453, London: Nature Publishing Group, Macmillan Publishers.
- Mishra, B., Daruwala, R., Zhou, Y., Ugel, N., Policriti, A., Antoniotti, M., Paxia, S., Rejali, M., Rudra, A., Cherepinsky, V., Silver, N., Casey, W., Piazza, C., Simeoni, M., Barbano, P. E., Spivak, M., Feng, J-W., Gill, O., Venkatesh, M., Cheng, F., Sun, B., Ioniata, I., Anantharaman, T. S., Hubbard, E. J. A., Pnueli, A., Harel, D., Chandru, V., Hariharan, R., Wigler, M., Park, F., Lin, S-C., Lazebnik, Y., Winkler, F., Cantor, C., Carbone, A., and Gromov, M. (2003). A sense of life: Computational and experimental investigations with models of biochemical and evolutionary processes. *OMICS* **7**(3):253–268.
- Paxia, S., Rudra, A., Zhou, Y., and Mishra, B. (2002). A random walk down the genomes: DNA evolution in VALIS. *Computer* **35**(7):73–79.
- SBML (System Biology Markup Language). (2002). www.sbml.org.
- Voit, E. O. (1991). *Canonical Nonlinear Modeling: S-system Approach to Understanding Complexity*. New York: Van Nostrand Reinhold.
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge: Cambridge University Press.

Standards, Platforms, and Applications

Herbert M. Sauro

Keck Graduate Institute, Claremont, California, USA

Chapter 6

ABSTRACT

With the sequencing of the human genome, it has become apparent that systems biology (the understanding of cellular networks through dynamical analysis) is becoming an important part of research for mainstream biologists. One of the indicative trends to emerge in recent years is the development of model interchange standards that permit biologists to easily exchange dynamical models between different software tools. In this chapter, two chief model exchange standards (SBML and CellML) are described. In addition, the development of extensible software frameworks (including SBW, BioSPICE, and BioUML) and the role they might play in stimulating the development of new tools and approaches are examined. Finally, the range of possible computational applications is described, highlighting the rich set of tools emerging as systems biology becomes a mainstream science.

I. INTRODUCTION

Although computational systems biology may seem to be a recent field of endeavor, its origins can be traced as far back as the 1920s and 1930s (Wright 1929). During this period, it was already believed by some that genes were responsible in some way for specifying enzymes. It was also about this time that glycolysis, the first metabolic pathway, was being elucidated. This period also saw the beginnings of the idea that enzymes formed linked sequences called pathways. It is therefore even more remarkable that given the infancy of these concepts Sewell Wright attempted at the time to give a physiological explanation for the occurrence of

genetic dominance and recessivity (Wright 1934). Wright argued that the explanation for the origin of dominance lay with the properties of catalytic networks, and laid out an initial mathematical theory that described the properties of enzyme networks (this early work later became significant during the development of metabolic control analysis [Kacser and Burns 1981]).

In the 1940s, as the first digital computers were being built, pioneering individuals such as Garfinkel, Higgins, and Chance began investigating the possibility of modeling the subtle behavior of biochemical pathways. Even before the advent of the digital computer, the same group had been using analog computers to model simple biochemical pathways for almost 15 years (Chance 1943; Higgins 1959; Garfinkel et al. 1961).

Since the work of further pioneers in the 1950s, there have been many small groups that have continued this line of inquiry and that together laid the foundation for many of the techniques and theory we use today and take for granted in contemporary systems biology. It should be noted that there is a large body of literature, particularly in the *Journal of Theoretical Biology*, dating back 50 years that many newcomers to the field will find useful to consider.

A. What is systems biology?

There are many conflicting opinions today on what exactly systems biology is. Historically, the answer seems clear. The chief aim of systems biology is to understand how individual proteins, metabolites, and genes contribute quantitatively to the phenotypic response. Lee Hood (president of the Institute of Systems Biology in Seattle, Washington) defines it similarly as “the identification of the elements in a system and the analysis of their interrelationships so as to explain the emergent properties of the system.” Even so, some believe systems biology to be concerned with the collection of high-throughput data, whereas others consider the elucidation of protein-to-protein networks and gene networks to be its hallmark. Certainly both are vital prerequisites for understanding systems, but neither alone can offer great insight into how networks operate dynamically. Systems biology is the natural progression of classical molecular biology from a descriptive to a quantitative science and is concerned with the dynamic response of biological networks.

B. Statement of the problem

Building models is not an entirely new approach to biology. If one examines any textbook on molecular biology or biochemistry, virtually every page has a diagram of a model. These models, which are often termed cartoon-based models, represent the culmination of years of painstaking research. They serve as repositories of accepted doctrine and the starting point for the generation of new hypotheses. There are, however, limits to what can be done with these models, their predictive value tends to be poor, and the ability to reason using qualitative models is limited. In other sciences, these limitations are avoided through the use of

quantitative models, which are described not just pictorially but mathematically. Quantitative models by their nature have much better predictive value compared to qualitative models, but their real usefulness stems from the capacity to carry out precise reasoning with them.

II. QUANTITATIVE APPROACHES

There is a wide range of mathematical representations that one can use to build quantitative models, the choice of approach depending on the type of biological question, the accessibility of experimental data, and the tractability of the mathematics. Modeling representations are depicted in Figure 6.1. Probably the most successful and widely used models are those based on differential equations (both ordinary and partial). These models assume a continuum of concentrations and rates. In reality, of course, cellular systems are discrete at the molecular level. However, because the number of molecules is very large the continuum approximation turns out to be very good. When the number of molecules drops to below a certain threshold, the continuum model can break down and in these cases one must revert to stochastic simulation.

The disadvantage of a stochastic simulation is that all analytical methods available for continuous models no longer apply. One should therefore use stochastic simulation only if it is absolutely necessary, and never in cases for which an ODE-based model adequately describes the data. This problem highlights the need to develop a new set of mathematical approaches in order to understand the dynamics of stochastic systems. There are other approaches (including Boolean, Bayesian, formal logic, and connectivity studies), but these have yet to show any overwhelming advantage over continuum-based models. In this chapter we are largely concerned with models based on differential equations, and to a lesser extent with those based on stochastic equations.

A. Quantitative models based on differential equations

It is probably fair to say that most of the successful models to be found in the literature are based on ordinary differential equations. Many researchers will express these models using Equation 6.1.

$$\frac{d\mathbf{S}}{dt} = \mathbf{N}\mathbf{v}(\mathbf{S}(\mathbf{p}), \mathbf{p}) \quad (6.1)$$

Here, \mathbf{S} is the vector of molecular species concentration, \mathbf{N} is the stoichiometry matrix, \mathbf{v} is the rate vector, and \mathbf{p} is a vector of parameters that can influence the evolution of the system. Real cellular networks have an additional property that is particularly characteristic of biological networks. This is the presence of so-called moiety-conserved cycles. Depending on the time scale of a study, there will be

- Boolean:** One of the simplest possible modelling techniques is to represent a network using Boolean logic (deJong 2002). This approach has been used to model gene networks.
- Ordinary differential equations (ODEs):** This is the commonest and arguably most useful representation. Although based on a continuum model, ODE models have proved to be excellent descriptions of many biological systems. Another advantage to using ODEs is the wide range of analytical and numerical methods that are available. The analytical methods in particular provide a means to gain a deeper insight into the workings of the model.
- Deterministic hybrid:** A deterministic hybrid model is one which combines a continuous model (e.g. ODE model) with discrete events. These models are notoriously difficult to solve efficiently and require carefully crafted numerical solvers. The events can occur either in the state variables or parameters and can be time dependent or independent. A simple example involves the division of a cell into two daughter cells. This event can be treated as a discrete event which occurs when the volume of the cell reaches some preset value at which point the volume halves.
- Differential-algebraic equations (DAEs):** Sometimes a model requires constraints on the variables during the solution of the ODEs. Such a situation is often termed a DAE system. The simplest constraints are mass conservation constraints, however these are linear and can be handled efficiently and easily using simple assignment equations (see equation 2). DAE solvers need only be used when the constraints are nonlinear.
- Partial differential equations (PDEs):** Whereas simple ODEs model well-stirred reactors, PDEs can be used model heterogenous spatial models.
- Stochastic:** At the molecular level concentrations are discrete, but as long as the concentrations levels are sufficiently high, the continuous model is perfectly adequate. When concentrations fall below approximately one hundred molecules in the volume considered (e.g. the cell or compartment) has to consider using stochastic modelling. The great disadvantage in this approach is that one loses almost all the analytical methods that are available for continuous models, as a result stochastic models are much more difficult to interpret.

Figure 6.1. A nonexhaustive selection of mathematical techniques for modeling biological systems.

molecular subgroups conserved during the evolution of a network. These are termed conserved moieties (Reich and Selkov 1981).

The total amount of a particular moiety in a network is time invariant and is determined solely by the initial conditions imposed on the system.¹

¹ There are rare cases when a “conservation” relationship arises out of a non-moiety cycle. This does not affect the mathematics but only the physical interpretation of the relationship. For example, $A \rightarrow B + C$; $B + C \rightarrow D$ has the conservation $B - C = \text{constant}$.

In metabolism, conserved cycles act as common conveyers of energy (ATP) or reducing power (NAD). In signaling pathways they occur as protein phosphorylation states, whereas in genetic networks they occur as bound and unbound protein states to DNA. These conserved cycles will often have a profound effect on the network behavior, and it is important that they be properly considered in computational models.

From the full set of molecular species in a model, it is customary to divide the set into two groups: the dependent (\mathbf{S}_d) and independent set (\mathbf{S}_i). This division is dependent entirely on the number and type of conserved cycles in the network. If there are no conserved cycles in a model, then the dependent set is empty and the size of the independent set equals the number of molecular species in the model. For details on how to compute \mathbf{S}_d and \mathbf{S}_i , the reader should consult Sauro and Ingalls (2004).

In many cases it is vital that the separation into dependent and independent species be made. For simple time-course simulations, the separation is not as important, but for most other analyses it is critical and for stiff integration methods highly desirable. The reason is that many numerical methods, including the stiff integrators, employ a measure called the Jacobian matrix as part of the numerical calculation. If the separation is not carried out, the Jacobian becomes singular and renders most analyses (e.g., steady-state location, bifurcation analysis, certain optimization methods, sensitivity methods, and so on) numerically unstable (if not impossible). Even when carrying out simple time-course simulations, the separation is also useful because it enables the number of differential equations to be reduced in number and thereby improves computational efficiency. Equation 6.1 is therefore better expressed as

$$\begin{aligned} \mathbf{S}_d &= \mathbf{L}_0 \mathbf{S}_i + \mathbf{T} \\ \frac{d\mathbf{S}_i}{dt} &= \mathbf{N}_r \mathbf{v}(\mathbf{S}_i(\mathbf{p}), \mathbf{S}_d, \mathbf{p}) \end{aligned} \quad (6.2)$$

In these equations, \mathbf{S}_i is the vector of independent species, \mathbf{S}_d is the vector of dependent species, \mathbf{L}_0 is the link matrix, \mathbf{T} is the total mass vector, \mathbf{N}_r is the reduced stoichiometry matrix, \mathbf{v} is the rate vector, and \mathbf{p} is the vector of parameters. Equation 6.2 constitutes the most general expression of an ODE-based temporal model (Heinrich and Schuster 1996; Hofmeyr 2001). The symbolism used in Equation 6.2 is the standard notation used by many in the systems biology community.

Although mathematically reaction-based models are given by Equations 6.1 and 6.2, many researchers are more familiar with expressing models in the form of a reaction scheme. For example, the following describes part of glycolysis (see Figure 6.2).

For brevity, the rate laws that accompany each reaction have been left out. Such notation is well understood by biologists. It is not straightforward, however, to convert this representation to the representation given by Equation 6.2. However, many software tools will permit users to enter models as a list of reactions and then

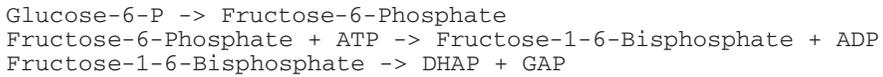


Figure 6.2. Part of glycolysis described using a reaction scheme notation.

automatically generate the mathematical model (Sauro and Fell 1991; Sauro 2000; Sauro et al. 2003).

B. Standards

In recent years, particularly since the sequencing of the human genome, there has been an ever-increasing list of wide-ranging cellular models published in the literature. Each author has a particular notation they use to publish the model. Some authors will publish the model as a reaction scheme, much like the notation given in Figure 6.2. Others will itemize the actual mathematical representation in the form of a list of differential equations. Some authors do not publish the model at all but provide the model as supplementary information. Until recently, there has been no way to publish models in a standard format. Without a standard format it has proved very difficult, if not impossible, in many cases to use published models without considerable effort.

As a result of this obvious shortcoming, a number of groups set out to gather community support to develop a standard that model developers would be happy to use. There was an early effort in 1998 by the BTK (BioThermoKinetics) group to standardize on a practical format for exchanging models between Gepasi (Mendes 1993) and SCAMP (Sauro and Fell 1991)—tools widely used at the time. About the same time, bioengineers at the University of Auckland began investigating the role XML (Harold and Means 2001) could play in defining a standard for exchanging computational models in order to reduce errors that appeared frequently in published models.

From the Auckland team emerged CellML (Lloyd et al. 2004). Members from the BTK group subsequently took their experience and contributed significantly to the other major model exchange standard, called SBML (Hucka et al. 2003). SBML was developed in 2000 at Caltech, Pasadena, as a result of funding received from the Japanese ERATO program. Both CellML and SBML are today viewed as the main standards for exchanging cellular network models. There are, however, fundamental differences between the approaches CellML and SBML take in the way models are represented.

1. CellML

CellML (Lloyd et al. 2004) represents cellular models using a mathematical description similar to Equation 6.1. In addition, CellML represents entities using a

component-based approach in which relationships between components are represented by connections. In many ways, CellML represents a literal translation of the mathematical equations, except that the relationship between dependent and independent species is implied rather than explicit. The literal translation of the mathematics, however, goes much further. In fact, the representation CellML uses is very reminiscent of the way an engineer might wire an analog computer to solve the equations (though without specifying the integrators).

As a result, CellML is very general and in principal could probably represent any system that has a mathematical description (not just the type indicated by Equation 6.1). CellML is also very precise in that every item in a model is defined explicitly. However, the generality and explicit nature of CellML also results in increased complexity, especially for software developers. Another side effect of the increased complexity is that models that are represented using CellML tend to be quite large. On average, my own analysis of a sample from the CellML repository www.cellml.org/examples/repository/ indicates that each reaction in a model requires about 5 Kbytes of storage.

Another key aspect of CellML is its provision for metadata support. The metadata can be used to provide a context for a model, such as the author name, when it was created, and what additional documents are available for its description. CellML uses standard XML-based metadata containers such as RDF (and within RDF the Dublin Core). The CellML team has amassed a very large suite (hundreds) of models, which provides many real examples of CellML syntax. This is an extremely useful resource for the community.

Owing to the complexity of CellML, one unfortunate side effect is that there are very few tools that can read and write CellML. As far as the author is aware, there are only two third-party tools that can read and write CellML. These are VCell (Loew and Schaff 2001) and COR (Garny et al. 2003). The CellML team has recently (2004, <http://cellml.sourceforge.net/>) begun to provide their own software tools to third-party developers. The delay in providing such tools to the community is probably one reason CellML (given its complexity) has not proved as popular relative to SBML.

2. SBML

Whereas CellML attempts to be highly comprehensive, SBML was designed to meet the immediate needs of the modeling community and is therefore more focused on a particular problem set. One result of this is that the standard is much simpler and much less verbose. Like CellML, SBML is based on XML. However, unlike CellML it takes a different approach to representing cellular models. The way SBML represents models closely maps the way existing modeling packages represent models. Whereas CellML represents models as a mathematical wiring diagram, SBML represents models as a list of chemical transformations (Figure 6.2).

Because every process in a biological cell can ultimately be broken down into one or more chemical transformations, this was the natural representation to use.

However, SBML does not have generalized elements such as components and connections. SBML employs specific elements to represent spatial compartments, molecular species, and chemical transformations. In addition to these, SBML has provision for rules that can be used to represent constraints, derived values, and general math that for one reason or another cannot be transformed into a chemical scheme. Like CellML, the dependent and independent species are implied.

3. SBML development tools

Early on in the development of SBML, the original authors decided to provide software tools almost immediately for the community. Because XML at the time was not well understood by many software developers, the provision of such assistance was crucial. In hindsight, this is probably one reason SBML has become a popular standard. Initially, the original authors provided a simple library for the Windows platform because the bulk of biology-based users tend to be Windows users. Today, this library is still used by a number of tools, including Gepasi, Jarnac, and JDesigner. With the growing popularity of SBML, the community has since developed a comprehensive cross-platform tool (<http://sbml.sourceforge.org>) that is now the recommended SBML toolkit to use (libSBML). libSBML was developed in C/C++ for maximum portability.

4. Extensibility

It was realized early on by the authors of SBML that as systems biology developed there would be pressure from the community to make additional functionality available in SBML. To address this issue, SBML has a formal means for adding extensions in the form of so-called annotations. There now exists a number of annotations used by software developers. Some of these address issues such as providing visualization information to allow software tools to render the model in some meaningful way (two examples of these are given in a later section).

Other extensions provide a means of storing information necessary for flux balance analysis or for stochastic simulations. Ultimately, some of the extensions will most likely be folded into the official SBML standard. This mechanism, a sort of Darwinian evolution, permits the most important and popular requests to be made part of SBML. It makes the process of SBML evolution more transparent and permits users to be more involved in the development of SBML.

5. Practical considerations

Whereas CellML is very general, SBML is more specific. As a result, the storage requirement for SBML is much less. It takes on average roughly 1.5 Kbytes to store a single chemical transformation in SBML Level 2 (compared to 5 K for CellML). Interestingly, it only takes roughly 50 to 100 bytes to store single transformations in

raw binary format where there is minimal extraneous syntax. Some readers may feel that with today's cheap storage technologies discussion of storage requirements is unnecessary.

Indeed, for small models storage is not an issue. However, in the future very large models are likely to be developed. There is, for example, a serious attempt (www.physiome.org) now underway to model in the long term entire organs and even entire organisms. The amount of information in these cases is huge, and the question of efficient storage is not trivial. Obviously, XML is highly compressible, and large models can be stored in this way. However, inefficient storage also increases the time taken to manipulate the models. Furthermore, in a modeling environment model authors tend to generate hundreds of variants while developing the model. For a large model, this clearly would generate huge amounts of XML-based data. One of the things yet to be addressed by either standard is how model variants can be efficiently stored.

6. Usage

Both SBML and CellML have been taken up by many software developers and implemented in their software. SBML is being used in more than 75 software projects. In addition, SBML is the official model interchange format for the BioSPICE project (www.biospice.org), the SBW project (www.sysbio.org), the international *E. coli* alliance, and the receptor tyrosine kinase consortium. Much of the SBML support is in standalone applications. However, a number of database vendors have also decided to provide export of SBML as an option. Examples include *reactome*, *stke*, *sigpath*, and *biomodels.net*.

A related standard that has been proposed by Yun et al. (2004) is for the storage of flux balance models. The proposed format is very similar to SBML but has the additional feature of storing the flux balance objective function.

C. Future considerations

The development of standards for systems biology is still at a very early stage. I have not, for example, considered the problem of standardizing the formats for the experimental data that will be required for modeling. For example, there are no current standards for representing *quantitative* proteomic or metabolomic data, though efforts for defining a quantitative microarray format are maturing (www.mged.org).

More pressing from a modeling perspective is that there is currently no agreed way to merge smaller submodels into larger models (composition). One of the few groups to have considered composition is Ginkel and Kremling (Ginkel et al. 2000). They have examined possible extensions to SBML to allow SBML to represent submodels and models composed of submodels. Additional issues include distinguishing different types of models, particularly ODE and stochastic models. Currently there is no means of identifying the type of model an SBML file repre-

sents other than to use specific annotations. One unfortunate side effect of using XML is the temptation to omit a detailed semantic specification. XML is often vaunted as a desirable technology because it is easily parsed. However, parsing and syntax checking are tasks easily implemented. The real difficulty comes when semantic checks are required, and current XML technology offers no assistance in this task.

D. Other standards

Apart from using XML to define an interchange format, there are two other mediums for representing models: human-readable text-based formats and visual formats.

1. Visualization of models

For many users, the ability to visualize models and to build models using visual tools is an important feature. There are currently a number of visualization formats in common use. One of the most comprehensive and freely available formats is the molecular interaction maps developed by Kohn (1999), and more recently by Mirit Aladjem (Kohn et al. 2004). The Kohn format emerged from the need to represent complex signaling networks in a compact way. Unlike metabolic networks, signaling networks can be extremely complex with multiple protein states and interactions. Therefore, an alternative and more concise approach is desirable. At the time of writing, there is no software for manipulating Kohn maps and no means of converting Kohn maps to SBML or any other standard. Hopefully, this will change in the future.

An early computer-based visual notation was proposed by Cook et al. (2001), who developed a notation called BioD. This notation has been implemented in a commercial software package called KineCyte (www.rainbio.com/Software.html).

Another proposal has been put forward by Kitano (2003). This is a more traditional approach in which various molecular entities (such as proteins, ions, transporters, and so on) have particular pictorial representations. The software tool called cellDesigner (Funahashi et al. 2003) implements this proposed format.

One of the first visualization tools, JDesigner (Sauro et al. 2003), also implements a traditional means of depicting networks (see Figure 6.3) using a pictorial representation to indicate substances and reactions. JDesigner also employs Bezier curves to represent arcs in an attempt to make the diagrams similar to the notation found in many molecular biology textbooks. CellDesigner and JDesigner connect to the Systems Biology Workbench (SBW) for simulation support.

Finally, there is a proposal from a commercial company called Gene Network Sciences, which has devised a derivative of the Kohn notation called DCL. However, this notation is proprietary and its utility to the general scientific community is not certain at this time.

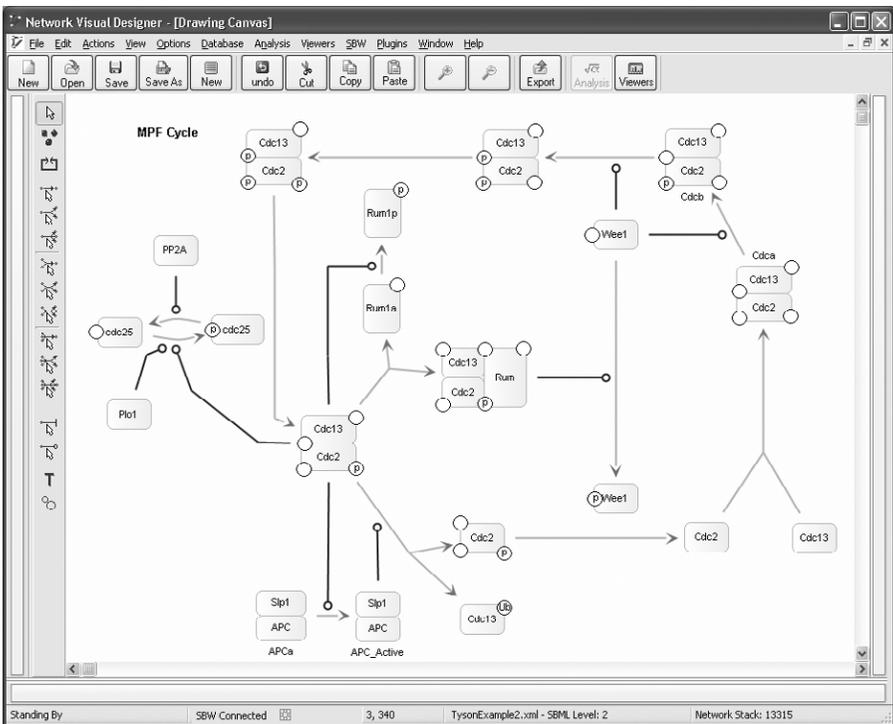


Figure 6.3. Example of JDesigner's visual format.

2. Human-readable formats

In addition to visualization approaches and the use of XML to represent models, there has been a long tradition in the field to describe models using human-readable text-based formats. Indeed, the very first simulator—BIOSIM (Garfinkel 1968)—allowed a user to describe a model using a list of reaction schemes. Variants of this have been employed by a number of simulators since, including SCAMP (Sauro and Fell 1991), Jarnac (Sauro 2000), ECell (Tomita et al. 1999), and more recently PySCeS (Olivier et al. 2005). Being able to represent models in a human-readable format offers many advantages, including conciseness, easily understood and manipulated using a simple editor, flexible, portable, and above all extremely easy to annotate.

E. Model databases

At the time of writing, there are very few model databases, exceptions include sigpath and biomodels.net.

Although model databases would be of great advantage to the community, the funding agencies have so far been reluctant to provide support. Instead, a number

of groups (including the original SBML group and the SBW group) are developing model databases as part of other projects. In particular, the Department of Energy is, through its GTL program, funding a small project to develop a database for microbial models. In addition the BioSPICE project funded by DARPA is supporting model curation for the biomodels.net project. What features of a database might be useful? Probably one of the most useful features for such a database (apart from the obvious ability to query the database for particular models, organisms, and so on) would be the ability to deliver models in different computationally ready formats.

F. Related standards

CellML and SBML are the primary formats used to store interchangeable dynamic models. Apart from the particular details on the model itself, there is also the need to consider data used to build the models. Most models are built by laboriously searching the literature and carrying out additional experiments as necessary to fill in gaps in the data. This has proved to be an extremely effective method of building reliable models (Tyson et al. 2001, 2002). However, many inexperienced researchers in systems biology feel that high-throughput data is the answer to the needs of the modeling community.

Unfortunately, much of the high-throughput data currently available is not appropriate. Much of the high-throughput data is very noisy and is probably more suitable for building qualitative models. More importantly, the bulk of high-throughput data is not generated with dynamic model building in mind and is therefore often not appropriate for this purpose. To date there has not been a single dynamic model that has been constructed as a result of high-throughput data. As systems biology and the construction of dynamic models become more important, it is very likely that the utility of high-throughput data will become much more significant. When this happens, a proposed standard called BioPAX (www.biopax.org) will most likely contribute.

BioPAX (Biological Pathway Exchange) is another proposed standard based on XML. BioPAX aims to integrate many of the incompatible pathway-related databases (such as BioCYC, BIND, WIT, aMAZE, KEGG, and others) so that data from any one of these databases can be easily interchanged. In the future it should be possible to extract data from many of the pathway databases and integrate the data directly into SBML (or CellML) via BioPAX. The BioPAX group proposes to embed BioPAX elements in SBML or CellML for unambiguous identification of substances (metabolites, enzymes) and reactions.

III. PLATFORMS

Much of the current software development in the systems biology community concentrates on the development of standalone applications. Most of these tools are not easily extensible and many of them offer nearly identical functionality. One of

the problems that currently plagues systems biology is the continual reinvention of the same type of tool (called YADS, for “yet another differential equation solver”). I believe it is not too unfair to suggest that in many cases our software capability today in systems biology is only marginally better than the first systems biology simulation package ever written (BIOSSIM) by David Garfinkel about 1960 (Garfinkel 1968).

In many cases, even the user interfaces are only marginally better. There are of course exceptions to this. VCell (Loew and Schaff 2001) in particular comes to mind, as well as tools such as Gepasi (Mendes 1993) and Jarnac/JDesigner (Sauro et al. 2003). VCell is particularly suited to spatial modeling. Gepasi is well known for its GUI user interface, the selection of optimization methods, and its ability to fit data to models. Jarnac was until very recently (e.g., Pysces (Olivier et al. 2005)) the only script-based programmable modeling tool with a fairly complete set of tools for the analysis of time-dependent ODEs and stochastic systems. JDesigner was the first visual design model tool.

The reason for the repetitive nature of software in systems biology is that almost every group engaged in computational systems biology writes its own simulation package. Given the time constraints on the project, the software will only reach a level of maturity that is often equivalent to BIOSSIM. As a result, the provision of software does not appear to advance.

A number of groups have recognized this problem and instead of developing single isolated applications they have chosen to develop a software infrastructure that permits and encourages extensibility and code reuse. Code reuse is extremely important because it allows developers to build on existing code, which in turn leads to new and interesting software tools. The following examines three such environments: SBW, BioSPICE, and BioUML. All three environments are open source.

A. SBW systems biology workbench

The SBW (Sauro et al. 2004) is an extensible software framework that is both platform and language independent. Its primary purpose is to encourage code reuse among members of the systems biology community. Developers can run SBW on Linux, Windows, or Mac OS and can develop software in a variety of different languages, including C/C++, Java, Delphi, FORTRAN, Matlab, Perl, Python, and any .NET language (e.g., Visual Basic or C#). The SBW was originally developed in parallel with SBML (Systems Biology Markup Language) as part of the Symbiotic Systems Project ERATO project at Caltech, Pasadena. (Subsequent development was supported by DARPA through the BioSPICE program, and development is now focused at the Keck Graduate Institute, with support from the Department of Energy.)

The central component of SBW is the broker, which is responsible for coordinating interactions among the various resources connected to it. These resources include simulation engines, model editors, SBML translators, databases, visualization tools, and a variety of analysis packages. All modules in SBW connect via

defined interfaces. This allows any one of the modules to be easily replaced if necessary. The key concept in SBW is that any new module may exploit resources provided by other modules. This dramatically improves productivity by allowing developers to build on existing tools rather than continuously reinvent.

Similar architectures have been developed, most notably CORBA. When SBW was being developed, CORBA was seriously considered but a number of problems arose. First, the learning curve for CORBA is very steep, which means that it is out of reach for most developers except highly skilled individuals. The aim of SBW was to allow the average computational biologist to develop new SBW modules, and hence the programming model had to be simple. In addition, there were very few open-source equivalents to the SBW broker and many of them were incompatible with each other.

An SBW module (the client) provides one or more interfaces or services. Each service provides one or more methods. Modules register the services they provide with the SBW broker. The module optionally places each service it provides into a category. By convention, a category is a group of services from one or more modules that have a common set of methods.

One of the key advantages of SBW is its language and OS neutrality. At a stroke, this eliminates the irrational language and operating systems “wars” that often plague software development. In addition to providing support for multiple languages there is the facility to automatically generate web services from any SBW module (Frank Bergmann, personal communication).

Messaging protocols

At the heart of SBW is the messaging protocol used to exchange information between the different modules. For efficiency reasons, messages that are exchanged between modules are simple sequences of binary data. For each programming language there is a language-binding library that takes care of much, if not all, of the housekeeping necessary to operate through SBW (Figure 6.4), including connection and transmission of data. In addition, issues such as little-endian and big-endian byte ordering need not concern the developer as this is taken care of automatically by the binding libraries. Each binding also provides the necessary message packing and unpacking logic and exposes functionality in the form of an easy-to-use API. Because SBW message passing is based on TCP/IP sockets, it is straightforward to run SBW across the Internet or more significantly across computational nodes on a supercomputer cluster.

B. BioSPICE

BioSPICE (www.biospice.org) is a DARPA-funded effort to develop an open-source framework and tool set for modeling dynamic cellular network functions. The central component of BioSPICE is the dashboard, which is used to construct workflows between BioSPICE-enabled applications. Both SBW and the dashboard

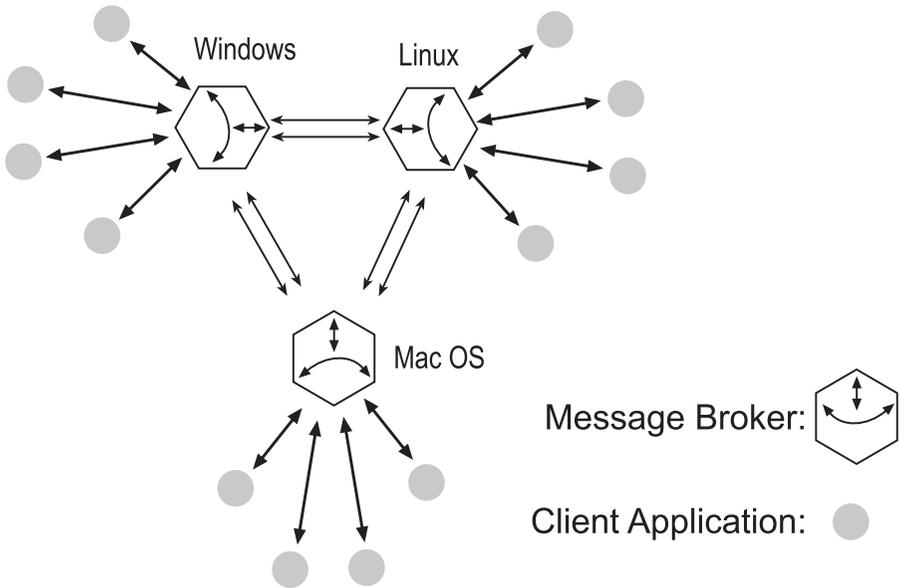


Figure 6.4. The Systems Biology Workbench (SBW) is a dynamic open-source distributed system. Client modules can attach and detach at runtime. Client modules can be written in a variety of languages, including C/C++, Java, Delphi, FORTRAN, Python, Perl, Matlab, and any .NET language. Data is exchanged between modules via binary messages that can include any combination of bytes, integers, floating points, complex numbers, strings, arrays, and lists. Currently, the available modules include simulators, model editors, SBML manipulation, math library, frequency analyzer, bifurcation discover and analysis modules, structural analysis modules, and others. Further details are found at www.sysbio.org.

encourage code reuse, although in different ways. In the dashboard, code reuse is through the construction of workflows. In SBW, code reuse is via programmatic interfaces and a pluggable runtime architecture. The unit components in SBW tend to be more fine-grained compared to BioSPICE modules.

For example, SBW provides modules such as SBML support, frequency analysis, simulation methods, and bifurcation analysis, which can be tied together at runtime to give the impression of a single application. The BioSPICE dashboard, on the other hand, allows the user to construct fixed workflows prior to a run. The workflow configurations cannot be changed during runtime. In addition, whereas SBW connects modules via interface specifications the dashboard connects modules via data types. The BioSPICE dashboard is based on the Java "net beans" application, which makes it highly Java centric. Interaction with applications written in other languages, though not impossible, are difficult. It is possible to easily connect SBW modules to the dashboard (via the SBW Java interface), which greatly increases the flexibility of the dashboard combining the advantages of a workflow approach to the free-flow approach of SBW.

C. BioUML

BioUML (www.biouml.org), developed by Fedor Kolpakov and his team, is a Java framework based around eclipse and targeted at the systems biology community. The authors state that the utility of BioUML covers access to databases with experimental data, tools for formalized description of biological systems structure and functioning, and tools for their visualization and simulation. BioUML is at an early stage of development, but the central idea is of a pluggable environment in which plug-ins written in Java are used to extend the functionality of the framework. Much work remains to make the BioUML usable for the average biologist, but the idea is interesting (although the requirement to write all code in Java is limiting and some means to permit alternative language bindings would be useful). Recently, the BioUML team developed an SBW interface that permits access to plug-ins written in many different languages.

IV. APPLICATIONS

In recent years there has been a proliferation of software applications for the systems biology community (See Figure 6.5). On the whole, many of these applications provide very similar functionality. The distinguishing feature among them is how easy they are to install and use. The more mature applications tend to be easier to install and have a much richer repertoire of functionality. Many of the applications are simple wrappers around standard ODE or Gillespie solvers and provide a simple means of loading models and runtime courses. Some of the applications fall by the wayside because the author has lost interest or funding has stopped. It is important therefore that whatever tool one uses the ability to export and import a recognized standard (or at least a documented format) such as SBML and/or CellML be available.

The original intention in this section was to list as many of the applications as possible, together with their capabilities, but given the large number now available

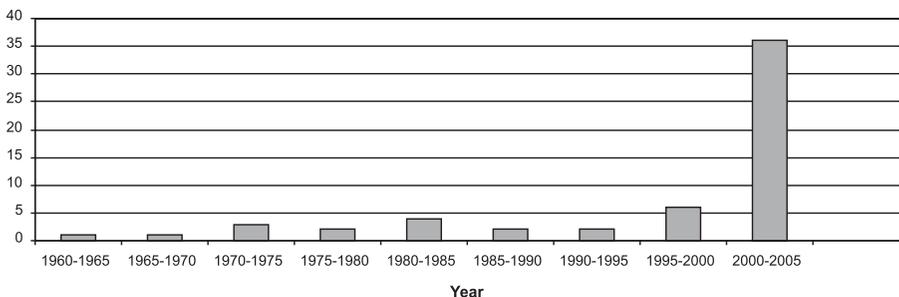


Figure 6.5. The release of software tools for computational systems biology over time. Note the spike in the last five years.

Table 6.1. Mature and easily accessible tools for modeling cellular networks.

Application	Description
VCell	A very mature server-based application specialized to build and simulate large-scale spatial (PDE) models. Open-source, multiplatform (Loew and Schaff 2001).
Gepasi	This is a form-based application that has been maintained for many years by a dedicated author. The tool is particularly adept at carrying out optimizations of ODE-based models to data. Closed-source, Windows, Linux (Mendes 1993).
WinSCAMP	A script-based GUI application that like Gepasi has a long tradition. Specialized for time-course, steady-state, and metabolic control analysis of ODE-based models. Source available upon request, multiplatform (Sauro and Fell 1991; Sauro 1993).
Pysces	A very complete ODE-based simulation environment built around the scripting language Python. Open-source, multiplatform (Olivier et al. 2005).
Jarnac/JDesigner	Jarnac is a script-based application. JDesigner (see Figure 6.3) is a visual design tool that can use Jarnac via SBW to carry out simulations. The simulation capabilities of Jarnac are quite extensive, offering both ODE and stochastic Simulation. Open-source, Windows, Linux (Sauro 2000; Sauro et al. 2003).

We do not have facilities for representing models in a way familiar to most biologists. Instead, users are required to derive the differential equations explicitly. Platforms such as SBW make available translators from SBML to a variety of formats, including Matlab, and in a number of cases users employ tools such as JDesigner to maintain the model, but use a translator to generate Matlab (or any other supported format such as C or Java).

it soon became clear that this task would be too great. Instead, I refer the reader to the recent paper by Hucka et al. (2004), in which the authors describe almost 40 applications. An even larger list can be found at the *sbml.org* web site.

There are some applications, however, that are worth mentioning specifically because they have some special characteristic. Table 6.1 lists a number of applications being actively maintained, have a reasonably large user base, and offer facilities that are either unique or well done. I have not mentioned any stochastic simulators in Table 6.1 because many of these are still immature.

There are also more general-purpose tools available, both commercial and open-source, which are worth considering. Probably the most well known commercial tool is Matlab (www.mathworks.com). Although Matlab is an excellent prototyping tool, it suffers from poor performance when simulating systems larger than about 30 species if the model is not specified in the correct way. In fact, a number of the open-source tools are orders of magnitude faster than Matlab. This stems from the fact that Matlab is a general-purpose tool, whereas the open-source tools are specialists and are therefore more heavily optimized for their specific application. The commercial tools require a high degree of programming skill because they do not have facilities for representing models in a way familiar to most biologists, instead users are required to derive the differential equations explicitly. Platforms such as SBW make available translators from SBML to a variety of formats including Matlab,

and in a number of cases, users employ tools such as JDesigner to maintain the model, but use a translator to generate Matlab (or any other supported format such as C or Java). In addition to generic commercial modelling tools there are also now available a number of commercial tools specifically geared for modelling cellular networks. The most well known include Gene Networks Sciences, Berkeley Madonna and Teranode (these can easily be located on the web by using a reliable search engine).

A. Model analysis

As a user, one of the most important aspects I consider is the range of techniques available for analyzing the model. The purpose of building a model is not simply to generate a predictive tool. If it were, we could probably get away with using empirical statistical techniques or machine learning approaches such as neural nets. An additional important role of model building is to gain a deeper understanding into the properties of the model and to understand how the structure of the model leads it to behave the way it does. To answer these types of questions, one needs techniques that can interrogate the model in a variety of ways. Table 6.2 lists some of the most important techniques available for analyzing models. Without these techniques, a model will often be as difficult to understand as the real system it attempts to model. The application of these techniques is therefore important.

All of these techniques are extremely useful in gaining insight into how a model operates. The connectionist and structural analyses focus on the network proper-

Table 6.2. Model analysis methods.

Approach	Description
Connectionist theory	Connectivity studies are centered around the search for patterns in the way cellular networks are physically connected (Barabasi and Oltvai 2004).
Structural analysis	There is a wide range of useful techniques that focus on the properties of the networks that depend on the mass conservation properties of networks. These include conservation analysis, flux balance, and elementary mode analysis (Heinrich and Schuster 1996).
Cellular control analysis	CCA (also known as metabolic control analysis) is a powerful technique for analyzing the propagation of perturbations through a network. There exists a very large literature describing applications and theory (Fell 1997).
Frequency analysis	Closely related to CCA is the analysis of how signals propagate through a network (Ingalls 2004; Rao et al. 2004).
Bifurcation analysis	Bifurcation analysis is concerned with the study of how the qualitative behavior of steady-state solutions change with changes in model parameters (Tyson et al. 2001).

ties of the model. That is, they do not explicitly consider the dynamics of the model but how the network connectivity sets the stage for generating the dynamics of the model. The last three techniques (CCA, frequency analysis, and bifurcation analysis) focus on the dynamical aspects of a model and are crucial to gaining a deep insight into the model (Bakker et al. 1997; Tyson et al. 2001).

B. Model fitting and validation

An important activity in systems biology modeling is the need to fit experimental data to models. There is not sufficient space to cover in any great detail this topic, but as time series data from microarray, proteomic, and metabolomic data becomes more readily available the need to fit models to experimental data will become more acute. There are a number of issues related to this topic, one of which concerns the nature of the data generated by most of the current experimental techniques.

In particular, most current techniques generate normalized data (i.e., absolute values are not given). This poses a number of problems for a fitting algorithm, in that the underlying model is established in terms of absolute quantities. A number of solutions are potentially available. However, none is entirely satisfactory and ultimately the models generated by normalized data will most likely be only capable of reproducing trends in the data. Whether such models will have great predictive value is open to question, and much research remains to be done in this area.

Another issue is the intensive nature of computations required to fit even a moderately sized model. One of the necessary requirements for fitting a model is estimating the confidence limits on the fitted parameters and the range of parameter space that describes the experimental data. This information is crucial to determine the validity of the model, and can be used to design additional experiments to either refute the model or increase the precision of the model parameters. As a result of these requirements, computing a global optimization can take considerable time.

For example, in a recent study Vijay Chickarmane (unpublished) estimated that the time required to fit a model of approximately 300 parameters would be on the order of seven years on a normal desktop computer. Luckily, global optimization can be easily parallelized given a suitable optimizer (for example, a genetic-algorithm-based optimizer) and the computation time can be reduced by hosting the problem on a cluster machine. Chickarmane estimates that using a thousand-node cluster the optimization of a 300-parameter model can be reduced to approximately two days of computation time. Such a computation can be easily set up using SBW. A single node on the cluster would act as the primary optimizer. In turn, this node would farm out the time-consuming simulation computations to the remaining nodes on the computer. For very large models, grid computing (Abbas 2004) may be very appropriate for solving this type of problem.

V. CONCLUSIONS

The systems biology field has been developing rapidly in recent years, but much remains to be done. One of the most useful developments must undoubtedly go to the development of standards such as SBML and CellML. Indeed, the most recent of a long list of new systems biology journals (*Molecular Systems Biology*) has stipulated that SBML is the preferred format for contributing models. It is hoped that other journals will follow. However, one aspect that still remains to be dealt with is to formalize the semantic rules for SBML. At the moment, there is no guarantee that models written by different tools can be interchanged. If one focuses on the core specification in SBML, this is generally not an issue. However, it is vital that semantic validators be developed for SBML.

The other area that has received a lot of attention in recent years is the development of tools for systems biology. However, much of what is being developed is repetitious and little true advancement is being made. This is probably due to the large number of newcomers to the field who are inexperienced and inevitably repeat what has gone before. A number of solutions exist to solve this problem. One is to develop extensible frameworks such as SBW, BioSPICE, or BioUML. The other is to develop a suite of open-source libraries that can carry out specific functionality. An example of this is libSBML, being developed by the SBML team. This library, written in C/C++ for maximum portability, enables other developers to concentrate on simulation capability rather than waste unnecessary effort developing their own SBML parser.

In terms of other possible libraries, examples include open-source Gillespie-based stochastic solvers and ODE solvers. In both cases, there is also the need to develop scalable and robust methods for computing the dependent and independent species. Furthermore, hybrid methods combining continuous and stochastic methods are a pressing need at the current time. Many biological systems interface noisy sensory apparatus (e.g., ligand binding to the surface of a cell membrane) to internal continuous analog networks (Sauro and Kholodenko 2004). In addition to the core solvers, we also need scalable analysis tools, particularly bifurcation analysis tools and sensitivity analysis tools.

On the model validation front, much remains to be done, particularly the relationship between model validation and how this can direct future experimentation. This leads to the need for development of new methods and algorithms for analyzing complex networks. In particular, methods should be developed to modularize large networks (in that understanding an entire network is virtually impossible without some recourse to a hierarchical modularization).

Finally, the role of high-performance computing in systems biology is still very novel. In fact, there appear to be very few applications to date of high-performance computing to systems biology. One of the few useful applications is model fitting to data. When done correctly, this is an extremely computationally intensive calculation and is an ideal candidate for large cluster machines. In fact, one wonders whether this is the application for systems biology that could benefit from grid computing.

ACKNOWLEDGMENTS

I would first like to acknowledge the generous support from the Japan Science and Technology Agency, DARPA (BAA0126 BioComputation), and the U.S. Department of Energy GTL program, without which the bulk of the work described in this chapter would not have been carried out. I would also like to acknowledge Mike Hucka, Andrew Finney, and Hamid Bolouri for their initial work on the Systems Biology Workbench, and in more recent years the tremendous programming work done by Frank Bergmann and the critical support given by Sri Paladugu and Vijay Chickarmane to the development of novel computational methods in systems biology.

RESOURCES

The following are four web sources of interest to readers of this chapter.

www.cellml.org This is the main CellML site. It has a very rich set of models expressed in CellML, including specifications for the standard and pointers to software toolkits.

www.sbml.org This is the main SBML site. The site has ample documentation, and examples illustrating how SBML is used and should be used. In addition, it has a rich set of software tools—in particular, libSBML, which allows developers to easily add SBML support to their tools.

www.sysbio.org This is the main SBW (Systems Biology Workbench) site. The latest versions for SBW, developer documentation, example models, screen shots, and user guides can be obtained from this site. A link to the main source-forge site is given, where all source code for SBW is made available.

www.biospice.org This is the main BioSPICE site. This site includes a description of BioSPICE and the large number of tools now available for the BioSPICE dashboard (including SBW itself).

REFERENCES

- Abbas, A. (2004). *Grid Computing: A Practical Guide to Technology and Applications*. Hingham, MA: Charles River Media.
- Bakker, B. M., Michels, P. A. M., Opperdoes, F. R., and Westerhoff, H. V. (1997). Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. *J. Biol. Chem.* **272**:3207–3215.
- Barabasi A. L., and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev Genet.* **5**(2):101–113.
- Chance, B. (1943). The kinetics of the enzyme substrate compound of peroxidase. *J. Biol. Chem.* **151**:553–577.
- Cook D. L., Farley, J. F., and Tapscott, S. J. (2001). A basis for a visual language for describing, archiving, and analyzing functional models of complex biological systems. *Genome Biol.* **2**(4):110.

- deJong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *J. Comp. Biol.* **9**:67–103.
- Fell D. (1997). *Understanding the Control of Metabolism*. London: Portland Press.
- Funahashi, A., Tanimura, N., Morohashi, M. H., and Kitano, H (2003). CellDesigner: A process diagram editor for gene regulatory and biochemical networks. *BIOSILICO* **1**:159–162.
- Garfinkel, D. (1968). A machine-independent language for the simulation of complex chemical and biochemical systems. *Comput. Biomed. Res.* **2**:31–44.
- Garfinkel, D., Rutledge, J. D., and Higgins, J. J. (1961). Simulation and analysis of biochemical systems: I. representation of chemical kinetics. *Communications of the ACM* **4**(12):559–562.
- Garny, A., Kohl, P., and Noble, D. (2003). Cellular Open Resource (COR): A public CellML-based environment for modeling biological function. *Int. J. Bif. Chaos* **13**(12):3579–3590.
- Ginkel, M., Kremling, A., Trankle, F., Gilles, E. D., and Zeitz, M. (2000). Application of the process modeling tool ProMot to the modeling of metabolic networks. In *Proceedings of the IMACS Symposium on Mathematical Modeling*, pp. 525–528.
- Harold, E. R., and Means, E. S. (2001). *XML in a Nutshell*: O'Reilly.
- Heinrich, R., and Schuster, S. (1996). *The Regulation of Cellular Systems*: Chapman and Hall.
- Higgins, J. J. (1959). Kinetic properties of sequential enzyme systems. Ph. D. thesis, University of Pennsylvania.
- Hofmeyr, J. H. S. (2001). Metabolic control analysis in a nutshell. In *Proceedings of the Second International Conference on Systems Biology*.
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., and Arkin, A. P. (2003). The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19**:524–531.
- Ingalls, B. P. (2004). A frequency domain approach to sensitivity analysis of biochemical systems. *Journal of Physical Chemistry B* **108**:1143–1152.
- Kacser, H., and Burns, J. A. (1981). The molecular basis of dominance. *Genetics* **97**:1149–1160.
- Kitano, H. (2003). A graphical notation for biochemical networks. *BIOSILICO* **1**:169–176.
- Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell.* **10**:2703–2734.
- Kohn, K. W., Aladjem, M. I., Pasa, S., Parodi, S., and Pommier, Y. (2004). *Cell Cycle Control: Molecular Interaction Map*. London: Nature Publishing Group.
- Lloyd, C. M., Halstead, M. D., and Nielsen, P. F. (2004). CellML: Its future, present, and past. *Prog. Biophys. Mol. Biol.* **85**:433–450.
- Loew, L. M., and Schaff, J. C. (2001). The Virtual Cell: A software environment for computational cell biology. *Trends Biotechnol.* **19**:401–406.
- Mendes, P. (1993). GEPASI: A software package for modeling the dynamics, steady states, and control of biochemical and other systems. *Comput. Applic. Biosci.* **9**:563–571.
- Olivier, B. G., Rohwer, J. M., and Hofmeyr, J. H. (2005). Modeling cellular systems with PySCeS. *Bioinformatics* **21**:560–561.
- Rao, C. V., Sauro, H. M., and Arkin, A. P. (2004). Putting the control in metabolic control analysis. *DYCOPS*
- Reich, J. G., and Selkov, E. E. (1981). *Energy Metabolism of the Cell*. London: Academic Press.
- Sauro, H. M. (1993). SCAMP: A general-purpose simulator and metabolic control analysis program. *CABIOS* **9**(4):441–450.

- Sauro, H. M. (2000). Jarnac: A system for interactive metabolic analysis. In (J. H. S. Hofmeyr, J. M. Rohwer, and J. L. Snoep eds.), *Animating the Cellular Map: Proceedings of the 9th International Meeting on BioThermoKinetics*. Stellenbosch University Press.
- Sauro, H. M., and Fell, D. A. (1991). SCAMP: A metabolic simulator and control analysis program. *Mathl. Comput. Modeling* **15**:15–28.
- Sauro, H. M., Hucka, M., Finney, A., Wellock, C., Bolouri, H., and Doyle, J. (2003). Next generation simulation tools: The Systems Biology Workbench and BioSPICE integration. *OMICS* **7**(4):355–372.
- Sauro, H. M., and Ingalls, B. (2004). Conservation analysis in biochemical networks: Computational issues for software writers. *Biophys. Chem.* **109**:115.
- Sauro, H. M., and Kholodenko, B. N. (2004). Quantitative analysis of signaling networks. *Prog. Biophys. Mol. Biol.* **86**(1):543.
- Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F. S. K., and Tanida, S. (1999). ECELL: Software environment for whole-cell simulation. *Bioinformatics* **15**:72–84.
- Tyson, J. J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. *Nature* **2**:908–916.
- Tyson, J. J., CsikaszNagy, A., and Novak, B. (2002). The dynamics of cell cycle regulation. *BioEssays* **24**:1095–1109.
- Wright, S. (1929). Fisher's theory of dominance. *The American Naturalist* **63**:274–279.
- Wright, S. (1934). Physiological and evolutionary theories of dominance. *The American Naturalist* **68**:24–53.
- Yun, H., Lee, D., Lee, S., Jeong, J., and Lee, S. (2004). MFAML: An XML-based standard format for metabolic flux analysis.

Introduction to Computational Models of Biochemical Reaction Networks

**Frank J. Bruggeman, Barbara M. Bakker*,
Jorrit J. Hornberg**, and Hans V. Westerhoff†**

Molecular Cell Physiology, Faculty of Earth and Life Sciences, Biocentrum Amsterdam, Vrije Universiteit, Amsterdam, The Netherlands.

**Molecular Cell Physiology, Faculty of Earth and Life Sciences, Biocentrum Amsterdam, Vrije Universiteit, Amsterdam, The Netherlands.*

***Molecular Cell Physiology, Faculty of Earth and Life Sciences, Biocentrum Amsterdam, Vrije Universiteit, Amsterdam, The Netherlands.*

†Molecular Cell Physiology, Faculty of Earth and Life Sciences, Biocentrum Amsterdam, Vrije Universiteit, Amsterdam, The Netherlands, and Mathematical Biochemistry, Swammerdam Institute for Life Sciences, Biocentrum Amsterdam, Universiteit van Amsterdam, The Netherlands.

ABSTRACT

Cellular and molecular biology have led to the understanding that cells resemble highly-organized spatiotemporal biochemical reaction networks composed of interacting gene networks, metabolic networks, and signaling networks. Within such networks many types of molecular processes take place on a wide range of time scales. These include diffusive and mediated (active and passive) transport, complex formation/dissociation, and chemical conversions. Systems biology aims at understanding the functioning of cells that stems from the interactions between their constituent (macro) molecules. Its research typically combines experiment, theory, and computation to analyze cellular behavior. This chapter deals with the computational and theoretical components of systems biology research. It gives an overview of the methods available to (1) analyze structural, regulatory, and kinetic models of the networks, (2) simulate the behavior of the networks in kinetic models, and (3) perform metabolic control analysis of these kinetic models.

I. INTRODUCTION

In the last few decades molecular biology has been especially successful in elucidating the basic molecular-network nature of all life. Organisms have proven to be sophisticated molecular systems composed of large numbers of molecules that interact with a high level of specificity, forming a huge biochemical reaction network organized in space and time. Nearly all cellular reactions are catalyzed by dedicated enzymes and the amino acid sequences of those proteins are all encoded by the cell's genome. The puzzle that remains to be solved by the biosciences is the elucidation of how molecules jointly bring about cellular behavior.

Molecular biology and genetics have led to the identification of the proteins and control mechanisms that make up living cells (e.g. Watson and Crick 1953; Pardee and Yates 1956; Umbarger 1956; Monod 1966). Developments in nonequilibrium thermodynamics (Glansdorff 1971; Westerhoff, 1987) and dynamical systems theory (Nicolis, 1977; Guckenheimer 1983) led to the consideration that the nonlinear dependency of biochemical rates on the concentrations of substrates, products, and effectors (which becomes evident for systems displaced from thermodynamic equilibrium) complicates reductive research aiming to unravel causes in complicated networks (Westerhoff and Palsson 2004).

An example of such a nonlinear relationship is the Michaelis—Menten rate equation that relates the rate of an enzyme v to the concentration of its substrate S and product P (i.e., as $v = V_{\text{MAX}} * (S/K_S)/(1 + S/K_S + P/K_P)$, with V_{max} the maximal rate of the enzyme and K_S and K_P the affinity of the enzyme for its substrate and product, respectively). One of the corollaries of this nonlinearity for the functioning of biochemical reaction networks is that there need not be a single enzyme (the rate-limiting step) solely in control of some flux (or other systemic property) in a biochemical network (Kacser and Burns 1973; Heinrich and Rapoport, 1974; Groen et al. 1982). In contrast, many processes may contribute in varying degrees to cellular phenomena such as fluxes, and certainly to the magnitudes of concentrations.

In addition, as a result of nonlinearity even small networks can display complicated behavior (such as oscillations and multiple steady states) that cannot be understood without the aid of sophisticated mathematical models and tools for their analysis. Moreover, qualitative changes in behavior can take place upon changes in parameters that characterize properties of constituents and the environment of cells (Hess 1973; Teusink et al. 1996; Goldbeter 1997; Tyson et al. 2003; Boogerd et al. 2005). The main problem nonlinearity poses for reductionistic research is that systemic properties of cells (such as growth rate or a nutrient uptake flux) are not readily understandable in terms of the properties of isolated cellular components (e.g. a K_M or a V_{max} of an enzyme). In other words, the emphasis should neither lie on the system as a whole nor on its components in isolation when aiming to intervene in and explain systemic behavior. It is the (what used to be) "no-man's-land" between that matters.

The lack of emphasis on how the system as a whole emerges from its components has slowed down the molecular understanding of cells and hampered our

ability to specifically change cellular behavior. To overcome this problem, quantitative molecular measurements in the systems at work and mathematical models that explain how cellular constituents jointly bring about system behavior have become increasingly important.

The promise of systems biology is that a molecular understanding of cellular behavior can be achieved by combined computational and experimental studies on cellular networks. Computational systems biology forms an important segment of systems biology because it supports experimental studies by offering hypothesis testing capabilities, theoretical tools, and quantitative concepts (Kitano 2002; Westerhoff and Palsson 2004). Whereas mathematical and theoretical biology have largely evolved outside experimental biology, computational systems biology should aim at becoming an integral component of experimental research.

In this chapter, we will give a broad overview of the tools available for performing computational systems biology of cellular networks. We discuss (1) methods of modeling cellular behavior in terms of signaling, metabolism, and gene expression, (2) models incorporating experimentally determined physicochemical and kinetic properties (silicon cells), and (3) metabolic control analysis as a tool to analyze control and regulation of cellular systems. We give illustrative examples of the model descriptions and analysis methods wherever possible.

II. ANALYSIS OF STRUCTURAL, REGULATORY, AND KINETIC MODELS

Many aspects of the structure and dynamics of biochemical reaction networks have been analyzed in the last decades with methods that can now be considered part of computational systems biology. Here we shall briefly give an overview of model descriptions and some of the methods for model analysis (Figure 7.1).

A. Structural models

At the lowest level of detail, we distinguish the stoichiometric structure of a biochemical reaction network. It is a description of all biochemical conversions that take place in the network (e.g., of catalysis, transport, and binding). It represents the topology of mass flow through the network. It identifies all substrates and products for all processes in the network; it does not incorporate inhibitory or activatory effects of allosteric effectors. Frequently, the end result is represented in terms of a stoichiometry matrix \mathbf{N} —the *stoichiometric model* (Figure 7.1). For a network with m intermediates and r reactions, the i,j -th entry of \mathbf{N} (n_{ij}) gives the number of moles of the i -th intermediate s_i of the network produced ($n_{ij} > 0$) or consumed ($n_{ij} < 0$) in the j -th reaction v_j of the network. Figure 7.2a shows an example of a biochemical reaction network, and Figure 7.2b displays its corresponding stoichiometric matrix.

In a biochemical reaction network, the rate of change of the concentration of the i -th intermediate s_i is given by the sum of the production and consumption rates

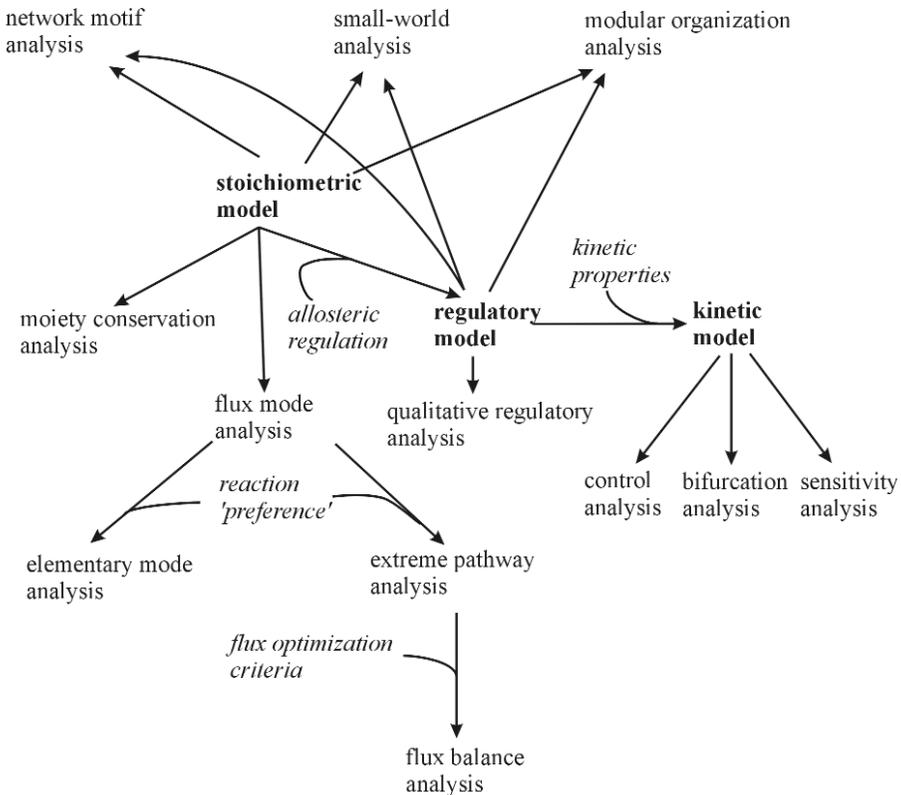


Figure 7.1. Biochemical reaction networks can be modeled with structural, regulatory, and kinetic models (shown in bold). These models can be analyzed with different methods (shown in normal font). Instances of the addition of new information are given in italic font.

of that intermediate; that is, $ds_i/dt = \sum_{j=1}^r n_{ij}v_j$. In a steady state, the rates of change of all intermediates are defined as zero (i.e., $\mathbf{Nj} = \mathbf{0}$, with the fluxes \mathbf{j} as the rates \mathbf{v} at steady state). In flux mode analysis, the fluxes \mathbf{j} are related to a particular property of the stoichiometric matrix (i.e., to its null space or kernel). Our interest is to obtain nontrivial solutions (i.e., $\mathbf{j} \neq \mathbf{0}$) for the fluxes from $\mathbf{Nj} = \mathbf{0}$. For this to be possible, the stoichiometric matrix \mathbf{N} should be singular, which is expressed by $\mathbf{NK} = \mathbf{0}$, with \mathbf{K} as the kernel or null space of \mathbf{N} . Each column of \mathbf{K} represents a relative flux distribution, a so-called flux mode that satisfies the steady-state condition (Schuster et al. 1999).

To each such flux mode one independent flux can be assigned. Its value is sufficient to recover all other flux values that occur in the same flux mode. Any steady-state flux distribution can be written as a linear combination of the flux modes in the system (Reder 1988; Schuster et al. 1999). For the example network displayed in Figure 7.2a, the \mathbf{K} matrix is displayed in Figure 7.2c. It identifies two independ-

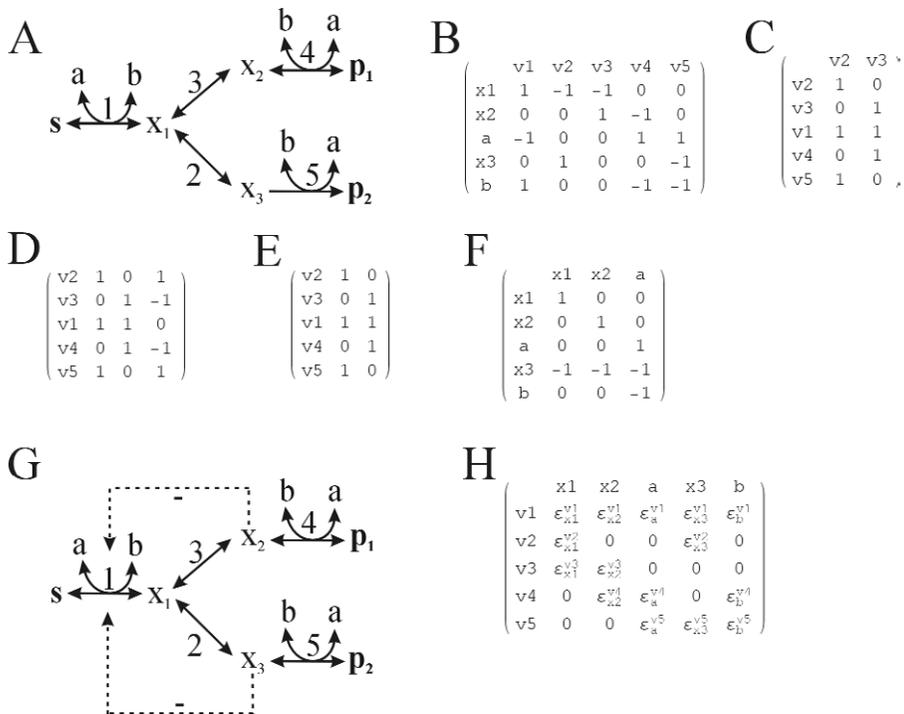


Figure 7.2. (a) Structural model of the network considered with five reactions (only reaction 5 is irreversible) and five metabolites: (b) stoichiometric matrix, (c) kernel matrix, (e) extreme pathway matrix (identical to **K**), (d) elementary mode matrix (different from **K** and extreme pathway matrix), (f) link matrix, (g) regulatory model, and (h) elasticity matrix. Metatool (Pfeiffer et al. 1999) was used to perform the structural analysis. Note that reaction 5 was chosen irreversible.

ent fluxes in the network. The values of these two fluxes suffice to describe all fluxes in the network. In this case, reactions 2 and 3 have been chosen as independent fluxes. Alternatively, reactions 4 and 5 could have been chosen. This analysis does not describe the actual magnitudes of the fluxes. The magnitudes of j_2 and j_3 still depend on the kinetics of all system components and the values for the boundary metabolites, but this is not considered in flux analysis. Analysis of steady-state fluxes in metabolic networks can also be used to determine unknown intracellular fluxes from experimentally measured import and export fluxes (Stephanopoulos 1998).

The drawback of the use of the kernel matrix is that it is not a unique description of the flux modes in the system. This is solved both by elementary flux mode analysis (EFMA) and by extreme pathway analysis (EPA) (Schuster et al. 2000; Schilling et al. 2000b), but in different ways. For a comparison of both methods and a historical overview, we refer the reader to Papin et al. (2004). Elementary flux modes are minimal sets of enzyme rates that can each generate valid steady states with all irreversible reactions proceeding in the direction suggested thermodynamically. In addition to the **N** matrix, the preferred directions of the reaction (on

the basis of its equilibrium constant) are considered (Figure 7.1). An elementary flux mode is elementary if it is non-decomposable (Schuster et al. 1999).

Any steady-state flux pattern can be expressed as a non-negative linear combination of the elementary flux modes (Schuster et al. 2000). To find the *smallest* number of flux modes that are unique, EPA can be used (Papin et al. 2004). It differs from EFMA in that it decomposes each import (and each net export) reaction into a forward and backward reaction and treats those as separate reactions. The elementary modes and the extreme pathways of the example network are shown in Figures 7.2d and 7.2e, respectively. Note that the kernel matrix and the elementary mode matrix are both a linear combination of the columns of the extreme pathway matrix. Neither EPA nor EFMA leads to the actual magnitudes of the fluxes. They merely show which flux distributions through all the steps in the network should be consistent with steady state.

For the simplest example of a linear pathway, the two methods only state that at steady states all enzyme rates should be equal, and that irreversible reactions should operate in their thermodynamically preferred direction. This is trivial for very simple networks, but for the complex and extensive networks encoded by entire genomes it is difficult to guess these steady-state solutions by simple inspection, and this is where the two systematic methods come in. They are useful when it turns out that with the available genomic information there is no elementary flux mode that leads to the production of an important compound, or indeed to biomass (growth).

The analyses also enable one to calculate yields of production of certain substances through elementary modes, and show that different such modes have different such yields (e.g., amount of ATP needed to produce a mole of product, often erroneously presented as “efficiencies”). In biotechnology, this can suggest genetic or other manipulation of the organism such that the modes with the higher yields prevail. To *understand* actual flux patterns in mechanistic terms, additional information on the regulation of gene expression and the regulation of enzyme rates *in situ* is required.

A combination of flux mode analysis and experimental determination of input and output fluxes to the systems (such as measured rates of nutrient consumption and biomass production) can lead to knowledge concerning magnitudes of network fluxes at steady state (Stephanopoulos 1998). Degeneracy of this problem due to the existence of a great many flux patterns that lead to the same input and output fluxes can be resolved by advanced isotope experiments, provided metabolic channeling is absent (Wiechert and Wurzel 2001; Isermann and Wiechert 2003).

And then there is the *teleologic* meaning of the “understanding” of flux patterns. This analyzes what flux patterns would correspond to certain “purposes,” or to certain optimalities. Flux balance analysis (FBA), largely used with EPA, determines flux distributions that are optimal relative to some criterion (Schilling et al. 2000a) (e.g., maximal growth rate). In addition to the \mathbf{N} matrix and the thermodynamic preferred direction of the reaction, an optimality criterion should be supplied to FBA (Figure 7.1). Because of the linear structure of the description, linear optimality criteria are met by extreme flux distributions. FBA makes use of linear program-

ming to determine the linear combination of extreme pathways that achieve some optimal network function.

Because FBA must assume the network to be optimal with respect to some functions, the method is speculative from the “mechanistic-explanation” point of view. This is because it is not at all understood whether biochemical reaction networks are indeed optimal for some readily understandable functionality, such as growth rate or efficiency. Indeed, neither efficiency nor growth rate seem to be functions that micro-organisms have been optimized for as a sole objective (Westerhoff 1987). Thus, results of FBA should always be presented as “this flux pattern should be expected if the network were optimal for the function . . .” FBA models of *E. coli* (Edwards and Palsson 2000) and *S. cerevisiae* (Forster et al. 2003) exist, and it is occasionally suggested that these two organisms are thereby understood. From the point of view of mechanistic understanding, which has been the most powerful force behind modern science, these suggestions are far from the truth.

We do not know how the organism achieves those flux patterns upon changes in nutrient levels. The organism should up- or down-regulate the activity of enzymes by metabolic, signaling, and genetic regulation, and the manner in which it does this is not given by FBA but would need other sophisticated analysis. Such type of analysis—for instance, with kinetic models (see material following)—would amount to an understanding of the mechanisms at work in the cell. From the teleological point of view, these suggestions may prove incorrect. The FBA models have so far merely calculated what should be optimal flux patterns, but there is no (complete) experimental validation that these flux actually occur in these organisms. On the contrary, the long-standing observation of quite significant growth-rate-dependent and growth-rate-independent maintenance metabolism in both organisms would by themselves invalidate these suggestions.

This is not to say that the FBA models are irrelevant or wrong. Quite on the contrary: (1) they open up an avenue toward a much more profound analysis of a long-standing issue (i.e., why microorganisms are not optimally efficient in energetic terms), (2) they allow for the prediction of intracellular flux patterns from a structural model alone, and (3) they are among the first tools that can analyze system-wide structural information of biochemical reaction networks.

Flux analyses show that not all flux patterns can be obtained at steady state. Similarly, the stoichiometric structure of the network limits the magnitudes of the concentrations many intracellular metabolites can obtain through metabolic regulation only. The concentration of NADH, for instance, cannot exceed the total concentration of NADH plus NAD at zero time, when only reactions interconverting the two exist in the network. Only one of the two is therefore dynamically independent, the other being dependent. Such dependencies between concentrations of molecular intermediates can be identified in the network by moiety conservation analysis.

What the kernel \mathbf{K} achieves for fluxes the link matrix \mathbf{L} achieves for metabolites. The latter relates the rates of change of metabolite concentrations one might wish to consider as independent dynamic variables (their concentration we will confine to the vector \mathbf{x}) to the rates of change of the dependent intermediates

(confined to \mathbf{x}^d); thus, $d\mathbf{s}/dt \equiv (d\mathbf{x}^i/dt, d\mathbf{x}^d/dt)^T = Ld\mathbf{x}^i/dt$ (Reder 1988; Heinrich 1996; Cornish-Bowden and Hofmeyr 2002). The identification of the conservation of moieties and their totals (confined to vector \mathbf{t}) can be identified from the relationship $[-L_0, I][\mathbf{x}^i, \mathbf{x}^d] = \mathbf{t}$.

The link matrix of the example biochemical network is displayed in Figure 7.2f. It shows that there exist three independent intermediates in the network and therefore two moiety conservation relationships ($a + x_1 + x_2 + x_3 = \text{constant}$; $a + b = \text{constant}$). Alternative methods exist that analyze systems in terms of more relevant independently variable properties (such as ratios of concentrations), which then map onto redox potentials and free energy differences, especially when coenzymes such as NAD and ADP are involved (Westerhoff and Van Dam 1987). For Figure 7.2, this could mean that a/b is chosen as one of independent dynamic variables.

So far, the methods all dealt exclusively with the mass-flow structure, and in case of EFMA and EPA with the thermodynamically preferred direction of the reactions. Actual fluxes could not be calculated (predicted) by EFMA or by EPA, only possible relative steady-state flux patterns. (Only if input and output values of fluxes and an optimality criterion were supplied could the magnitudes of fluxes be predicted with FBA.) This is understandable, as information about expression levels and kinetic properties of the system's catalysts was not used in these methods.

Indeed, this is sometimes proclaimed to be an advantage of the flux mode methods. That is, the information most difficult to obtain is not necessary for those methods. Of course, the other side of the coin is that if (according to the leitmotif of systems biology) functional properties arise in the dynamic interactions of the molecules, how much of the essence of systems biology could one expect to discover if one refrains from using the information about the dynamic nonlinear interactions between the molecules? If systems biology is about obtaining an understanding of how molecules jointly bring about cellular behavior, flux analysis does not by itself suffice to obtain an understanding of the molecular mechanisms at work in the cell.

The analysis of the type of organization of stoichiometric models of biochemical networks (small-world analysis) has focused on how the distance between two intermediates (nodes) in a biochemical reaction network chosen at random—measured as the number of reactions that need to be traversed in order to go from one node to the other—are scaled with the size of the network (Jeong et al. 2000; Wagner and Fell 2001; Barabasi and Oltvai 2004). It turned out that the intracellular biochemical networks have few nodes with a very high number of edges, where each reaction is considered an edge. In addition, they have short average path lengths between nodes (just like many random graphs have), and in some cases they display a high level of clustering (which is not found in random graphs) (Newman 2003). This is the basis for structural models to be termed small-world networks.

B. Regulatory models

Additional consideration of the dynamic interactions of substrates and products with their enzymes, of allosteric effectors with those enzymes, of transcription

factors with the DNA, of kinases with signal transduction proteins, and of feed-forward and feedback loops that convey regulatory influences rather than mass flow without forfeiting the stoichiometric structure leads to a description of all interactions in the biochemical reaction network (the *regulatory model*). This type of regulatory model leads to particularly strong results, and is one of the main objects of study in metabolic control analysis. This structure can be described in terms of the $r \times m$ elasticity matrix ϵ (defined as $\partial \ln v / \partial \ln s$) (Kacser and Burns 1973; Reder 1988). An i, j -th entry of this matrix gives the fractional sensitivity of the i -th rate v_i to the j -th intermediate s_j of the network achieved by substrate, product, or effector effects. The regulatory model of the example network and its description in terms of the elasticity matrix is shown in Figures 7.2g and 7.2h, respectively. Analysis of this matrix in a qualitative manner, by only considering the locations and the signs of the entries, allows qualitative analysis of the control distribution within the network (Hofmeyr 1989; Schuster and Schuster 1992; Teusink and Westerhoff 2000). In addition, it allows for the search for network motifs with network motif analysis (Milo et al. 2002; Shen-Orr et al. 2002). The rationale behind such a search is that biochemical reaction networks may contain network structures that appear more frequently than anticipated at random, and may therefore reflect a recurrent functional topology.

Most of these structural and regulatory analyses considered the intracellular biochemical networks as unweighted (directed or undirected) graphs, with the molecular species as nodes and the chemical conversions as the only edges. The limitation of these methods is that the strengths of the interactions, the actual magnitude of the flux through individual reactions, the concentrations of intermediates, and (perhaps worst of all) the allosteric interactions known to be crucial to the regulation of intracellular biochemistry are not considered. The essence of biochemical reaction networks (i.e., the nonlinear dynamic interactions) have not been taken into account in these network analyses, even though those should perhaps be the focus of systems biology.

An exception is the analysis of gene regulatory networks by Shen-Orr et al. (2002), which was largely based on a map of allosteric interactions. This led to an interesting set of regulatory motifs that were more dominant than others. Here, the analysis is still incomplete, precisely because the majority of regulatory interactions (i.e., those running through metabolism and signal transduction) were missing from the analysis. This should not be the case for complete kinetic models of biochemical reaction networks, but this type of model description (although much more definitive) is far more demanding in terms of experimental information.

C. Kinetic models

Incorporation of the kinetic properties of the processes and the total concentrations of moieties present in the network gives a *kinetic model* description of the network. This amounts to the characterization (parameterization) of all rate equations of all processes in the network, which then amounts to the determination of the type of function and the parameters of the rates $v_i = v_i(\mathbf{s}, \mathbf{p})$ for $i = 1 \dots r$ for all

reactions of the network as functions of all intermediate concentrations—not only those that are stoichiometrically adjacent. Such rate equations can be (ir-) reversible Michaelis–Menten; of more complicated type, such as ordered, sequential, or random mechanisms (Segel 1993); or for multi-subunit enzymes of cooperative mechanisms (Monod et al. 1965; Koshland et al. 1966).

Integrating with the stoichiometric models, the kinetic models ultimately have the description $ds/dt = \mathbf{N}v(\mathbf{x}(t, \mathbf{p}), \mathbf{x}^d(\mathbf{x}), \mathbf{p})$, where all parameters (kinetic properties, moiety-conserved totals, and boundary conditions) have been given an (experimentally) determined value (Chance et al. 1960; Bakker et al. 1997; Teusink et al. 2000). An example kinetic model (i.e., the kinetic model description of the network of Figure 7.2) can be downloaded from www.systemsbiology.net/compsysbiolbook/ in Jdesigner, Gepasi, or SBML format. Alternatively, it can be simulated online using the JWS online server of Stellenbosch University (where the model is available in the demo model section).

With these kinetic models, temporal profiles of the concentrations of intermediates and rates in the network can be calculated and compared to experimental flux analysis or X-omics data. In a promising approach to systems biology called the “silicon cell,” these kinetic models are solely based on experimentally determined kinetic properties, mechanisms, and interactions of the molecular components of the system. This makes them computer “replicas” of the real system, with the important tenet that any system property emanating from the nonlinear dynamic interaction of the components of the system should be calculable in such a computer replica (Bakker et al. 1997; Rohwer et al. 2000; Teusink et al. 2000; Hoefnagel et al. 2002; Bruggeman et al. 2005).

Models that incorporate phenomenological descriptions of processes and interactions, fitted to system rather than to component behavior, are referred to as core models. They can only function to illustrate the possible behavior of simplified networks (e.g. Selkov 1981; Teusink et al. 1998; Tyson et al. 2003). They cannot be subjected to validation in contrast to silicon cell models; compare the core models of glycolysis Goldbeter and Lefever (1972) and Selkov (1975) to the silicon cell models of glycolysis (Teusink et al. 2000; Hynne et al. 2001). Silicon cells (but not core models) can be used for purposes such as prediction of improved product formation (Hoefnagel et al. 2002), drug design (Bakker et al. 2000; Boros et al. 2002), and rigorous testing of proposed biochemical mechanisms (see the testing of oscillophoretic mechanisms in glycolysis by Reijenga et al. (2005b); the analysis of ammonium assimilation in *Escherichia coli* (Bruggeman et al. 2005)).

Kinetic models can be analyzed in many ways. Bifurcation analysis allows for the evaluation of changes in qualitative dynamics of systems as a function of one or more parameters; for example, changes from stable steady state (fixed points) to instable steady state showing oscillations (emergence of a limit cycle; Hopf bifurcation), to bistability (saddle node bifurcation), or to chaos (Goldbeter 1997; Heinrich 1996). Bifurcation analysis has been combined with quantitative experimentation (Hynne et al. 2001; Reijenga et al. 2002; Reijenga et al. 2005a). Sensitivity analysis (or parameter sensitivity analysis) can be used to determine the identity

of parameters to which a particular cellular phenomenon of interest is most sensitive, but only sensibly so in silicon-cell-type models (Saltelli et al. 2000). Such sensitivity analysis is an important tool in the analysis of kinetic models. When combined with an assessment of experimental error, it can be used to decide to what extent model predictions can be trusted.

III. SIMULATION METHODS FOR KINETIC MODELS

We will now overview the simulation methods most frequently used to perform calculations with kinetic models. Kinetic models can be evaluated with different description methods, as shown in Figure 7.3. The choice of the theoretical description depends on the nature of the experimental data available (spatiotemporal versus temporal, single-cell versus population resolution), the physicochemical characteristics of the process (reactions between macromolecules or small molecules; the copy numbers of the (macro) molecules; diffusive properties of molecules), and the type of question (quantitative or qualitative) of interest.

The complexity of the cell's interior (Goodsell 1991) illustrates that in principle many different simulation methods may apply, depending on the cellular phenomenon of interest. The high concentrations of the many different macromolecules present may give rise to macromolecular crowding, which can potentially introduce diffusion gradients (lead to channeling) and alter kinetic and physicochemical properties (Zimmerman and Minton 1993; Ovadi 1995; Brown and Kholodenko 1999; Elowitz et al. 1999; Ellis 2001). Some concentrations of (macro) molecules are so low within cells that stochastic fluctuations may affect systemic behavior (Arkin et al. 1998; Isaacs et al. 2003). Even at high copy numbers stochastic phenomena may become important when elasticities are small (Elf et al. 2003).

At the highest level of detail, such as single-molecule (or single-atom) resolution, microscopic modeling procedures can in principle be applied to study dynamic phenomena of biochemical reaction networks (Baras and Mansour 1996; Gorecki 1999). However, such methods are computationally too intensive at present to give results that apply at the level of a biochemical reaction network (for an application to membranes, see Mouritsen and Jorgensen (1997)). At present, systems up to 10^5 atoms/molecules on time scales up to nanoseconds can be simulated (Kaan-dorp; personal communication), and it will never be possible to simulate a complete living cell in complete single-molecule detail, just like it will never become possible to calculate the dynamic structure of a large protein in complete atomic detail without mesoscopic simplifications.

Even if it were feasible, such simulation in complete molecular detail foregoes many of the spectacular advances in statistical mechanics, leading to the realization of importances of ensemble averages and to higher-order concepts such as irreversibility and the second law of thermodynamics. These higher-order concepts and principles are in fact examples of systems theory principles, and many of the essential issues posed to biology (including "order out of chaos," symmetry break-

ing, hysteresis, evolutionary optimization, and key-lock fitting of substrates into enzymes) depend on such averaging and transition to higher-order concepts.

Indeed, we consider full-blown microscopic simulation of systems without first demonstrating the need for such detailed modeling, a detraction from scientific progress. Needless to say, we do advocate microscopic modeling for those cases where relevant molecule numbers are so small, interactions so nonlinear, and the issues so important that it really matters. Microscopic simulation is shown in gray in Figure 7.3 to indicate that these methods are not yet widely used in computational systems biology.

One method that is more feasible at present (and which has received a lot of attention) is a mesoscopic level of description, which originates from statistical physics (Keizer 1987; Van Kampen 1993). (Alternative methods deal with particle-based modeling or cellular automata, which will not be discussed here further (Weimar 1997)). This method uses the master equation description of the kinetic model, which describes the evolution of the probability density function of the state of the network (i.e., the rate of change of the probability that a biochemical network has a particular spatiotemporal state at a particular time). This amounts to a stochastic description for the reaction-diffusion processes that take place in the system. This description explicitly considers the effects of local (thermal) fluctua-

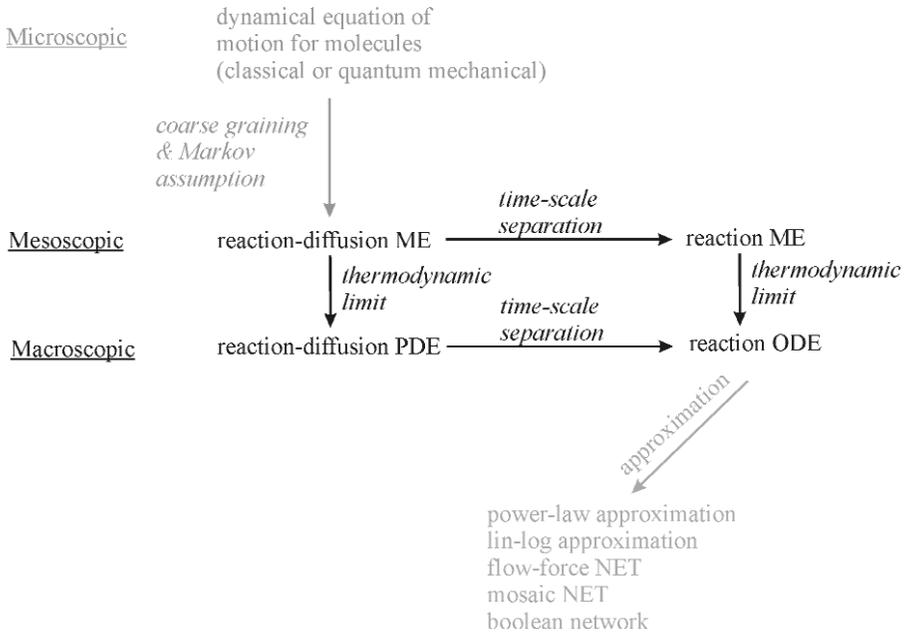


Figure 7.3. Overview of simulation methods for kinetic models of biochemical reactions networks. Three levels of simulation are depicted. We focus on mesoscopic and macroscopic simulation methods, which is why microscopic is shown in gray font. ME refers to “master equation,” PDE to “partial differential equation,” and ODE to “ordinary differential equation.”

tions on the reaction and diffusion rates that take place in the system. It is frequently assumed, however, that diffusion gradients are dissipated (equilibrated) before changes in concentrations can arise due to reactions (for an exception, see Baras and Mansour (1996) and Elf (2005)). If this does hold true, the stochastic description resembles a master equation that only refers to reaction processes and not to diffusion processes (Van Kampen 1993).

The pure-reaction description neglecting spatial heterogeneities has received much attention (Gillespie 1976; Arkin et al. 1998; Elowitz and Leibler 2000; Turner et al. 2004). In the so-called small-noise or linear-noise approximation (LNA), the evolution of the average of the probability distribution as described by the master equation for reaction-diffusion and pure reactive systems becomes (respectively) equal to a description in terms of partial differential equations (PDEs, reaction and diffusion processes) and ordinary differential equations (ODEs, only reaction processes), which are far more amenable to theoretical and numerical analysis (Keizer 1987; Van Kampen 1993; Elf and Ehrenberg 2003).

The importance of the LNA theory is that with the results obtained with PDE and ODE simulations the size of stochastic fluctuations can be approximated up to first order (Keizer 1987; Van Kampen 1993; Elf and Ehrenberg 2003). In other words, information on the mesoscopic stochastic behavior of the network can be obtained with LNA from the macroscopic behavior of the network alone. This makes LNA a potentially useful tool in the analysis of the dynamics of signaling and genetic networks in cases where differential equations are used, as has been done so frequently in the past, and a first-order guess of the stochastic behavior of the systems is needed.

When the kinetic parameters of processes within the biochemical reaction network have not been determined experimentally, they have to be estimated or fitted, or the kinetic model description has to be simplified. For systems without diffusion limitation, a reasonable number of approximate methods have been developed to deal with the simplification of the kinetic model to cope with the problem of parameter uncertainty. Biochemical systems theory (BST), S-systems, and power-law approximations have been specifically designed to deal with a simplified description of biochemical reaction networks that should overcome the analytical problems introduced by the nonlinearity of rate equations and deal with the problem of unknown kinetic parameters (Savageau 1976; Voit 2000).

The *linlog* approximation (Visser and Heijnen 2002) champions the flow-force relationships derived by mosaic non-equilibrium thermodynamics (flow-force relationships (Westerhoff 1987)) and offers, as does BST, many analytical possibilities (Visser and Heijnen 2002). Mosaic non-equilibrium thermodynamics (MNET) introduced mechanistic information into flow-force relationships of irreversible thermodynamics, and has been successfully applied to the analysis of mitochondrial oxidative phosphorylation, bacterial growth, and ion transport in bacteriorhodopsin liposomes (Westerhoff 1987). All of these approximate methods suffer from their qualitative nature, but this depends to a large extent on the question one wants to answer. For example, BST has proven very powerful in the analysis of different

biochemical reaction network topologies regarding their functional properties (Savageau 1991, 2001; Alves and Savageau 2003).

IV. ANALYSIS OF REGULATION AND CONTROL OF SYSTEMIC PROPERTIES OF BIOCHEMICAL REACTION NETWORKS

A theoretical framework frequently used in experimental and theoretical studies on control and regulation of biochemical reaction networks is metabolic control analysis (MCA). MCA provides a tool to calculate to what extent any systemic property of those biochemical reaction networks (e.g., a flux, a concentration, or any function thereof) is controlled by the activities of the processes (e.g., by the activity of an enzyme). It was pioneered in the 1970s by Kacser and Burns (1973) and Rapoport and Heinrich (1974). Later it was generalized by Kell (1986) and Reder (1988).

The theory continues to be extended to deal with different aspects of biochemical reaction networks. Whereas initially MCA dealt solely with control of steady-state fluxes and concentrations of metabolic networks, it has since been extended to encompass control of other variables such as Gibbs free energies, efficiencies, flux ratios (Westerhoff and Van Dam 1987), generalized variables (Schuster 1996), and transition times (Melendezhevia et al. 1990).

It has been developed also for systems involving quasi-equilibrium reactions and time-scale separation (Delgado and Liao 1995; Kholodenko et al. 1998); to address the control of frequencies and amplitudes of oscillatory systems (Kholodenko et al. 1996, 1997a; Ingalls 2004a, 2004b); the statics of signaling networks (Kahn and Westerhoff 1991; Bruggeman et al. 2002); the dynamics of signaling networks (Kholodenko et al. 1997b; Hornberg et al. 2005); channeling (Kholodenko et al. 1994); intra-enzymatic processes (Kholodenko and Westerhoff 1994); hierarchical networks with gene expression, signal transduction, and metabolism (Kahn and Westerhoff 1991; Hofmeyr and Westerhoff 2001; Bruggeman et al. 2002; Hornberg et al. 2005); modular networks (Schuster et al. 1993); reaction-diffusion networks (Peletier et al. 2003); and transient trajectories (Acerenza et al. 1989; Heinrich 1991; Ingalls and Sauro 2003). It has been applied frequently to the experimental analysis of cellular networks (Groen et al. 1982; Westerhoff 1987; Fell 1997; Ainscow and Brand 1999a, 1999b).

Control analysis focuses on the extent of control exerted by a process with the biochemical reaction network on a particular systemic property. This is quantified by a control coefficient of the activity a_i of a reaction i on the systemic property f as $C_{v_i}^f = d \ln f / d \ln a_i$. A control coefficient is a special case of the more generally applicable response coefficient (used in sensitivity analysis), which considers any parameter $R_{p_i}^f = d \ln f / d \ln p_i$. This special case is important as it comprises the control by all catalytic process in the cell.

Response coefficients are identical to control coefficients; that is, $R_{p_i}^f = C_{v_i}^f$ if the parameter p_i only affects one process i in the network and if the rate of this process depends linearly on p_i (i.e., if $\partial \ln v_i / \partial \ln p_i = 1$). Examples of such parameters are enzyme concentrations in metabolic networks considered without gene expression.

In proper descriptions, the rate equations of the individual enzymes are parameterized for their activity by introducing a parameter a_i or λ_i as a multiplier that equals 1 at the physiological state but can be modulated independently. The modulation corresponds to a simultaneous modulation of the forward and reverse V_{\max} of the enzyme by the same factor.

Classical control analysis focuses on the control of all processes in the network on steady-state fluxes and concentrations in a biochemical reaction network. The control coefficients can be concisely written in matrix format, using the matrices introduced previously (Westerhoff et al. 1994; Kholodenko et al. 1995; Heinrich 1996).

$$\underbrace{\begin{bmatrix} C_v^j \\ C_v^{x^i} \end{bmatrix}}_C \underbrace{[\bar{K} - \varepsilon \bar{L}]_{r \times r}}_E = I \Rightarrow$$

$$CE = I \Rightarrow C = E^{-1} (\Rightarrow E = C^{-1})$$

The matrices \bar{K} and \bar{L} are the scaled kernel and link matrix, respectively. These are $\bar{K} = \text{Dg}(j)^{-1} \cdot K \cdot \text{Dg}(j^i)$ and $\bar{L} = \text{Dg}(s)^{-1} \cdot L \cdot \text{Dg}(x^i)$ (with $\text{Dg}(a)$ as a diagonal matrix with the entries of vector a as its diagonal elements). For a good introduction to the matrix formulation of metabolic control analysis, the reader is referred to Hofmeyr (2001).

What makes metabolic control analysis so attractive, and where it differs from sensitivity analysis, is the existence of summation laws and (for some types of control coefficients) the existence of connectivity laws. Summation laws hold that the sum of all control coefficients on a given property of the system is an integer (1 for flux, 0 for concentration, and -1 for frequency). Indeed, the laws reflect the behavior of a systemic property of the network if all activities in the network are increased simultaneously by the same factor. They derive from a rescaling of the time dimension of the kinetic model description of the biochemical reaction network and can be understood in terms of Euler's theorem of homogeneous functions (Westerhoff 1987; Giersch 1988; Peletier et al. 2003).

For instance, the summation theorem for the concentration control coefficients of a particular intermediate x_j at a particular time t gives the scaled rate of change of that intermediate; that is, $\sum_{i=1}^r C_{v_i}^{x_j} = d \ln x_j / d \ln t$ (which yields the more familiar summation theorem in steady state, but is then also valid for the maximum attained in a transient response such as in signal transduction) (Acerenza et al. 1989; Kholodenko et al. 1997a; Hornberg 2004). Many such summation laws have been derived and can be found in many of the references to MCA literature in this section. Connectivity theorems are related in an interesting way to the stability properties of biochemical reaction networks and express the tendency of networks prevailing in an asymptotically stable steady state to dissipate any change in concentrations of independent intermediates (Westerhoff and Van Dam 1987).

V. CONCLUSIONS

In this chapter we could only give the reader a limited overview of methods and modeling descriptions available in computational systems biology. Many topics we did not cover, such as the existing field of the elucidation of the network structure from experimental data (de la Fuente et al. 2002; Kholodenko et al. 2002; Vlad et al. 2004; Crampin et al. 2004) or of robustness of biochemical reaction networks (Kitano 2004; Stelling et al. 2004). However, this is inevitable considering the present turbulent state of (computational) systems biology.

With the growing availability of sequenced genomes, quantitative spatiotemporal data sets on the cellular state, kinetic data on cellular processes, and precise determination of network topology, we anticipate that the need for sophisticated computational systems biology will become increasingly urgent in the future. Analysis methods such as the aforementioned structural analysis methods are anticipated to have some impact on the organization of cellular networks. The kinetic methods, however, should truly deliver the systems biology, as they do incorporate the dynamic nonlinear interactions between molecules.

These methods are, however, inherently much more difficult because they require more difficult nonlinear mathematics (but above all accurate experimental kinetic information). The development of modeling descriptions and analysis tools based on modularity may continue to alleviate this otherwise Herculean task. They remain promising and exciting for the understanding of the functional organization of cellular networks. At the same time, more quantitative and standardized experimental results are needed to build detailed quantitative models of biochemical network functioning. Only in this way, it seems to us, can one test whether our knowledge is accurate and whether it is complete by using detailed models.

Such detailed models can then be analyzed to understand basic principles of cell functioning and used as predictive tools to further guide experimental research. Only in this way can we come to understand the structure and (dynamic) functioning of cells in terms of their constituent macromolecules, which is the ultimate aim of (systems) biology. Only in this way will biology become a verifiable and falsifiable science without giving up its exciting territory: life.

ACKNOWLEDGMENTS

The authors would like to thank Boris Kholodenko, Herbert Sauro, and Jannie Hofmeyr for discussions. We thank Brett Olivier and Jacky Snoep for making the kinetic models available on the JWS online web site.

INTERNET REFERENCES

Internet references accompanying this chapter can be found at www.systemsbiology.net/compsysbiolbook/, which will be continuously updated frequently.

REFERENCES

- Acerenza, L., Sauro, H. M., and Kacser, H. (1989). Control analysis of time-dependent metabolic systems. *J. Theo. Biol.* **137**(4):423–444.
- Ainscow, E. K., and Brand, M. D. (1999a). Internal regulation of ATP turnover, glycolysis and oxidative phosphorylation in rat hepatocytes. *Eur. J. Biochem.* **266**(3):737–749.
- Ainscow, E. K., and Brand, M. D. (1999b). Top-down control analysis of ATP turnover, glycolysis and oxidative phosphorylation in rat hepatocytes. *Eur. J. Biochem.* **263**(3):671–685.
- Alves, R., and Savageau, M. A. (2003). Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: Differences in molecular structure and physiological function. *Mol. Microbiol.* **48**(1):25–51.
- Arkin, A., Ross, J., and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**(4):1633–1648.
- Bakker, B. M., Michels, P. A. M., Opperdoes, F. R., and Westerhoff, H. V. (1997). Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. *J. Biol. Chem.* **272**(6):3207–3215.
- Bakker, B. M., Westerhoff, H. V., Opperdoes, F. R., and Michels, P. A. (2000). Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs. *Mol. Biochem. Parasitol.* **106**(1):1–10.
- Barabasi, A. L., and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**(2):101–113.
- Baras, F., and Mansour, M. M. (1996). Reaction-diffusion master equation: A comparison with microscopic simulations. *Physical Review* **54**(6):6139–6148.
- Boogerd, F. C., Bruggeman, F. J., Richardson, R., Stephan, S. (2005). Emergence and its place in Nature: A case study of biochemical networks. *Synthese* **145**(1):131–164.
- Boros, L. G., Cascante, M., and Lee, W. N. (2002). Metabolic profiling of cell growth and death in cancer: applications in drug discovery. *Drug Discov. Today* **7**(6):364–372.
- Brown, G. C., and Kholodenko, B. N. (1999). Spatial gradients of cellular phospho-proteins. *FEBS Lett.* **457**(3):452–454.
- Bruggeman, F. J., Westerhoff, H. V., Hoek, J. B., and Kholodenko, B. N. (2002). Modular response analysis of cellular regulatory networks. *J. Theo. Biol.* **218**(4):507–520.
- Bruggeman, F. J., Boogerd, F. C., Westerhoff, H. V. (2005). The multifarious short-term regulation of ammonium assimilation of *Escherichia coli*: dissection using an *in silico* replica. *Febs J* **272**(8):1965–1985.
- Chance, B., Garfinkel, D., Higgins, J., and Hess, B. (1960). Metabolic control mechanisms: A solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells. *J. Biol. Chem.* **235**:2426–2439.
- Cornish-Bowden, A., and Hofmeyr, J. H. (2002). The role of stoichiometric analysis in studies of metabolism: An example. *J. Theo. Biol.* **216**(2):179–191.
- Crampin, E. J., Schnell, S., McSharry, P. E. (2004). Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog Biophys Mol Biol* **86**(1):77–112.
- de la Fuente, A., Brazhnik, P., and Mendes, P. (2002). Linking the genes: Inferring quantitative gene networks from microarray data. *Trends Genet.* **18**(8):395–398.
- Delgado, J., and Liao, J. C. (1995). Control of metabolic pathways by time-scale separation. *Biosystems* **36**(1):55–70.

- Edwards, J. S., and Palsson, B. O. (2000). Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics* **1**(1):1.
- Elf, J., and Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research* **13**(11):2475–2484.
- Elf, J., and Ehrenberg, M. (2005). Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. *Systems Biology* (Accepted for publication in Issue 2).
- Elf, J., Paulsson, J., Berg, O. G., and Ehrenberg, M. (2003). Near-critical phenomena in intracellular metabolite pools. *Biophysical Journal* **84**(1):154–170.
- Ellis, R. J. (2001). Macromolecular crowding: Obvious but underappreciated. *Trends in Biochemical Sciences* **26**(10):597–604.
- Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* **403**(6767):335–338.
- Elowitz, M. B., Surette, M. G., Wolf, P. E., Stock, J. B., and Leibler, S. (1999). Protein mobility in the cytoplasm of *Escherichia coli*. *J. Bacteriol.* **181**(1):197–203.
- Fell, D. A. (1997). *Understanding the Control of Metabolism*. London/Miami: Portland Press.
- Forster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research* **13**(2):244–253.
- Giersch, C. (1988). Control analysis of metabolic networks: Homogeneous functions and the summation theorems for control coefficients. *Eur. J. Biochem.* **174**(3):509–513.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22**:403–434.
- Glansdorff, P., and Prigogine, I. (1971). *Thermodynamic Theory of Structure, Stability, and Fluctuations*. New York: John Wiley & Sons.
- Goldbeter, A. (1997). *Biochemical Oscillations and Cellular Rhythms: The Molecular Basis of Periodic and Chaotic Behavior*. Cambridge, UK: Cambridge University Press.
- Goldbeter, A., and Lefever, R. (1972). Dissipative structures for an allosteric model—Application to glycolytic oscillations. *Biophysical Journal* **12**(10):1302–1315.
- Goodsell, D. S. (1991). Inside a living cell. *Trends Biochem. Sci.* **16**(6):203–206.
- Gorecki, J., Kawczynski, A. L., and Nowakowski, B. (1999). Master equation and molecular dynamics simulations of spatiotemporal effects in a bistable chemical system. *Journal of Physical Chemistry* **103**:3200–3209.
- Groen, A. K., Wanders, R. J., Westerhoff, H. V., van der Meer, R., and Tager, J. M. (1982). Quantification of the contribution of various steps to the control of mitochondrial respiration. *J. Biol. Chem.* **257**(6):2754–2757.
- Guckenheimer, J., and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. New York: Springer-Verlag.
- Heinrich, R., and Rapoport, T. A. (1974). A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* **42**(1):89–95.
- Heinrich, R., and Reder, C. (1991). Metabolic control analysis of relaxation processes. *J. Theo. Biol.* **151**:57–61.
- Heinrich, R., and Schuster, S. (1996). *The Regulation of Cellular Systems*. New York: Chapman & Hall.
- Hess, B. (1973). Organization of glycolysis: Oscillatory and stationary control. *Symp. Soc. Exp. Biol.* **27**:105–131.
- Hoefnagel, M. H., Starrenburg, M. J., Martens, D. E., Hugenholtz, J., Kleerebezem, M., Van, S. II, Bongers, R., Westerhoff, H. V., and Snoep, J. L. (2002). Metabolic engineering of lactic

- acid bacteria, the combined approach: Kinetic modeling, metabolic control and experimental analysis. *Microbiology* **148**(4):1003–1013.
- Hofmeyr, J. H. S. (1989). Control-pattern analysis of metabolic pathways: Flux and concentration control in linear pathways. *Eur. J. Biochem.* **186**(1–2):343–354.
- Hofmeyr, J. H. S. (2000). Metabolic control analysis in a nutshell. In *Proceedings of the Second International Conference on Systems Biology*, Pasadena, California, 291–300.
- Hofmeyr, J. H. S., and Westerhoff, H. V. (2001). Building the cellular puzzle: Control in multi-level reaction networks. *J. Theo. Biol.* **208**(3):261–285.
- Hornberg, J. J., Binder, B., Bruggeman, F. J., Schoeberl, B., Heinrich, R., and Westerhoff, H. V. (2005). Control of MAPK signalling: from complexity to what really matters. *Oncogene* **24**:5533–5542.
- Hornberg, J. J., Bruggeman, F. J., Binder, B., Geest, C. R., de Vaate, A. J., Lankelma, J., Heinrich, R., and Westerhoff, H. V. (2005). Principles behind the multifarious control of signal transduction: ERK phosphorylation and kinase/phosphatase control. *FEBS J.* **272**(1): 244–258.
- Hynne, R., Dano, S., and Sorensen, P. G. (2001). Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophysical Chemistry* **94**(1–2):121–163.
- Ingalls, B. (2004a). Autonomously oscillating biochemical systems: parametric sensitivity of extrema and period. *Systems Biology* **1**:62–70.
- Ingalls, B. (2004b). A frequency domain approach to sensitivity analysis of biochemical networks. *Journal of Physical Chemistry* **108**:1143–1152.
- Ingalls, B. P., and Sauro, H. M. (2003). Sensitivity analysis of stoichiometric networks: An extension of metabolic control analysis to non-steady state trajectories. *Journal of Theoretical Biology* **222**(1):23–36.
- Isaacs, F. J., Hasty, J., Cantor, C. R., and Collins, J. J. (2003). Prediction and measurement of an autoregulatory genetic module. *Proc. Natl. Acad. Sci. USA* **100**(13):7714–7719.
- Isermann, N., and Wiechert, W. (2003). Metabolic isotopomer labeling systems. Part II: structural flux identifiability analysis. *Math Biosci.* **183**(2):175–214.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**(6804):651–654.
- Kacser, H., and Burns, J. A. (1973). The control of flux. *Symp. Soc. Exp. Biol.* **27**:65–104.
- Kahn, D., and Westerhoff, H. V. (1991). Control theory of regulatory cascades. *J. Theo. Biol.* **153**(2):255–285.
- Keizer, J. (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. Berlin: Springer-Verlag.
- Kell, D. B., and Westerhoff, H. V. (1986). Metabolic control theory: Its role in microbiology and biotechnology. *FEMS Microbiol. Rev.* **39**:305–320.
- Kholodenko, B. N., and Westerhoff, H. V. (1994). Control theory of one enzyme. *Biochim. Biophys. Acta.* **1208**(2):294–305.
- Kholodenko, B. N., Cascante, M., and Westerhoff, H. V. (1994). Control theory of metabolic channelling. *Mol. Cell Biochem.* **133**(134):313–331.
- Kholodenko, B. N., Westerhoff, H. V., Puigjaner, J., and Cascante, M. (1995). Control in channeled pathways: A matrix method calculating the enzyme control coefficients. *Biophysical Chemistry* **53**(3):247–258.
59. Kholodenko, B. N., Demin, O. V., and Westerhoff, H. V. (1996). The metabolic control theory of biochemical oscillating systems: Definitions of the quantitative characteristics and their simplest properties. *Biochemistry-Moscow* **61**(4):423–434.
- Kholodenko, B. N., Demin, O. V., and Westerhoff, H. V. (1997a). Control analysis of periodic phenomena in biological systems. *Journal of Physical Chemistry* **101**(11):2070–2081.

- Kholodenko, B. N., Hoek, J. B., Westerhoff, H. V., and Brown, G. C. (1997b). Quantification of information transfer via cellular signal transduction pathways. *FEBS Lett.* **414**(2):430–434.
- Kholodenko, B. N., Schuster, S., Garcia, J., Westerhoff, H. V., and Cascante, M. (1998). Control analysis of metabolic systems involving quasi-equilibrium reactions. *Biochim. Biophys. Acta.* **1379**(3):337–352.
- Kholodenko, B. N., Kiyatkin, A., Bruggeman, F. J., Sontag, E., Westerhoff, H. V., and Hoek, J. B. (2002). Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA* **99**(20):12841–12846.
- Kitano, H. (2002). Computational systems biology. *Nature* **420**(6912):206–210.
- Kitano, H. (2004). Biological robustness. *Nat. Rev. Genet.* **5**(11):826–837.
- Koshland, D. E. Jr., Nemethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5**(1):365–385.
- Melendezhevia, E., Torres, N. V., Sicilia, J., and Kacser, H. (1990). Control analysis of transition times in metabolic systems. *Biochemical Journal* **265**(1):195–202.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298**(5594):824–827.
- Monod, J. (1966). From enzymatic adaptation to allosteric transitions. *Science* **154**(748):475–483.
- Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**:88–118.
- Mouritsen, O. G., and Jorgensen, K. (1997). Small-scale lipid-membrane structure: Simulation versus experiment. *Curr. Opin. Struct. Biol.* **7**(4):518–527.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review* **45**:167–256.
- Nicolis, G., and Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations*. New York: John Wiley & Sons.
- Ovadi, J. (1995). *Cell Architecture and Metabolic Channeling*. New York: Springer-Verlag.
- Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends Biotechnol.* **22**(8):400–405.
- Pardee, A. B., and Yates, R. A. (1956). Control of pyrimidine biosynthesis in *Escherichia coli* by a feed-back mechanism. *J. Biol. Chem.* **221**(2):757–770.
- Peletier, M. A., Westerhoff, H. V., and Kholodenko, B. N. (2003). Control of spatially heterogeneous and time-varying cellular reaction networks: A new summation law. *Journal of Theoretical Biology* **225**(4):477–487.
- Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J. C., Montero, F., and Schuster, S. (1999). META-TOOL: For studying metabolic networks. *Bioinformatics* **15**(3):251–257.
- Reder, C. (1988). Metabolic control theory: a structural approach. *J. Theo. Biol.* **135**(2): 175–201.
- Reijenga, K. A., Bakker, B. M., van der Weijden, C. C., and Westerhoff, H. V. (2005a). Training of yeast cell dynamics. *FEBS J.* **272**(7):1616–1624.
- Reijenga, K. A., van Megen, Y. M., Kooi, B. W., Bakker, B. M., Snoep, J. L., van Verseveld, H. W., and Westerhoff, H. V. (2005b). Yeast glycolytic oscillations that are not controlled by a single oscillophore: A new definition of oscillophore strength. *J. Theo. Biol.* **232**(3):385–398.
- Reijenga, K. A., Westerhoff, H. V., Kholodenko, B. N., and Snoep, J. L. (2002). Control analysis for autonomously oscillating biochemical networks. *Biophys. J.* **82**(1/1):99–108.
- Rohwer, J. M., Meadow, N. D., Roseman, S., Westerhoff, H. V., and Postma, P. W. (2000). Understanding glucose transport by the bacterial phosphoenolpyruvate: Glycose phos-

- phototransferase system on the basis of kinetic measurements *in vitro*. *J. Biol. Chem.* **275**(45):34909–34921.
- Saltelli, A., Chan, K., Scott, E.M., eds. (2000). Sensitivity analysis. Chichester: John Wiley and Sons.
- Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. New York: Addison-Wesley.
- Savageau, M. A. (1991). Metabolite channeling: Implications for regulation of metabolism and for quantitative description of reactions *in vivo*. *J. Theo. Biol.* **152**(1):85–92.
- Savageau, M. A. (2001). Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos* **11**(1):142–159.
- Schilling, C. H., Edwards, J. S., Letscher, D., and Palsson, B. O. (2000a). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.* **71**(4):286–306.
- Schilling, C. H., Letscher, D., and Palsson, B. O. (2000b). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theo. Biol.* **203**(3):229–248.
- Schuster, S. (1996). Control analysis in terms of generalized variables characterizing metabolic systems. *Journal of Theoretical Biology* **182**(3):259–268.
- Schuster, S., and Schuster, R. (1992). Decomposition of biochemical reaction systems according to flux control insusceptibility. *Journal De Chimie Physique Et De Physico-Chimie Biologique* **89**(9):1887–1910.
- Schuster, S., Kahn, D., and Westerhoff, H. V. (1993). Modular analysis of the control of complex metabolic pathways. *Biophys. Chem.* **48**(1):1–17.
- Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology* **17**(2):53–60.
- Schuster, S., Fell, D. A., and Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**(3):326–332.
- Segel, I. H. (1993). *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-state Enzyme Systems*. New York: John Wiley & Sons.
- Selkov, E. (1975). Stabilization of energy charge, generation of oscillations and multiple steady states in energy metabolism as a result of purely stoichiometric regulation. *European Journal of Biochemistry* **59**(1):151–157.
- Selkov, E. E., and Reich J. G. (1981). *Energy Metabolism of the Cell*. London: Academic Press.
- Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**(1):64–68.
- Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. III, and Doyle, J. (2004). Robustness of cellular functions. *Cell* **118**(6):675–685.
- Stephanopoulos, G., Aristidou, A., and Nielsen, J. (1998). *Metabolic Engineering: Principles and Methodologies*. London: Academic Press.
- Teusink, B., and Westerhoff, H. V. (2000). "Slave" metabolites and enzymes: A rapid way of delineating metabolic control. *Eur. J. Biochem.* **267**(7):1889–1893.
- Teusink, B., Bakker, B. M., and Westerhoff, H. V. (1996). Control of frequency and amplitudes is shared by all enzymes in three models for yeast glycolytic oscillations. *Biochim. Biophys. Acta.* **1275**(3):204–212.
- Teusink, B., Walsh, M. C., van Dam, K., and Westerhoff, H. V. (1998). The danger of metabolic pathways with turbo design. *Trends Biochem. Sci.* **23**(5):162–169.

- Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., Walsh, M. C., Bakker, B. M., van Dam, K., Westerhoff, H. V., and Snoep, J. L. (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.
- Turner, T. E., Schnell, S., and Burrage, K. (2004). Stochastic approaches for modeling *in vivo* reactions. *Comput. Biol. Chem.* **28**(3):165–178.
- Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology* **15**(2):221–231.
- Umbarger, H. E. (1956). Evidence for a negative-feedback mechanism in the biosynthesis of isoleucine. *Science* **123**(3202):848.
- Van Kampen, N. G. (1993). *Stochastic Processes in Physics and Chemistry*. (rev. ed.). Burlington, MA: Elsevier Science.
- Visser, D., and Heijnen, J. J. (2002). The mathematics of metabolic control analysis revisited. *Metab. Eng.* **4**(2):114–123.
- Vlad, M. O., Arkin, A., and Ross, J. (2004). Response experiments for nonlinear systems with application to reaction kinetics and genetics. *Proc. Natl. Acad. Sci. USA* **101**(19):7223–7228.
- Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK: Cambridge University Press.
- Wagner, A., and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. R. Soc. Lond. B. Biol. Sci.* **268**(1478):1803–1810.
- Watson, J. D., and Crick, F. H. C. (1953). Molecular structure of nucleic acids—a structure for deoxyribose nucleic acid. *Nature* **171**(4356):737–738.
- Weimar, J. R. (1997). *Simulation with Cellular Automata*. Berlin: Logos-Verlag.
- Westerhoff, H., and Van Dam, K. (1987). *Thermodynamics and Control of Biological Free-energy Transduction*. Amsterdam: Elsevier Science Publishers B. V. (Biomedical Division)
- Westerhoff, H. V., and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nat. Biotechnol.* **22**(10):1249–1252.
- Westerhoff, H. V., Hofmeyr, J. H., and Kholodenko, B. N. (1994). Getting to the inside of cells using metabolic control analysis. *Biophys. Chem.* **50**(3):273–283.
- Wiechert, W., and Wurzel, M. (2001). Metabolic isotopomer labeling systems. Part I: global dynamic behavior. *Math Biosci.* **169**(2):173–205.
- Zimmerman, S. B., and Minton, A. P. (1993). Macromolecular crowding—Biochemical, biophysical, and physiological consequences. *Annual Review of Biophysics and Biomolecular Structure* **22**:27–65.

Biological Foundations of Signal Transduction and the Systems Biology Perspective

Ursula Klingmüller

Theodor Boveri Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

Chapter 8

ABSTRACT

Cellular communication is mediated by extracellular stimuli that bind cellular receptors and activate intracellular signaling pathways. Principal biochemical reactions used for signal transduction are transient phosphorylation of proteins or lipids, proteolytic cleavage, and degradation and formation of complexes (mediated by specific protein-to-protein interactions). Within the nucleus, signaling pathways orchestrate the activity of transcription factors and regulate gene expression. Cells differ in their competence to respond to extracellular stimuli. For example, different intracellular signaling pathways are activated in hepatocytes during the priming, proliferation, and termination stages of regeneration.

A deeper understanding of complex biological responses cannot be achieved by traditional approaches but requires the combination of experimental data with mathematical modeling. Following a systems biology approach, data-based mathematical models describing sub-modules of signaling pathways have been established. By combining computer simulations with experimental verification systems, properties of signaling pathway such as cycling behavior or threshold response could be successfully identified. However, to analyze complex growth and maturation processes at a systems level and to quantitatively predict the outcome of perturbations will require further advances in both experimental and theoretical methodologies.

I. INTRODUCTION

Cells do not live in isolation, but have evolved mechanisms to communicate. Principal signals used are direct cell-to-cell contact and secreted molecules that bind

to cell surface receptors. Arrays of intracellular proteins form signal transduction pathways and connect to receptors, facilitating signal transmission from the extracellular compartment to the nucleus and thereby triggering various biological responses. A key mechanism used for signal transduction is phosphorylation due to its simplicity, flexibility, and reversibility. In the late 1970s, it was discovered that the oncogene *v-Src* can transform cells, possesses protein kinase activity, and causes an increase in tyrosine phosphorylation (Hunter 1980). This led to an intense hunt for the underlying mechanisms facilitating signal transduction. As a consequence, many components of signaling pathways were discovered but it remained unknown how information is processed and how cellular responses are regulated.

Signaling pathways do not operate in isolation but form complex cellular networks that regulate biological functions in a context-dependent manner. It became evident that to identify regulatory mechanisms and to predict the behavior of these networks mathematical models could be very helpful. The initial attempts to model signaling were primarily based on qualitative data, reflecting the possible interactions between the components, and on computer simulations with ad hoc fixed parameters or parameters extracted from the literature (Bhalla and Iyengar 1999; Fussenegger et al. 2000). However, these parameters frequently rely on experiments performed in different cellular settings or on *in vitro* studies. From these studies it could not be decided whether the model structure was incorrect or whether the parameters were ill chosen if the computational simulations did not fit experimental observations. Thus, to understand the dynamic behavior of signaling pathways at a systems level it is essential to combine mathematical model building with experiments (Kitano 2002; Eungdamrong and Iyengar 2004) and establish data-based models.

II. CONCEPTS AND PRINCIPLES OF SIGNAL TRANSDUCTION

A. The cell: structural organization

Complex organisms are highly organized assemblies of specialized cells. Despite these differences, all cells share common fundamental properties and represent a "unit" in living organisms. They are surrounded by a plasma membrane, use DNA as their genetic material, and employ the same basic mechanisms for energy metabolism. There are two types of cells: the *eukaryotic cell* (which participates in the formation of complex organisms and contains a nucleus, cytoplasmic organelles, and a cytoskeleton) and the *anuclear prokaryotic cell* (bacteria), which lacks these components.

To maintain integrity, cells are surrounded by *lipid membranes* that form a shell and separate the cell interior from the environment (Figure 8.1). The principal building blocks of membranes are phospholipids, which are amphipathic molecules consisting of two hydrophobic long fatty acid chains linked to a phosphate-containing

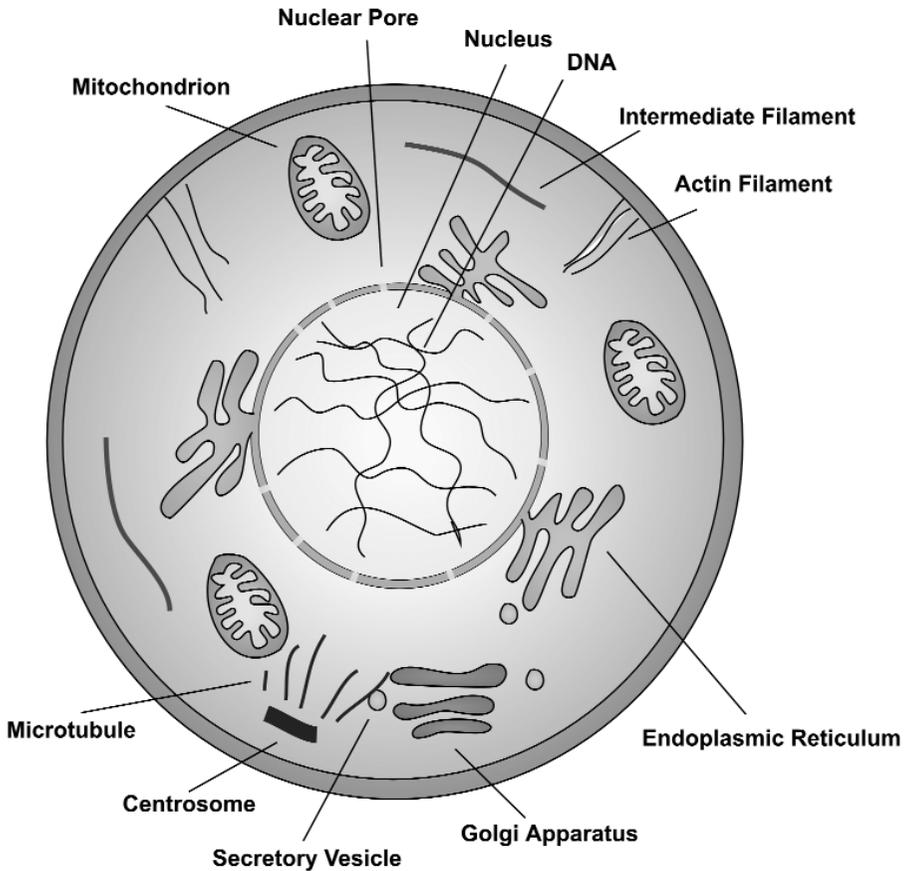


Figure 8.1. Structural organization of the mammalian cell. The major elements are schematically indicated.

hydrophilic head group. These molecules spontaneously form a bi-layer by orienting the charged head groups to interact with the surrounding water, and fatty acid tails to form a hydrophobic interior. In addition, the membrane of mammalian cells contains glycolipids and cholesterol, which increase the rigidity. An important property of membranes is that they behave as 2D fluids and their fluidity is influenced by temperature and lipid composition.

In mammalian cells, membranes not only segregate the cell interior from the environment but surround intracellular organelles. This facilitates extensive sub-cellular compartmentalization and enables mammalian cells to function efficiently. The largest organelle is the *nucleus*, which harbors the cell's genome (DNA) and is the site of transcription (RNA synthesis). Only the final stages of gene expression, the synthesis of proteins (translation), take place in the cytoplasm. Hence, the

nucleus not only serves to store genetic information but controls cellular responses. By separating the genome from the cytoplasm, post-transcriptional modifications such as RNA splicing can take place before the messenger (m)RNA is transported to the cytoplasm, where protein synthesis occurs and the access of proteins to the genetic material is limited.

In contrast to bacteria, which lack a nucleus, this opens novel opportunities for the regulation of gene expression in mammalian cells, including the selected transport of transcription factors from the cytoplasm to the nucleus. Another large organelle present in multiple copies in the cytosol of mammalian cells are *mitochondria*, in which most of the cellular ATP is generated by oxidation of small molecules. Therefore, mitochondria are regarded as the “power plant” of the cell. A large network of interconnected membrane enclosed tubules forms the *endoplasmic reticulum* (ER), which extends from the nuclear membrane throughout the cytosol. The major task of the ER is sorting of proteins destined for secretion from the plasma membrane. Polypeptide chains are translocated into the ER, where protein folding and processing takes place. From the ER, proteins are transported within membranous vesicles to the *Golgi apparatus* and further delivered to the cell surface membrane or are secreted.

In addition to the membrane-enclosed organelles, a network of protein filaments extends through the cytoplasm, forming the *cytoskeleton* and providing another level of organization. The cytoskeleton provides a structural framework determining the cell shape and cellular movements, including transport of organelles. In contrast to the rigid implications, the cytoskeleton undergoes constant remodeling and thus reflects a highly dynamic entity. There are three principal types of protein filaments: actin filaments, intermediate filaments, and microtubules. Actin filaments are generated by head-to-tail polymerization of actin monomers forming a helical structure. Assembly and disassembly of these filaments is tightly regulated by actin binding proteins.

Upon interaction with the motor protein myosin, actin filaments support a variety of movements of cells. Intermediate filaments are polymers of different proteins expressed in various cell types and possess a rope-like structure. They are not involved in cellular movement, but provide mechanical support. Microtubules are formed by reversible polymerization of tubulin in dependence of GTP hydrolysis. They are extended outward from a centrosome and the mitotic spindle forms during mitosis that is responsible for chromosomal separation. Two families of motor proteins, kinesins and dyneins, associate with microtubules and promote movement as well as positioning of organelles in the cytoplasm.

B. Signal transmission from the cell surface to the nucleus

In multicellular organisms, cells do not live in isolation but rely on specific mechanisms to communicate (Figure 8.2). In close proximity, direct cell-to-cell contact is used, whereas soluble ligands also permit communication over distances. However, integral membrane proteins (receptors) in the cell membrane are essential because

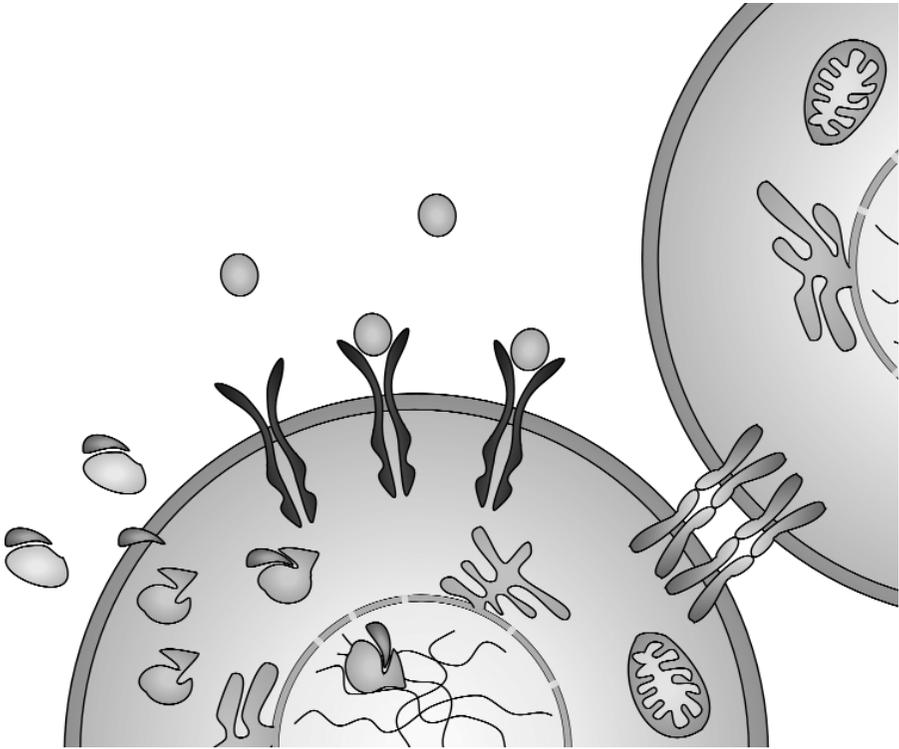


Figure 8.2. Signals used for communication of cells. Secreted soluble ligands (light blue) bind to cell surface receptors (blue). In the extracellular space, hydrophobic ligands (purple) are bound to carrier proteins (rose). In proximity to cells, they dissociate from the carrier, migrate through the cell surface membrane, and bind to receptors (light purple) present in the cytoplasm or nucleus. Alternatively, signals are transmitted by direct cell-to-cell contact mediated by cell surface proteins (green).

cells are surrounded by a lipid membrane that cannot be penetrated by hydrophilic ligands such as hormones and growth factors. They bind the ligand in the extracellular space and mediate signal transmission into the cell interior by activating specific signaling cascades.

Finally, the signal is transported across the nuclear membrane and gene expression is modulated. Alternatively, hydrophobic ligands such as steroid hormones or thyroxine are transported by carrier proteins and diffuse after dissociation from the carrier into the cytosol or nucleus, where they bind to specific receptors that regulate transcription of target genes. The principal modes used for intracellular communication are phosphorylation, second messengers, degradation, and complex formation.

Phosphorylation: To convey an intracellular signal, modifications introduced have to be transient. The most general regulatory device adopted by eukaryotic cells is

protein phosphorylation because it is simple and reversible, and because ATP is readily available as a phosphoryl donor. The key enzymes for protein phosphorylation in target proteins are protein kinases (which transfer a phosphoryl group from ATP to the hydroxyl group of tyrosine), serine, or threonine residues, whereas protein phosphatases counter-balance the reaction by removing phosphate groups from proteins. Reversible phosphorylation of proteins regulates nearly every aspect of cell life by increasing or decreasing the biological activity of enzymes, stabilizing or marking proteins for destruction, facilitating or inhibiting movements between subcellular compartments, and initiating or disrupting protein-to-protein interaction. Abnormal phosphorylation is the cause or the consequence of many human diseases.

Protein kinases possess a highly conserved overall structure (Huse 2002) and operate as molecular switches. The “on” state (which represents maximal activity) is highly similar in different protein kinases, whereas in the “off” state kinases have minimal activity and adopt a conformation that is distinct for different protein kinase classes. The transition between the two states is highly regulated by phosphorylation, interaction with additional domains, and/or binding of regulatory proteins.

This tight regulatory mechanism was first identified in cytoplasmic *tyrosine protein kinases* of the *src*-family (Harrison 2003), which in addition to the protein kinase domain possess an *src*-homology (SH)2 domain facilitating binding to specific phosphotyrosine residues localized within certain binding motifs and an SH3 domain mediating binding to proline-rich motifs. In addition to cytoplasmic tyrosine kinases, several cell surface receptors possess a tyrosine kinase domain in their cytoplasmic part. *Receptor tyrosine kinases*—such as the epidermal growth factor receptor (EGF-R) (Schlessinger 2002) and the platelet-derived growth factor receptor (PDGF-R) (Heldin 1992)—are characterized by specific domains within the extracellular portion that interacts with the ligand, by a single transmembrane domain, and by a tyrosine kinase domain in part exposed to the cell interior.

The tyrosine kinase activity is tightly regulated by multiple autoinhibitory mechanisms, including an inhibitory conformation of the extracellular domain, the transmembrane domain, the juxtamembrane domain, and the activation loop. Ligand binding to the extracellular domain causes a conformational switch that leads to the activation of the tyrosine kinase domain. Other cell surface receptors (such as the hematopoietic *cytokine receptors*, including the interleukin receptors) lack enzymatic activity (D’Andrea 1989) but couple with cytoplasmic tyrosine kinases of the Janus kinase family. Ligand binding to the cytokine receptors causes activation of the receptor-associated Janus kinase and results in tyrosine phosphorylation of the receptor on multiple tyrosine residues.

Phosphorylation on serine or threonine residues occurs much more frequently than tyrosine phosphorylation but is less inducible. The overall structure of *serine/threonine protein kinases* is very similar to tyrosine protein kinases but the regulation is mediated by additional subunits that bind second messengers or vary in their expression level (Johnson 1996). Another mode of regulation is achieved by phosphorylation or dephosphorylation on multiple residues. For example, cell cycle

control is performed by protein serine/threonine kinases of the cyclin-dependent kinase family that are inactive as monomers but activated by cyclin binding.

Regulation of the cell cycle is achieved by synthesis and destruction of cyclines, phosphorylation of the activation loop and the ATP binding loop in the cyclin-dependent kinases, and binding of an inhibitor. Counterintuitive is the regulation of the protein serine/threonine kinase glycogen synthase kinase 3 (GSK-3), which lies at the crossroads of metabolism and signal transduction (Dajani 2001). GSK-3 is active as kinase in the absence of signal and processively phosphorylates substrates at multiple residues that are already prephosphorylated at a C-terminal residue. Upon growth factor binding to cell surface receptors, GSK-3 is phosphorylated at the N-terminus, which turns the N-terminus into a pseudosubstrate and thereby blocks the catalytic cleft of the kinase.

The mitogen-activated protein (MAP) kinases form a signaling cascade consisting of an array of protein serine/threonine kinases (Raman 2003). These protein kinases are characterized by their ability to use protein kinases as substrate and phosphorylate them at two residues, which is required for full activation. Contrary to receptor tyrosine kinases, only one *receptor serine/threonine kinase family* is known (Shi 2003). The transforming growth factor (TGF β) beta receptors type I and II possess serine/threonine kinase activity in their cytoplasmic domain, which is regulated by autophosphorylation and inhibitor binding.

The activation of signal transduction is counter-balanced by the activation of *protein phosphatases* (Tonks 1996), which remove the phosphoryl group from tyrosine, serine, or threonine residues by a cystein-catalyzed mechanism. Characteristic of protein tyrosine phosphatases is the multidomain substructure. Protein tyrosine phosphatases that are located at the cell membrane contain tandem protein phosphatase domains autoregulated by wedge-like structures. The cytoplasmic protein tyrosine phosphatases of the SHP1/SHP2 family harbor two N-terminal SH2 domains that block the protein tyrosine phosphatase domain in the inactive state.

Upon activation of signal transduction, the SH2 domains mediate recruitment to tyrosine-phosphorylated receptors and thereby open the phosphatase domain. Tyrosine phosphorylation within cells is rapidly induced by stimulation of cells, but declines soon after. Serine/threonine protein phosphatases share a homologous catalytic domain and are regulated by multiple regulatory subunits controlling phosphatase activity and selection of substrate (Janssens 2001). The most prominent examples are protein phosphatase PPI and PPIIa.

In addition to phosphorylation on proteins, phosphorylation of phospholipids (in particular, *phosphoinositides*) is used for signal transduction. Phosphoinositides are characterized by an inositol head group that can be phosphorylated by phosphoinositide kinases on multiple hydroxyl groups and that serves as a lipid-derived second messenger (playing a role in vesicle trafficking and signal transduction). The central enzyme for signal transduction is the phosphoinositide 3 (PI3) kinase, which phosphorylates phosphoinositides at the D-3 position of the inositol ring structure (Cantley 2002).

Best studied is the class IA PI3 kinase, which is composed of a regulatory subunit (p85) and a catalytic subunit (p110). Growth factor stimulation results in a transient increase in phosphoinositide-3,4-bisphosphate (PtdIns-3,4-P₂) or phosphoinositide-3,4,5-trisphosphate (PtdIns-3,4,5-P₃), which is rapidly counteracted by phosphoinositide phosphatases—such as the SH2-domain, containing inositol 5-phosphatase SHIP, and the phosphatase and tensin homolog deleted on chromosome 10 (PTEN)—which removes specific phosphate groups of phosphoinositides.

Another mode used for intracellular communication is protein-bound guanosine triphosphate (GTP). *GTP-binding proteins* such as Ras belong to the GTPase superfamily and are molecular switches that alternate between the GTP-bound activated state and a GDP-bound off state (Downward 1997). The activation is accelerated by a guanine nucleotide-exchange factor (GEF) that promotes dissociation of GDP from Ras and thus the formation of a Ras-GTP complex. Binding of a GTPase-activating protein (GAP) to the Ras-GTP complex results in GTP hydrolysis and GAP dissociation and thus the formation of the inactive Ras-GDP complex.

C. Complex formation

To ensure intracellular communication, *modular interaction domains* have evolved that recognize transient modifications (Pawson et al. 2004). These domains fold independently, are incorporated in larger polypeptides, and recognize exposed sites on their protein or lipid partners. The first modular interaction domain discovered was the src-homology (SH)2 domain in the N-terminus of the cytoplasmic tyrosine kinase Src. This domain comprises a block of 100 amino acids and recognizes phosphotyrosine residues in conjunction with a C-terminally localized short recognition motif. Closely related is the phosphotyrosine-binding (PTB) domain that recognizes phosphotyrosine residues localized with an N-terminal NPXY motif.

Less frequent are domains that specifically recognize phosphoserine/threonine residues. Best characterized are the 14.3.3 proteins, which are highly abundant dimeric proteins binding phosphoserine within RXXpSXP motifs. The phosphorylation of phosphoinositides in the cellular membrane at the D-3 position is recognized by pleckstrin homology domains and thereby mediates translocation of signaling proteins to the cellular membrane. Direct protein-to-protein interaction is mediated by several modular interaction domains, such as the src-homology (SH)3 domain that recognizes PXXP motifs. Other examples are the WW-domain (which interacts with PPXY motifs) and the PDZ domain, which binds to ES/TDV motifs. A class of signaling molecules entirely composed of modular interaction domains and lacking enzymatic activity are adapter proteins or scaffolds.

D. Proteolytic cleavage and degradation

A common mechanism for regulating the activity of enzymes is mediation by *proteolytic cleavage*, which ensures processing of hormones from larger precursor pro-

teins and mediates activation of enzymes involved in blood coagulation, digestion, or programmed cell death. The ultimate effectors and executors of programmed cell death are caspases, a family of proteases (characterized by a cysteine in the active site) that cleave after an aspartic acid residue in their substrate. A large transmembrane protein that controls cell fate during development is notch. Ligand-binding is mediated by cell-to-cell contact and results in proteolytic cleavage of notch and translocation of the cytoplasmic domain into the nucleus.

The activity of proteins is not only controlled by synthesis and processing but by the rate of *degradation*, which determines the life span of intracellular proteins. Whereas membrane proteins or aged organelles are primarily degraded within lysosomes, the degradation of cytosolic proteins is mediated by chemical modification of lysine residues by the addition of *ubiquitin*, a 76-residue polypeptide (Bonifacino 1989). The process involves three consecutive steps. A ubiquitin-activating enzyme (E1) is activated by the addition of ubiquitin. Ubiquitin is transferred to a cysteine residue in the ubiquitin-conjugating enzyme (E2). Finally, the peptide bond formation between ubiquitin and lysine in the target protein is catalyzed by a ubiquitin ligase (E3). These steps are repeated multiple times, resulting in the formation of polyubiquitinated proteins that are recognized by the *proteasom* machinery and cleaved into short peptides.

E. Second messenger

Binding of ligands (first messengers) results frequently in the production of short-lived small molecules (*second messengers*). The first identified second messenger was cyclic AMP (cAMP), which regulates the activity of protein kinase A. The binding of cAMP to the regulatory subunit results in the dissociation of the inactive tetramer and activation of the catalytic subunit. Because the binding is positively cooperative, small changes in cAMP concentration are translated into large changes of protein kinase A activity. Similarly, cyclic GMP (cGMP) regulates the activity of protein kinase G and the opening of rod channels.

Lipid-derived second messengers are diacylglycerol (DAG)—which contributes to the activation of protein kinase C—and inositol-1,4,5-trisphosphate (IP₃), which triggers the opening of Ca²⁺ channels in the endoplasmic reticulum. The release of Ca²⁺, another second messenger, into the cytosol facilitates binding of protein kinase C to the cell membrane and activation by DAG. Phosphoinositides phosphorylated at the D-3 position of the inositol ring structure are not cleaved but remain imbedded in the cell membrane and act as second messengers.

F. MicroRNA

Most recently identified as a modulator of signal transduction are microRNAs (Ambros 2004). They are a family of 21–25-nucleotide small non-coding RNAs that regulate gene expression in a sequence-specific manner. In mammalian cells, microRNAs are expressed in a developmentally regulated or tissue-specific manner

and affect protein synthesis from their complementary target RNA. Bioinformatic prediction of microRNA targets has been used to examine the function of microRNAs (Lewis et al. 2003; Rajewsky and Socci 2004), but these predictions remain to be experimentally validated.

III. SIGNALING PATHWAYS: FORMATION OF NETWORKS

Intracellular proteins form signaling cascades that use the described modes of communication to transmit signals from the cell surface to the nucleus. A rather simple and fast signaling cascade is the *JAK-STAT pathway* (Figure 8.3, left-hand panel), which mediates signal transduction primarily through hematopoietic cytokine receptors, but also hepta-helix receptors and receptor tyrosine kinases (Rawlings 2004). The key enzyme of this cascade is a member of the cytoplasmic protein tyrosine kinase family of the Janus type, which harbors two protein kinase domains (one catalytically active and the other with regulatory functions).

Upon activation, the receptor-associated Janus kinase (JAK) is activated—leading to tyrosine phosphorylation of the receptor cytoplasmic domain. This mediates recruitment of signal transducer and activator of transcription (STAT) proteins to specific phosphotyrosine residues in the receptor via their SH2 domain. Then tyrosine phosphorylation of STATs occurs, facilitating STAT dimerization. STAT dimers depart from the receptor and migrate to the nucleus, where target gene expression is activated.

Upon dephosphorylation, STATs recycle to the cytoplasm and engage in further activation cycles. Thus, by multiple consecutive activation cycles the phosphorylation level of the receptor is constantly monitored and translated into appropriate levels of target gene expression. Among the induced genes are genes encoding suppressor of cytokine signaling (SOCS) proteins, which inhibit signaling through hematopoietic cytokine receptors and thereby constitute a negative feedback loop that ensures tight regulation of the JAK-STAT pathway.

Another signaling cascade that operates very similarly is the *SMAD signaling pathway* (Figure 8.3, right-hand panel), which is activated by TGF β receptors type II and type I possessing protein serine/threonine kinase activity (Shi 2003). Ligand binding is facilitated by the type III TGF β receptor, a proteoglycane lacking enzymatic activity that delivers the ligand to the type II and type I receptors. Upon oligomerization, the type II receptor phosphorylates the type I receptor at the glycine/serine (GS) motif located in the juxtamembrane domain. This leads to the activation of the serine/threonine kinase activity of the type I receptor and consequently to receptor recruitment of the receptor (R) SMAD-3 and -2.

The R-SMADs are serine-phosphorylated, depart from the receptor, form a trimeric complex with the common SMAD-4, translocate to the nucleus—where they bind to nuclear transcription factors and activate target gene transcription, including the negative regulatory SMAD7. SMAD7 has a higher affinity to the activated type I receptor and thereby displaces the R-SMADs, resulting in down-

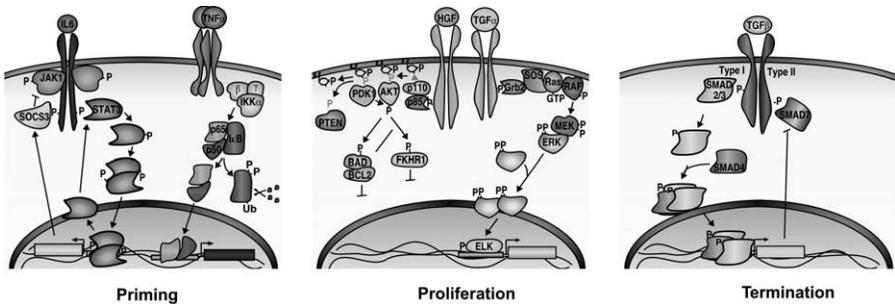


Figure 8.3. Signaling pathways involved in hepatocyte regeneration. During the priming stage interleukine (IL)6 binds to its receptor. The receptor-associated tyrosine kinase JAK1 is activated by phosphorylation (P), and the cytoplasmic domain of the IL6 receptor (blue) is phosphorylated at multiple tyrosine residues. The signal transducer and activator of transcription is (STAT)3. STAT3 is tyrosine phosphorylated, dimerizes, migrates to the nucleus, and binds to the promoter of target genes.

Upon dephosphorylation, STAT3 leaves the nucleus and re-enters the cytoplasm. One of the target genes is the gene encoding the suppressor of cytokine signaling (SOCS)3, which inhibits JAK1. The tumor necrosis factor (TNF) alpha binds as trimer to the trimeric TNF receptor, resulting in the activation of the nuclear factor NFκB signaling cascade. NFκB consisting of p65 and p50 binds to the inhibitor IκB and is sequestered as inactive complex in the cytoplasm. Upon activation of the inhibitor kinase (IKK)—composed of an α, β and γ subunit—IκB is serine phosphorylated (P), dissociates from the complex, is ubiquitinated (Ub), and degraded by the proteasom (scissor).

During the proliferation stage, the receptor tyrosine kinases, met (receptor for the hepatocyte growth factor), and epidermal growth factor receptor EGF-R (receptor for transforming growth factor TGF alpha) activate the phosphoinositide (PI)3 kinase cascade and the mitogen-activated (MAP) kinase pathway. The regulatory subunit of PI3 kinase p85 is recruited to the tyrosine phosphorylated receptor. The catalytic subunit of PI3 kinase p110 phosphorylates phosphoinositides at the D-3 position. The product is recognized by the pleckstrin homology domain of the protein serine/threonine kinases Akt/protein kinase B and phosphoinositide-dependent kinase (PDK1)—and PDK1 phosphorylates Akt, resulting in full activation of Akt.

Akt phosphorylates the apoptosis-promoting protein BAD and the fork-head-related transcription factor 1 (FKHR1), and inhibits the activity of the proteins. Activation of the MAP kinase cascade is initiated by recruitment of the growth factor receptor (Grb)2, associated protein to the tyrosine phosphorylated receptor. This promotes cell membrane recruitment of son-of-sevenless (SOS) guanine exchange factor, activation of Ras in the GTP bound form, and membrane translocation as well as activation of the serine/threonine kinase Raf. Raf triggers phosphorylation of MEK on two serine residues, and activated MEK phosphorylates the extracellular signal regulated kinase (ERK) on a tyrosine and a threonine residue.

The phosphorylated ERK dimerizes, translocates to the nucleus, and phosphorylates (for example) the transcription factor ELK, which modifies the DNA-binding activity of ELK. The transforming growth factor (TGF) β participates in termination of hepatocyte regeneration. TGF binds to the type II and type I receptor, which possesses serine/threonine kinase activity. The phosphorylated type I receptor mediates recruitment of the receptor SMAD proteins 2 and 3. SMAD2/3 is serine phosphorylated, dissociates from the receptor, and forms a complex with the common SMAD4. The complex migrates to the nucleus, activates target genes, is dephosphorylated, and relocates to the cytoplasm. Among the target genes is the inhibitory SMAD7, which competes with SMAD2/3 for receptor binding and inhibits the pathway (see color plate 4).

modulation of the pathway. Additional regulation is ensured by the transcriptional repressors SNON and SKI, which bind to SMADs in the nucleus and form an inhibitory complex. Like the JAK-STAT signaling cascade, the SMAD pathway shows cycling behavior.

The *MAP kinase cascade* (Figure 8.3, middle panel) is formed by the consecutive activation of three serine/threonine protein kinases (Raman 2003). The MAP kinase Raf is represented by three isoforms (Raf-1, B-Raf, and A-Raf), which are regulated by various inhibitory and activating phosphorylation events and phosphorylate MAP kinase (MEK) at two serine residues. MEK is a dual-specificity protein kinase and phosphorylates MAP kinase (ERK1 and ERK2) at a tyrosine and a threonine residue within the YPT motif, which in turn activates the MAP kinases ERK1 and ERK2. The activated MAP kinases dimerize and translocate to the nucleus, where they phosphorylate transcription factors such as ELK-1. The kinase cascade is organized by scaffold proteins, and at multiple levels negative feed-back loops ensure regulated activation.

The *NF κ B signaling pathway* (Figure 8.3, right-hand panel) combines signal transduction through phosphorylation with complex formation and selective degradation. In the absence of signaling, NF κ B is sequestered by the inhibitory subunit I κ B in the cytoplasm (Chung 2002). Upon the activation of signal transduction through the TNF receptor 2/interleukin-1 receptor family (trimeric receptors lacking endogenous enzymatic activity connecting to trimeric TNF receptor-associated factor (TRAF)), the I κ B kinase is activated and phosphorylates I κ B on serine residues—thereby marking the inhibitory subunit for destruction through proteasomal degradation.

NF κ B is released, which migrates to the nucleus and activates target gene transcription. On the contrary, receptors of the TNF receptor 1 type harbor death domains in their cytoplasmic part and can promote programmed cell death (apoptosis). Ligand-induced receptor trimerization facilitates the assembly of a death-inducing signaling complex (DISC), leading to caspase 8 recruitment and activation. Caspase 8 is an initiator caspase that activates other caspases and thereby promotes signal amplification through the cascade.

Another signaling cascade that operates by selective complex formation and phosphorylation is the *Wnt/ β -catenin signaling cascade* (Dajani 2001). In the absence of a Wnt, signaling of the scaffold protein axin forms a complex with APC, β -catenin, and the protein serine/threonine kinase GSK3, resulting in constitutive phosphorylation of β -catenin. Phosphorylated β -catenin is marked for proteasomal degradation and therefore does not accumulate in cells. Upon Wnt binding to its receptor Frizzled, Disheveled is recruited and inhibits GSK3. Unphosphorylated β -catenin accumulates in the cytosol, subsequently migrates to the nucleus, and mediates target gene activation in conjunction with transcription factor TCF.

Several signaling pathways use phosphoinositides as mediators, and phosphoinositide-4,5-bisphosphate is the common precursor used. As part of the *canonical inositol triphosphate (IP₃) signaling cascade*, the activation of several receptors leads to phospholipase C activation—resulting in cleavage of phospho-

inositide-4,5-bisphosphate to DAG and IP_3 . IP_3 diffuses through the cytoplasm and triggers opening of Ca^{2+} channels in the endoplasmic reticulum. The rise in cytosolic Ca^{2+} facilitates binding of protein kinase C to the membrane and activation of the protein kinase activity by DAG. In contrast, the phosphoinositides modified by the *lipid kinase PI3 kinase* (Figure 8.3, middle panel) are not cleaved and function as membrane-embedded second messengers (Cantley 2002).

Ligand-induced receptor tyrosine phosphorylation mediates recruitment of PI3 kinase via the SH2 domains of the regulatory subunit p85. This places the catalytic subunit p110 in proximity to substrates and results in the phosphorylation of phosphoinositide 4 phosphate and phosphoinositide-4,5-bisphosphate at the D-3 position of the inositol ring. Phosphoinositide-3,4-bisphosphate and phosphoinositide-3,4,5-triphosphate is recognized by the pleckstrin homology domain of two cytosolic serine/threonine kinases, the phosphoinositide-dependent kinase (PDK)1, and Akt/protein kinase B. PDK1 has a low basal activity and is fully activated by engagement of the PH domain, at the cell membrane. Mediated by the PH domain, Akt is translocated to the cell membrane and requires phosphorylation by PDK for full activation. Multiple downstream targets have been identified for both kinases, including factors involved in protein synthesis, cell survival, and metabolism.

The described signaling pathways do not operate in isolation but form complex signaling networks that regulate biological functions in a context-specific manner, as shown for hepatocyte regeneration in Figure 8.3. Different intracellular signaling pathways are activated in hepatocytes during the priming, proliferation, and termination stages of regeneration—regulating cell cycle re-entry, proliferation, and re-differentiation. Due to the complexity of biological responses, a deeper understanding cannot be achieved by traditional approaches but requires the combination of experimental data with mathematical modeling.

A. Mathematical modeling of signaling pathways

1. Modeling approaches

Dynamic growth and differentiation processes, such as hepatocyte regeneration, are regulated by the coordinated activation of multiple signaling pathways that form complex signaling networks. In addition to a particular pathway being triggered, critical are timing, amplitude, and duration of activation. However, insight into the characteristic dynamic behavior and design principles of signaling pathways cannot be achieved by intuitive approaches alone but requires the combination of experimental and theoretical approaches, including *model building* (Eungdamrong and Iyengar 2004). Typically, signaling pathways are represented by simple cartoons that qualitatively indicate the connection between the individual components but lack information about the dynamic behavior.

To translate these graphical representations into mathematical descriptions, first the involved biochemical reactions have to be specified. Non-covalent interactions

such as complex formation mediated by protein-to-protein interaction show a linear response relationship and can be described using the laws of mass action. The formation of dimers or oligomers is frequently used by signaling pathways to generate signaling complexes with changed biological activity, such as the STAT dimers that acquire high-affinity DNA-binding capacity or the assembly of the death-inducing signaling complex (DISC) at the cytoplasmic domain of the TNF-receptors. Enzyme-mediated reactions such as phosphorylation/dephosphorylation or protein synthesis/degradation show in the simplest case a hyperbolic response and are governed by Michaelis—Menten kinetics.

Sigmoidal response curves result from allosteric regulation of enzymatic activity; for example, due to increased activity of enzymes in response to binding of multiple ligands. This permits graded responses and enables the enzyme to react to small changes in ligand concentration, resulting in a switch-like behavior. A well-studied example for allosteric regulation and positive cooperativity is the binding of the second messenger cAMP to the inhibitory subunit of the serine/threonine protein kinase A, permitting the activation of the protein kinase in response to small local changes in cAMP.

Alternatively, the existence of positive feedback loops in which a downstream component of a signaling pathway accelerates the activity of an upstream component can lead to sigmoidal response curves. Furthermore, negative feedback loops or a certain combination of negative and positive feedback loops within signaling pathways can result in oscillatory responses. For the **NFκB** signaling cascade, oscillation of the nuclear localization of **NFκB** in dependence of the expression of the inhibitory signaling component **IκB** has been observed.

For the mathematical representation of signaling pathways, currently two major approaches are being applied. The most frequently used is a *deterministic* representation that considers bulk concentrations of pathway components (not individual molecules) and that assumes the cell is a well-stirred reactor. If the molecules are present in sufficient concentrations, the reactions can be described by chemical kinetic models based on ordinary differential equations (ODEs) representing the concentration as a function of time.

Critical for this type of modeling approach are the starting concentrations of all reactants and the rate constants of the reactions. If these are specified, the changes in concentration of reactants over time can be quantitatively predicted. The other approach is to apply *stochastic models* for reactants that exist in small concentrations. In this case, a reaction might or might not occur during a given time period. The fluctuations (or “noise”) inherent in such stochastic systems are exploited for cellular functions, resulting in switch-like behavior.

The majority of mathematical models so far established for signaling pathways disregard as a first approximation the spatial organization of cells. To capture this additional level of complexity, *compartmental models* can be established that also use ODEs but treat the same molecule in different compartments as distinct species and model the flux of the molecules between the compartments. In models considering the concentration as a function of time and space, variables require the

use of partial differential equations (PDEs). Finally, delay differential equations (DDEs) model delayed reactions (e.g., during which one reactant rests in a different compartment, such as the nucleus), and differential algebraic equations (DAEs) also comprise algebraic side conditions (such as the total concentration of protein A summed over all reactions equals a positive constant).

2. *Parameter estimation and data-based mathematical models of signaling pathways*

Parameter estimation and *sensitivity analysis* have been identified as key components for model identification. Parameter estimation refers to the determination of values of unknown model parameters to provide an optimal fit between the simulation and experimental data (Deuflhard 1983). The identification of critical system parameters can be achieved by sensitivity analysis. Sensitivities describe the relative changes of molecule concentrations (and therefore of the system behavior) as a result of changes of the parameters.

Notably, sensitivities can be determined for specific sets of parameters only (local sensitivity analysis). Thus, sensitivity analysis can usually only be applied if most parameters are known or can be estimated. The number of assessable parameters, and therefore the maximum size of the model, has been very limited due to the large amount of experimental data required for high-dimensional parameter estimation problems and the curse of dimensionality. Curse of dimensionality refers to the problem that the space of possible sets of parameter values grows exponentially with the number of unknown parameters, severely impairing the search for the globally optimal parameter values.

Whereas parameter estimation of ODE systems has been greatly advanced, the necessary procedures for PDEs are much more complex and require further development. Critical for the estimation of meaningful parameters is on the one hand the existence of high-quality quantitative highly sampled experimental data, which currently represents one of the major bottle-necks in systems biology approaches.

Recent examples show that new biological knowledge regarding general design principles of signaling pathways operating in mammalian cells can be generated by combining *quantitative experimental data* with *mathematical modeling* of sub-modules of signaling networks. For the core module of the JAK-STAT signaling cascade, a deterministic mathematical model consisting of coupled ODEs has been established (Swameye et al. 2003; Nicolas et al. 2004). The dynamical parameters of the model were estimated from time-course experiments measuring receptor and STAT5 activation using Bock's multiple shooting technique (Bock 1981, 1983).

During parameter estimation, time-series values given by a parameterized model are compared to measured data (e.g., via the mean square distance). By changing the parameter values, this distance increases or decreases. The parameter set leading to the global minimum of the mean square distance is the least square estimate of the model parameters. Unfortunately, many estimation algorithms finish

the estimation procedure in local minima rather than in the global minimum. This occurs especially in ODE parameter estimation, wherein after each parameter change the model has to be integrated to obtain model time series. The integration needs initial values, which are usually the measured data points at the first time point. Hence, the integration step uses only a small fraction of the given information, often leading to a local minimum for the mean square estimate.

The multiple-shooting approach divides the measured time series into several parts and integrates the ODE model on the subsections based on the first data point of the corresponding measurements. The resulting model time series is by construction in every part at least for one time point close to the measured data set, but possesses discontinuities between the subsection. Removing the discontinuities and decreasing at the same time the mean square distance requires sophisticated numerical algorithms. It has been shown in many applications that the multiple-shooting algorithm is well suited to finding global optimal parameters (Timmer 1998). For initial hypothesis testing, a model based on the traditional assumption of signaling pathways as linear feed-forward cascade was compared to a model capturing a cycling behavior of STAT5.

Independent of the parameter values, it was not possible to fit the experimental data with the linear feed-forward cascade model, but only with the model including the cycling behavior of STAT5, which was modeled using a delay term reflecting the sojourn time of STAT5 in the nucleus. By applying the fitted model, the parameters of nuclear export and import were identified as most sensitive to perturbation, a prediction that was experimentally verified. *In silico* investigations revealed that STAT5 undergoes rapid nucleocytoplasmic cycling, continuously coupling receptor activation and target gene transcription. The identification of rapid nucleocytoplasmic cycling as a general design principle of signaling pathways was confirmed by studies on the SMAD signaling pathway using fluorescence microscopy (Nicolas et al. 2004).

Other examples are mathematical models based on experimental data describing the temporal control of **NFκB** activation by the coordinated degradation and synthesis of **IκB** proteins. Hoffmann et al. (2002) reported a computational model of the **NFκB**–**IκB** module consisting of ODE that involved two compartment kinetics of **NFκB** and **IκB** activation. The model was based on the analysis of genetically reduced systems. Some of the parameters were taken from the literature, but several were determined by model fitting to quantitative time-course data determining the DNA-binding activity of **NFκB** in embryonic fibroblast harboring a single **IκB** isoform.

The modeling approach revealed that **IκBα** ensures rapid turn-off of **NFκB** responses, thereby representing a negative feedback loop—whereas **IκBβ** and ϵ stabilize nuclear **NFκB** localization during longer responses and dampen the oscillation of **NFκB**. Furthermore, a bimodal signal-processing characteristic with respect to stimulus duration was revealed regulating selective target gene expression. To explore modeling of the **NFκB** signaling pathway at the single-cell level, Nelson et al. (2004) combined time-lapse imaging of fluorescent-labeled **NFκB** and

I κ B in the cytosol and nucleus with use of the computational model established by Hoffmann et al.

This analysis revealed that just two variables, the concentration of free IKK and I κ B α , were intimately coupled to the oscillatory dynamics of nuclear NF κ B. Computational simulation and experimental variation of the I κ B expression rate gave comparable results, demonstrating that increased I κ B expression damped the oscillation in NF κ B localization. The expression of target genes was dependent on the oscillation persistence, amplitude, and period of NF κ B nuclear localization.

An even more complex task was the mathematical modeling of the CD95-induced apoptosis pathway (Bentele et al. 2004). The previously mentioned attempts to model signal transduction pathways were limited to small systems. The signaling cascade of CD95-induced apoptosis consists of more than one hundred reactions involving molecule species and reaction parameters in the same order of magnitude. To address the complexity of apoptotic signaling, we subdivided the entire system into subsystems of different information qualities. A new approach for sensitivity analysis within the mathematical model was key for the identification of critical system parameters and two essential system properties (modularity and robustness). The model well described the regulation of apoptosis on a systems level, and revealed a threshold mechanism for the regulation of apoptosis. The model predictions were verified experimentally.

B. Challenges for performing kinetic measurements on a large scale

These examples demonstrate that important biological knowledge can be generated by data-based mathematical models. However, the establishment of models critically depends on the generation of *high-quality quantitative data*. The majority of techniques currently used aim at the generation of qualitative data. However, to reproducibly generate quantitative temporal and possibly spatially resolved data several adjustments have to be considered.

The techniques that have been successfully used for data-based mathematical modeling include (1) quantitative immunoblotting that combines separation of proteins according to their molecular weight, with detection by specific antibodies followed by chemiluminescence-detection (Swameye et al. 2003; Bentele et al. 2004), (2) electrophoretic mobility shift assay (EMSA), which enables one to measure the DNA-binding capacity of proteins by mixing protein extracts with radioactive-labeled DNA probes (Hoffmann et al. 2002), and (3) imaging of green fluorescent protein (GFP)-tagged proteins in live cells by fluorescence microscopy (Nelson et al. 2004).

The currently applied techniques can be further advanced by establishing procedures to convert the obtained relative values into absolute numbers (such as molecules per cell) and to remove systematical errors. However, the limitations of quantitative immunoblotting are that only a limited number of samples can be processed at the same time and live-cell imaging is restricted by the availability of

only a limited number of spectral variants of GFP. Cell-based microarrays are being developed to study the perturbation of the function of genes in a systematic high-throughput fashion (Wu et al. 2002). However, to generate quantitative data mere overexpression or disruption of function is not sufficient, but precisely identifiable expression levels must be achieved. Thus, new techniques (for example, protein arrays with high specificity and sensitivity or advanced fluorescence microscopy techniques) have to be developed to facilitate large-scale systems biology approaches.

IV. CONCLUSIONS

To generate informative mathematical models of signaling pathways, knowledge of the involved biochemical reactions, the concentration of the components, and possibly their subcellular organization is critical. Computer simulations can be used to rapidly test different hypotheses (for example, regarding system behavior), but experimental validation is required before conclusions can be drawn. The majority of present high-throughput data results in qualitative information that is not quantitative and therefore not suited for quantitative mathematical modeling.

In general, it will be important to establish quality standards for the experimental data used for parameter estimation. This will be facilitated by the development of standard operating procedures for the techniques used for data acquisition and the data-processing procedure, and an agreement on a limited number of cellular systems that are initially analyzed by systems biology approaches.

Mathematical models of signaling pathways are initially based on hypothesis and are further refined by iterative cycles of model adjustments and experimental validation. Thus, essential for the success of systems biology is a close cooperation of model builders and experimentalists. The advancement in theoretical tools and quantitative techniques will determine whether systems biology will be able to fulfill the promise to decipher mechanisms leading to diseases and to enhance the identification of efficient therapeutic targets.

ACKNOWLEDGMENTS

I thank Sebastian Bohl, Marcel Schilling, Andrea Pfeifer, Verena Becker, Clemens Kreutz, and Thomas Maiwald for critically reading the manuscript and for many helpful suggestions.

REFERENCES

- Ambros, V. (2004). The functions of animal microRNAs. *Nature* **431**(7006):350–355.
Bentele, M., Lavrik, I., et al. (2004). Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. *J. Cell. Biol.* **166**(6):839–851.

- Bhalla, U. S., and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science* **283**(5400):381–387.
- Bock, H. G. (1981). Numerical treatment of inverse problems in chemical reaction kinetics. In *Modelling of Chemical Reaction Systems*. New York: Springer 102–125.
- Bock, H. G. (1983). Recent advances in parameter identification for ordinary differential equations. In *Progress in scientific computing*. New York: Springer 95–121.
- Bonifacino, J. S., and Weissman, A. M. (1989). Ubiquitin and the control of protein fate in the secretory and endocytic pathways. *Ann. Rev. Cell Dev. Biol.* **14**:418–428.
- Cantley, L. C. (2002). The phosphoinositide 3-kinase pathway. *Science* **296**(5573):1655–1657.
- Chung, J. Y., Park Y. C., Ye, H., and Wu, H. (2002). All TRAFs are not created equal: Common and distinct molecular mechanisms of TRAF-mediated signal transduction. *J. Cell. Sci.* **115**(4):679–688.
- Dajani, R., Fraser, E., Roe, S. M., Young, N., Good, V., Dale, T. C., and Pearl, L. H. (2001). Crystal structure of glycogen synthase kinase 3 beta: Structural basis for phosphate-primed substrate specificity and autoinhibition. *Cell* **115**(4):721–723.
- D'Andrea, A. D., Fasman, G. D., and Lodish, H. F. (1989). Erythropoietin receptor and interleukin-2 receptor beta chain: A new receptor family. *Cell* **58**(6):1023–1024.
- Deuffhard, P. (1983). *Numerical Treatment of Inverse Problems in Differential and Integral Equations*. Switzerland: Birkhäuser.
- Downward, J. (1997). Cell cycle: Routine role for Ras. *Curr. Biol.* **7**(4):258–260.
- Eungdamrong, N. J., and Iyengar, R. (2004). Computational approaches for modeling regulatory cellular networks. *Trends Cell Biol.* **14**(12):661–669.
- Fussenegger, M., Bailey, J. E., and Varner, J. (2000). A mathematical model of caspase function in apoptosis. *Nat. Biotechnol.* **18**(7):768–774.
- Harrison, S. C. (2003). Variation on an Src-like theme. *Cell* **112**(6):737–740.
- Heldin, C. H. (1992). Structural and functional studies on platelet-derived growth factor. *EMBO J.* **11**(12):4251–4259.
- Hoffmann, A., Levchenko, A., Scott, M. L., and Baltimore, D. (2002). The I κ B-NF- κ B signaling module: Temporal control and selective gene activation. *Science* **298**(5596):1241–1245.
- Hunter T., and Serton, B. (1980). Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. *Proc. Natl. Acad. Sci. USA* **77**(3):1311–1315.
- Huse, M., and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell* **109**(3):275–282.
- Janssens V., and Goris, J. (2001). Protein phosphatase 2A: A highly regulated family of serine/threonine phosphatases implicated in cell growth and signaling. *Biochem. J.* **353**(3):417–439.
- Johnson L. N., Noble, M. E., and Owen, D. J. (1996). Active and inactive protein kinases: Structural basis for regulation. *Cell* **85**(2):149–155.
- Kitano, H. (2002). Computational systems biology. *Nature* **420**(6912):206–210.
- Lewis, B. P., Shih, I. H., and Jones-Rhoades, M. W. (2003). Prediction of mammalian microRNA targets. *Cell* **115**(7):787–798.
- Nelson, D. E., Ihekwaba, A. E., et al. (2004). Oscillations in NF- κ B signaling control the dynamics of gene expression. *Science* **306**(5696):704–708.
- Nicolas, F. J., De Bosscher, K., Schmierer, B., and Hill, C. S. (2004). Analysis of SMAD nucleocytoplasmic shuttling in living cells. *J. Cell Sci.* **117**(18):4113–4125.
- Pawson, T. (2004). Specificity in signal transduction: From phosphotyrosine-SH2 domain interactions to complex cellular systems. **116**(2):191–203.

- Rajewsky, N., and Socci, N. D. (2004). Computational identification of microRNA targets. *Dev. Biol.* **267**(2):529–535.
- Raman M., and Cobb, M. H. (2003). MAP kinase modules: Many roads home. *Curr. Biol.* **13**(22):886–888.
- Rawlings, J. S., Rosler, K. M., and Harrison, D. A. (2004). The JAK/STAT signaling pathway. *J. Cell Sci.* **117**(8):1281–1283.
- Schlessinger, J. (2002). Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell* **110**(6):669–672.
- Shi, Y., and Massague, J. (2003). Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell* **113**(6):685–670.
- Swameye, I., Muller, T. G., Timmer, J., Sandra, O., and Klingmüller, U. (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *Proc. Natl. Acad. Sci. USA* **100**(3):1028–1033.
- Timmer, J. (1998). Modeling noisy time series: Physiological tremor. *International Journal of Bi-furcation and Chaos* **8**(7):1505–1516.
- Tonks, N. K., and Neel, B. G. (1996). From form to function: Signaling by protein tyrosine phosphatases. *Cell* **87**(3):365–368.
- Wu, R. Z., Bailey, S. N., and Sabatini, D. M. (2002). Cell-biological applications of transfected-cell microarrays. *Trends Cell Biol.* **12**(10):485–488.

Reconstruction of Metabolic Networks from Genome Information and Its Structural and Functional Analysis

Hong-Wu Ma and An-Ping Zeng

Experimental Bioinformatics, Research Group Systems Biology, GBF—German Research Center for Biotechnology, Braunschweig, Germany

ABSTRACT

Understanding the complex interactions among cellular components (genes, proteins and metabolites) at a network level is a key issue in systems biology. In this chapter, we give an overview of metabolic network reconstruction from genome information and its structural analysis. First, existing databases for gene-enzyme and enzyme-reaction relationships needed for or applicable to the reconstruction of metabolic networks are discussed. Various approaches to reconstructing organism-specific metabolic networks are then briefly illustrated. The various means of mathematical representation of metabolic networks are explained, with particular emphasis on the problem arising from currency metabolites.

In the second part of the chapter, we summarize and discuss some major results of structural analysis of large-scale metabolic networks. Comparative analysis of a large number of fully sequenced organisms has revealed several intriguing topological properties, such as the power law connection degree distribution and the “bow-tie” global connectivity structure, which are explained as fundamental organizational principles of both biological and physical networks. Finally, we show an example of how structural analysis can be used for functional analysis of metabolic networks, especially for a modular network analysis, along with the challenges that face us for a more integrated and functional analysis of metabolic networks at a genome level.

I. INTRODUCTION

One of the key issues in systems biology is to decipher the metabolic and regulatory networks involved in cellular processes. The rapid development in genome

sequencing and functional genomic studies provides a large amount of information on the constituents (genes, mRNA, proteins/enzymes, and metabolites) of these biological networks and their activities in different organisms and under different environmental conditions. This makes it feasible to understand cell physiology and to compare different organisms at a system level. To this end, the reconstruction and analysis of genome-wide metabolic networks is of particular importance because it ultimately determines the metabolic activities and thus the physiology of cells. In fact, the study of genome-scale metabolic networks has gained much attention in recent years (Jeong et al. 2000; Wagner and Fell 2001; Ravasz et al. 2002; Stelling et al. 2002; Palsson et al. 2003; Forster et al. 2003; Ma and Zeng 2003a; Covert et al. 2004; Hatzimanikatis et al. 2004).

In this chapter, we first illustrate the major approaches and available databases for the reconstruction of genome-scale metabolic networks. We then introduce various means of mathematical representation of metabolic networks. Subsequently, methods for the structural and functional analysis of metabolic networks are explained and discussed. Emphasis is placed on methods based on graph theory and found useful in deciphering the global organization principle and the local modular hierarchical structure of metabolic networks. For a more pathway-orientated and stoichiometric analysis of metabolic networks, readers are referred to several recent excellent reviews on this aspect (Klamt and Stelling 2003; Palsson et al. 2003; Hatzimanikatis et al. 2004; Papin et al. 2004).

II. RECONSTRUCTION AND REPRESENTATION OF METABOLIC NETWORKS

A. Genome-scale metabolic network reconstruction: from parts to the whole

An important step in the reconstruction of organism-specific metabolic networks from genome information is to obtain the gene-enzyme and enzyme-reaction relationships. Enzymes are the key players that link a gene with a specific metabolic reaction. The IUBMB (International Union of Biochemistry and Molecular Biology) assigns an Enzyme Commission classification number (EC number, such as 1.1.1.1) to each enzyme. The EC number provides a unique and consistent representation of an enzyme and thus is widely used in genome annotation and metabolic network reconstruction.

There are several publicly available enzyme databases, such as BRENDA (Schomburg et al. 2004), KEGG (Kanehisa et al. 2004), and ExPASy enzyme nomenclature database (Gasteiger et al. 2003), which describe not only the basic information of an enzyme but the genes found to code for that enzyme. These databases can be used to obtain the gene-enzyme relationships. For example, part of the information for the enzyme 1.1.1.81 in the KEGG enzyme database follows.

```
ENTRY EC 1.1.1.81
NAME hydroxypyruvate reductase
beta-hydroxypyruvate reductase
```

NADH:hydroxypyruvate reductase
D-glycerate dehydrogenase
REACTION D-glycerate + NAD(P) = hydroxypyruvate + NAD(P)H
GENES YPE: YPO2536
PAE: PA1499
PPU: PP4300
RSO: RS03094(ttuD1) RS05749(ttuD2)
NEU: NE2456(ttuD2)
MLO: mlr5146

The corresponding genes in different organisms for the enzyme are listed in the section "GENES." From this information, one can know if an enzyme is coded in a specific organism. Therefore, one can obtain the lists of enzymes for all organisms in the database at one time.

For information on a newly sequenced organism that has not yet been included in the enzyme databases one can obtain an enzyme list for it directly from its genome annotation information. Based on sequence similarity or protein domain (motif) analysis, many genes (ORFs) in the genome are annotated as enzymes and assigned EC numbers. Then, from the annotation information one can know which enzymes are in the metabolic network of the organism.

It is often not straightforward to reconstruct the metabolic network from the obtained enzyme list for a specific organism because there are often no simple one-to-one enzyme-reaction relationships. One enzyme may catalyze several different reactions, and the same reaction may be catalyzed by different enzymes. For example, the enzyme fatty-acid synthase (2.3.1.85) catalyzes about 30 reactions in the fatty acid synthesis pathway, whereas the reaction



can be catalyzed by five different enzymes (2.3.1.85, 2.3.1.86, 4.2.1.58, 4.2.1.60, and 4.2.1.61). The various enzymes may exist in different organisms or be active under different environmental conditions. Unfortunately, in most enzyme databases only the main reaction catalyzed is listed for each enzyme. Therefore, some reactions that happen in reality may not be included in the reconstructed metabolic network. As far as we are aware, the KEGG LIGAND database is the most complete metabolic reaction database (Goto et al. 2002). It includes more than 6,000 enzyme-catalyzed or nonenzyme-catalyzed biochemical reactions. Most of the known reactions catalyzed by a specific enzyme are listed, allowing for reconstruction of more complete metabolic networks. However, an important source of information missing in the LIGAND database (and most other reaction databases) is the reaction reversibility.

It is generally recognized that many important metabolic reactions only occur in one direction under real physiological conditions. Therefore, information on reaction reversibility is important in network analysis. However, there is no metabolic reaction database available that gives clear and sufficient information about it. The

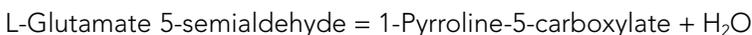
reaction direction is shown in the KEGG metabolic maps (direction inconsistencies in different maps exist). However, this information is not included in the reaction database file. Ma and Zeng (2003a) manually checked the reactions in the KEGG LIGAND database and added the reversibility information according to certain physiological rules. For example, one rule is that all oxygen-consuming reactions in organisms are irreversible.

Based on these rules, about 2,000 reactions are identified as irreversible. It is found that some irreversible reactions have a wrong reaction direction in the KEGG LIGAND reaction database. We corrected the wrong direction for these reactions. In addition, we also corrected some mistakes in the original reaction database, such as inconsistencies in compound names and mistakes in the reaction equations. The complemented and corrected metabolic reaction database is freely available from our web site (genome.gbf.de/bioinformatics/). The enzyme-gene and reaction-enzyme relations in the database are continually updated by incorporating the most up-to-date gene annotation information from the newest KEGG database files.

The conventional method for reconstructing metabolic networks is based on annotated genome sequencing (e.g., via ORFs). Because annotation of genomes is time consuming and often only 50 to 60% of the sequences can be accurately annotated with the present techniques, the necessity of using annotated sequences means a time delay and incompleteness. To solve this problem, and to explore data from a large number of on-going sequencing projects, Sun and Zeng (2004) recently developed an algorithm called *IdentiCS* to identify protein-coding sequences and thus to reconstruct strain-specific metabolic networks directly from unfinished raw genomic data. Compared with the conventional method (which needs more than an 8x coverage of the genome sequences), our method needs only a 3 to 4x coverage for bacteria. The method is being extended to eucaryotes.

The metabolic network reconstruction methods described previously are based on enzyme and reaction databases and can be called high-throughput reconstruction because they only make use of information available from databases. This allows an automatic approach in reconstructing networks for several organisms at the same time. The general workflow of this reconstruction method is summarized in Figure 9.1 (the solid arrows). This high-throughput method is necessary for comparative analysis of large-scale metabolic networks. However, there is a trade-off between the high productivity and the high quality. For example, the networks reconstructed in such a high-throughput way may be not complete due to the following.

- There are some nonenzyme-catalyzed reactions occurring spontaneously in metabolic networks. For example, the reaction



is an important nonenzyme-catalyzed reaction in proline synthesis pathways. These reactions should be added to the reactions lists obtained from genome information to avoid artificial missing links in the reconstructed metabolic network.

- EC numbers are often used in linking an annotated gene with one or more metabolic reactions. However, only chemically well-characterized enzymes are given an EC number by IUBMB (International Union of Biochemistry and Molecular Biology). For this reason, many enzymes are often found to have an incomplete EC number (such as 1.2.-.-) in the genome annotation database. Such incomplete EC numbers appear in almost all metabolic maps in KEGG (including the well-studied glycolysis pathway). It is necessary to develop a set of new IDs for these unclear enzymes to correctly map a reaction to a gene.
- Many enzymes for which the reactions catalyzed have been experimentally determined are not found in any fully sequenced genomes. Among the 4,223 enzymes in KEGG database, 2,572 are not found to be coded by any gene in any fully sequenced organism. The reason for this may be that the functions of a large part of the genes in a genome are unknown. For this reason, Karp (2004) recently called for an Enzyme Genomics Initiative to find coding sequences for these enzymes.

To address the problems mentioned previously for the high-throughput network reconstruction, one needs to extend the network with reactions from biochemistry and physiological studies and inferred from the literature (as illustrated in Figure 9.1). This is typically desired for an in-depth functional analysis of the metabolic network of a specific organism. However, this reconstruction process is relatively time consuming. EcoCyc is a well-known metabolic database that provides a high-

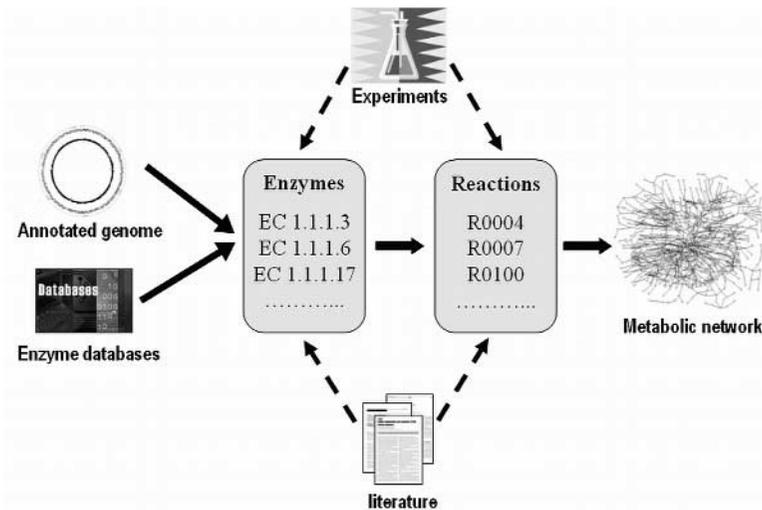


Figure 9.1. The reconstruction of metabolic networks from genome information. The high throughput reconstruction method (shown by the solid arrows) directly extracts information from enzyme or genome databases to obtain a list of reactions included in the metabolic network of one organism. The high-quality metabolic network reconstruction also adds new enzymes or reactions from biological experiments or literature (shown by the dashed arrows) in addition to those from databases.

quality metabolic network for *Escherichia coli* based on firsthand literature (Karp et al. 2000). High-quality metabolic networks for a small number of organisms (such as *Helicobacter pylori* and *Saccharomyces cerevisiae*) have been reconstructed by several groups (Schilling et al. 2002; Forster et al. 2003).

These well-defined metabolic models have been used to quantitatively analyze possible metabolic phenotypes of the organisms and/or to predict metabolic flux distributions under specific environmental conditions (Famili et al. 2003; Almaas et al. 2004). By integrating heterogeneous experimental data (such as those from microarray, proteomic, and metabolomic measurements), these models can be consolidated or improved by including newly discovered interactions (missing links in metabolic networks), which in turn can guide strain improvement process to reach a desired metabolic phenotype (Edwards et al. 2001; Ibarra et al. 2002; Covert et al. 2004).

B. Mathematical representation of metabolic networks

A proper mathematical representation of the large number of reactions obtained for a specific organism is necessary for any structural analysis of metabolic networks. Two approaches are generally used: the stoichiometric matrix and the connectivity graph (Figure 9.2). In the stoichiometric representation, the rows and columns of the so-called stoichiometry matrix represent reactions and metabolites, respectively. A cell with a nonzero value in the matrix represents the stoichiometric coefficient of the corresponding metabolite in the corresponding reaction. A positive value means that it is a product, whereas a negative value indicates a substrate. The stoichiometric representation is a full representation of the network structure. Several quantitative analysis methods have been developed based on the stoichiometric matrix of metabolic networks, such as flux balance analysis, elementary flux mode analysis, and extreme pathway analysis (Schuster et al. 1999; Edwards et al. 2002; Price et al. 2002; Klamt et al. 2003; Papin et al. 2004).

However, when dealing with large-scale genome-based metabolic networks these methods often face serious computational problems. For example, the combinatorial explosion problem resulting from huge numbers of pathways often makes it difficult or even impossible to calculate all elementary modes or extreme pathways in genome-scale metabolic networks (Klamt and Stelling 2002; Schuster et al. 2002). For a detailed description of these stoichiometric-matrix-based methods, one can refer to the chapter by Bruggemann et al. in this book.

In contrast to the stoichiometric representation, graph representation is a simplified way of representing the metabolic network. As shown in Figure 9.2, two types of graphs can be generated from a metabolic network: the metabolite graph (in which the nodes are metabolites and the links are reactions) and the reaction graph (in which the nodes are reactions and two reactions are linked if a metabolite is the substrate of one reaction and the product of another reaction). The metabolite graph is similar to the classical way of metabolic pathway illustration in biochemistry textbooks, and is thus often used in structural analysis of metabolic networks.

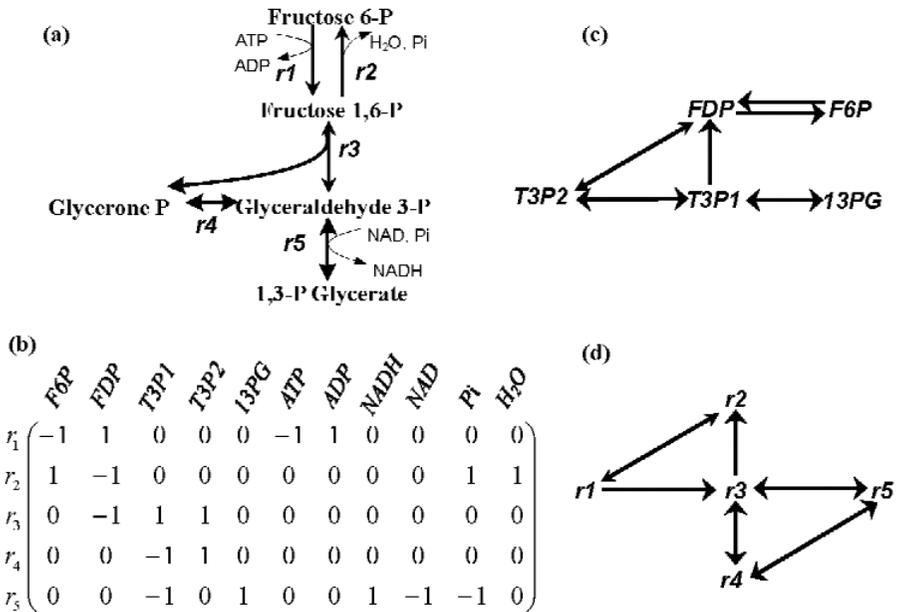


Figure 9.2. Mathematical representation of metabolic networks. (a) The upper part of the glycolysis pathway, (b) the stoichiometric matrix of the pathway, (c) the metabolite graph representation of the pathway, and (d) the reaction graph representation of the pathway. Metabolite abbreviations: F6P (D-Fructose 6-phosphate); FDP (D-Fructose 1,6-bisphosphate); T3P1 (D-Glyceraldehyde-3-phosphate); T3P2 (Glycerone phosphate); and 13PG (1,3-Bisphospho-D-glycerate).

Considering that many reactions are irreversible, many links in the graph are directed (called arcs in graph theory, and correspondingly the undirected links are called edges), resulting in a directed graph.

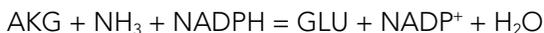
Compared with the stoichiometric representation, the graph representation is more suitable for visualization and structural analysis of large-scale metabolic networks. However, we should keep in mind that the graph representation is a simplified way of network representation that loses some information, such as stoichiometric coefficients of reactions. A reaction often has several links in the graph (sometimes in very different parts) because most reactions have multiple substrates and products. On the other hand, one link in the graph may represent several different reactions. Therefore, a reverse step to map a link to its corresponding reaction(s) is required when providing biological interpretation for the results from graph analysis of metabolic networks.

C. Currency metabolites in graph representation of metabolic networks

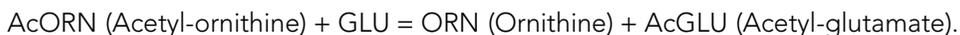
An important issue in graph representation of metabolic networks is how to deal with the currency metabolites such as H₂O, CO₂, ATP, and so on (Ma and Zeng

2003a). Currency metabolites are normally used as carriers for transferring electrons and certain functional groups (phosphate group, amino group, one carbon unit, methyl group, and so on). In a relatively early and most-often cited study on the structure of genome-scale metabolic networks based on graph theory, Jeong et al. (2000) regarded all metabolites (including currency metabolites) as nodes. In this way, they calculated one of the network topology parameters, average path length (APL), which is defined as the shortest path length averaged for every connected pair of metabolites in the entire network. They found that APL is almost the same (about 3.2) for all 43 organisms studied. This means that most of the metabolites can be converted to each other in about three steps. These results are surprising, and in fact unexpected, in view of the often long pathways for the synthesis of many metabolites. The reason for this unrealistic short path length is that most of the apparent shortest paths are actually linked through currency metabolites. For example, in the glycolysis pathway the path length (number of reaction steps in the pathway) from glucose to pyruvate should be nine in terms of biochemistry. However, if ATP and ADP are considered as nodes in the network, the path length between glucose and pyruvate becomes only two (the first reaction uses glucose and produces ADP, whereas the last reaction consumes ADP and produces pyruvate).

This calculation of path length is obviously biologically not meaningful. Therefore, the connections through currency metabolites should be avoided in finding the shortest path from one metabolite to another. It should be mentioned that currency metabolites cannot be defined per se by compounds but should be defined according to the reaction. For example, glutamate (GLU) and 2-oxoglutarate (AKG) are currency metabolites for transferring amino groups in many reactions, but they are primary metabolites in the following reaction.



The connections through them should be considered. The same situation holds for NADH, NAD⁺, ATP, and so on. Another problem involves reactions such as the following.



Here, the acetyl group is transferred between GLU and ORN. Only the connections AcORN-ORN and GLU-AcGLU are included, but AcORN-AcGLU and GLU-ORN are excluded. If the latter two connections are considered, the path length from GLU to ORN will be one, and this is not in accordance with the pathway in real biochemistry.

From this discussion we can see that it is difficult to remove the connections through currency metabolites automatically by a program. Therefore, in a recent study (Ma and Zeng 2003a) we manually checked the reactions that appear in the KEGG metabolic maps and added corresponding connections one by one. In this way, the reaction-connection relationships can be more accurately obtained and used to generate metabolite graphs from the lists of reactions of different organ-

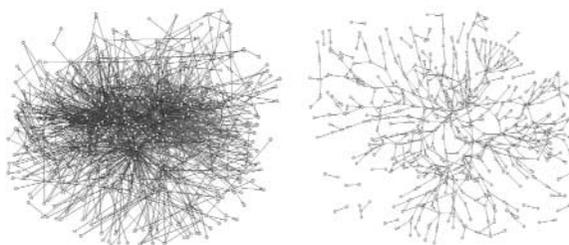
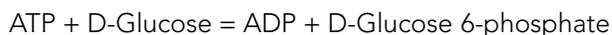


Figure 9.3. The metabolite graph representation of metabolic networks of *Streptococcus pneumoniae*. The left-hand network includes the connections through currency metabolites, and the right-hand network does not. Links with arrows represent irreversible reactions, and those without arrows represent reversible reactions (see color plate 5).

isms. As an example, Figure 9.3 depicts the two graphs (with and without connections through currency metabolites) for the reconstructed metabolic network of *Streptococcus pneumoniae*. It can be seen that the one without currency metabolites is more realistic and more amenable to analysis. In contrast, the true network structure in the graph with currency metabolites is masked by the large number of links through currency metabolites. Therefore, the removal of connections through currency metabolites is an essential step in drawing biologically meaningful conclusions from graph analysis of metabolic networks.

Arita (2003) proposed a different approach, called atomic reconstruction of metabolism for graph representation of metabolic networks. In this approach, the atomic flow in a metabolic reaction is traced and a substrate is only connected to the product(s) that contains at least one atom from it. An example is shown here for the following reaction.



In this reaction, the link from D-glucose to ADP is not included in the graph because there is no atomic flow between these two metabolites. However, the other three links (ATP to ADP, ATP to D-glucose 6-phosphate, and D-glucose to D-glucose 6-phosphate) are included in the resulting graph. Therefore, although this approach can avoid certain connections through currency metabolites there are still biologically not meaningful connections in the graph.

III. STRUCTURAL ANALYSIS OF METABOLIC NETWORKS

A. Degree distribution and average path length

An important structure characteristic of metabolic networks and many other complex networks is the power law degree distribution: most of the nodes in the network have a low connection degree, whereas few nodes have a very high

connection degree (Albert et al., 2000; Jeong et al. 2000, 2001; Strogatz 2001; Wolf et al. 2002; Bray 2003). The high-degree nodes dominate the network structure and are called hubs of the network. Most of the nodes are connected through the hubs by a relatively short path, and the average path length is insensitive to the network scale. Therefore, this type of network is called a scale-free network in several studies (Strogatz 2001). The scale-free property makes the network robust against random errors because most errors on the less connected nodes do not affect the network connectivity very much. Therefore, such a robust structure may be the result of a long evolutionary selection process.

The metabolite graphs (with or without connections through currency metabolites) were found to follow the power law degree distribution, implying that both are scale-free networks (Jeong et al. 2000; Ma and Zeng 2003a). However, the hubs identified for the graphs with or without currency metabolites are very different. Most of the hub metabolites in the metabolite graph with currency metabolites are currency metabolites such as H_2O , ATP, ADP, and so on due to their frequent appearance in many reactions. Excluding these currency metabolites, one normally finds several major primary metabolites as hubs. These include glycerate-3-phosphate, pyruvate, and D-fructose-6-phosphate and D-glyceraldehyde-3-phosphate (which are intermediates in the glycolysis pathway); D-ribose-5-phosphate and D-xylulose-5-phosphate (which are intermediates in the pentose phosphate pathway); acetyl-CoA (which is the metabolite linking the glycolysis pathway, the citric acid cycle, and the fatty acid synthesis pathway); 5-phospho-D-ribose 1-diphosphate (which is the precursor for purine and histidine synthesis); and L-glutamate and L-aspartate (two important amino acids directly produced from precursors in the citrate acid cycle and convertible to many other amino acids). These metabolites are in the central metabolic network across organisms (Schilling et al. 2002; Stelling et al. 2002) and thus are the true hubs in the organization of metabolic networks.

One feature of the scale-free network is the somehow invariable average path length with increasing network size. This phenomenon has been found by Jeong et al. (2000) in the metabolite graphs with currency metabolites (about 3.2 for all 43 studied organisms). However, for metabolite graphs without currency metabolites much longer and variable path lengths are obtained (as shown in Figure 9.4).

Generally, APL tends to increase with network scale. Furthermore, quantitative differences exist among the three domains of organisms; namely, the metabolic networks of eukaryotes and archaea generally have a longer APL than those of bacteria. The average APL values for networks of these three domains of organisms are 9.57, 8.50, and 7.22, respectively. This result indicates that there are true structure differences between the metabolic networks of different organisms, which can only be revealed by removing the connection through currency metabolites. The network structural differences are the result of a long evolutionary process.

To explore this, we constructed evolutionary trees based on the reaction content of metabolic networks for 82 fully sequenced organisms (Ma and Zeng 2004). We found that the major results from phylogenetic trees based on metabolic networks are surprisingly in good agreement with the tree based on 16S rRNA, despite the

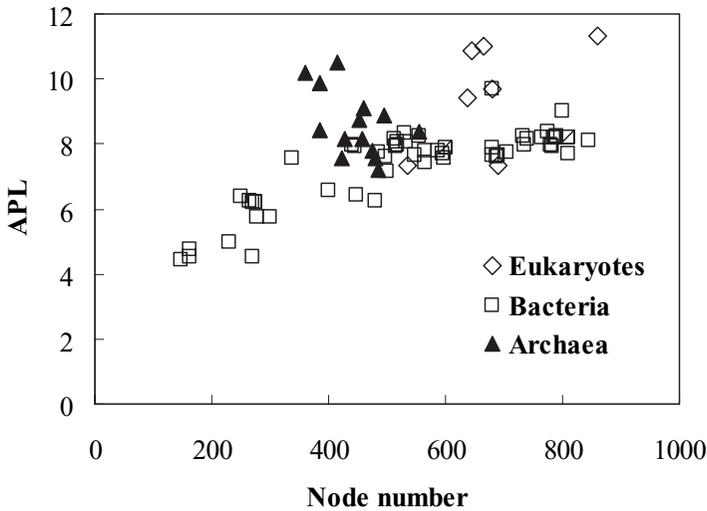


Figure 9.4. Calculated average path lengths for metabolic networks of fully sequenced organisms.

prevalence of horizontal transfer of metabolic genes among organisms—confirming the three-domain classification and the close relationship between eukaryotes and archaea at the level of metabolic networks. This indicates that the gene transfer events are constrained by some system-level organizational principle(s).

B. Network global connectivity: the “bow-tie” structure

The scale-free property revealed by the power law connection degree distribution is regarded as an important finding in the study of complex networks, and has been found in many different types of networks (Albert et al. 2000; Jeong et al. 2000; Jeong et al. 2001; Strogatz 2001). However, this property only reflects one aspect of the network structure. Actually, it only shows the local connectivity of a network but does not tell us anything about the global network structure. For example, both networks shown in Figure 9.5 indicate a power law degree distribution. However, the left-hand network is a fully connected one, whereas the right-hand network consists of several disconnected subgraphs. The same problem exists for the parameter average path length.

Normally, a short APL means that the network is more efficiently connected. However, the right-hand network shown in Figure 9.5 has a shorter APL than the left-hand network, though most of the nodes in the right-hand network are not connected with each other at all. Therefore, new method(s) and parameter(s) are needed to investigate the global network connectivity, which cannot be described by the parameters degree distribution and APL. To this end, we used the breadth-

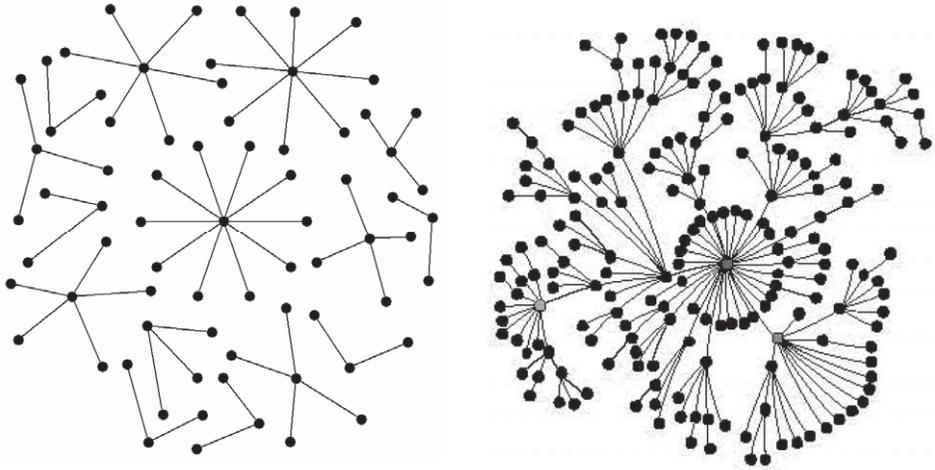


Figure 9.5. Two simple network examples showing the limitation of connection degree distribution. Both networks show power law degree distributions, but apparently have different network connectivity.

first searching method to find all connected pairs of metabolites (Broder et al. 2000). We found that in most of the metabolic networks about half the metabolites can be converted to only a very limited number (usually less than 10) of metabolites. Although the number of metabolites reachable by the other metabolites is much higher, it is still not more than half the metabolites. For a randomly chosen pair of substrate and product, the probability that a path exists between them is less than 20% (10% for metabolic networks of certain organisms).

This indicates that metabolic networks are far from fully connected networks. At the same time, we found that there exist several fully connected sub-networks in which all metabolites can be converted to each other. These fully connected sub-networks are called strong components of the metabolic network (Ma and Zeng 2003b). In graph theory, a strong component of a network is defined as a subset of nodes such that for any pair of nodes u and v in the subset there is a path from u to v (Batagelj and Mrvar 1998). The size distribution of the strong components in the metabolic network of *E. coli* is shown in Figure 9.6. It can be seen that the largest component is much larger than other components and thus is called the "giant strong component (GSC)." Then we found that there are an IN subset in which all metabolites can be converted to metabolites in the GSC, and an OUT subset in which all metabolites can be produced from metabolites in the GSC. All other metabolites not connected with metabolites in the GSC form an isolated subset (IS).

In this way, we obtained a "bow-tie" connectivity structure of metabolic networks, as shown in Figure 9.7. This bow-tie connectivity structure was found in the metabolic networks of all other organisms studied. The bow-tie structure has also been found in the web page graph in which web pages represent nodes and hyperlinks

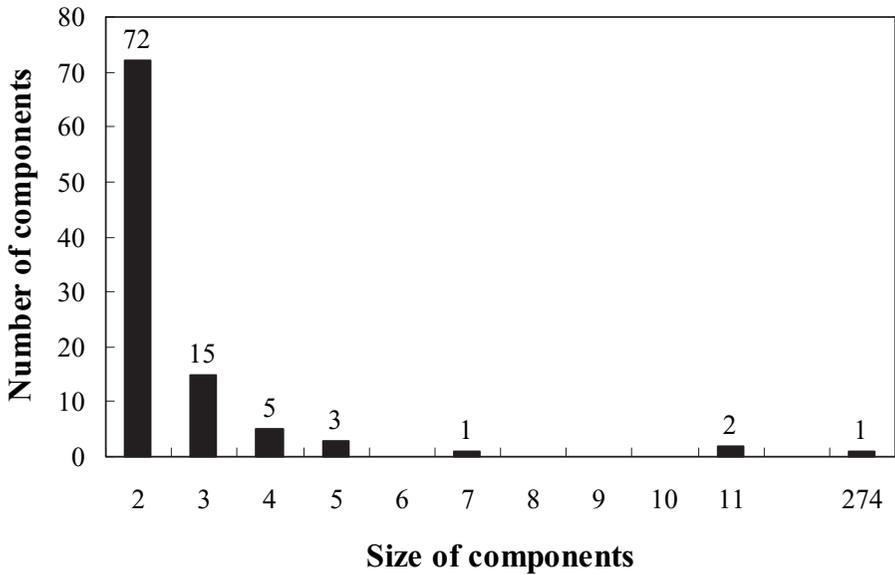


Figure 9.6. The distribution of the size of the strongly connected components in the metabolic network of *E. coli*.

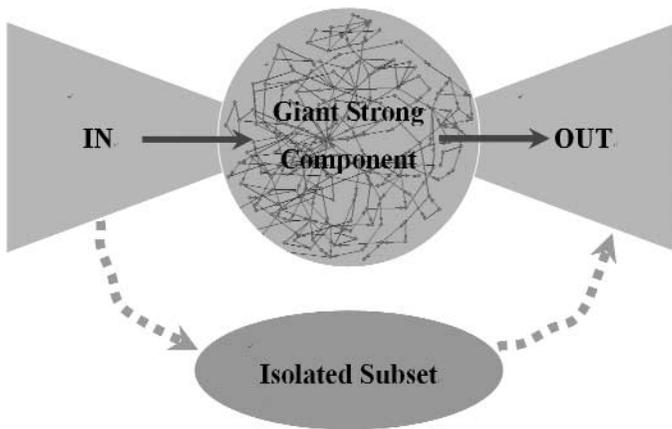


Figure 9.7. The bow-tie connectivity structure of metabolic networks (see color plate 6).

represent links (Broder et al. 2000). The discovery of the bow-tie structure in different types of networks implies that it is a common structure in large-scale networks. Organization as a bow-tie may be important for the complex system to be robust and evolvable under variable and undetermined environmental conditions (Csete and Doyle 2004; Kitano 2004). The core part of the network (the GSC), which

represents the universal mechanism across organisms, may be highly conserved in the evolutionary process. On the other hand, new biological functions such as utilizing a new substrate or producing a new product can be evolved by innovation in the IN and OUT subsets.

C. GSC in the bow-tie structure

The giant strong component is the most complex and core part of a metabolic network. We found that the GSC follows a power law connection degree distribution similar to that of the entire network. Furthermore, the average path length of the entire network (ALW) was found to be determined by that of the GSC (ALG), as depicted in Figure 9.8. Because of the large scale, it is often difficult to achieve a comprehensive understanding of biological features of genome-based metabolic networks. A certain form of reduction or classification of the entire network is desired to make the network more amenable to functional analysis. The connectivity structure of metabolic networks as revealed in Figure 9.7 represents a step forward in this direction.

The most important part of the network, the GSC, normally contains less than one-third of the nodes of the entire network but conserves the main features of the entire network. Here, we use *S. pneumoniae* (an important Gram-positive pathogen) as an example to analyze the functional feature of its network structure. The entire network of *S. pneumoniae* consists of 486 metabolites, whereas its GSC contains only 87 metabolites. To further reduce the complexity of the GSC, we

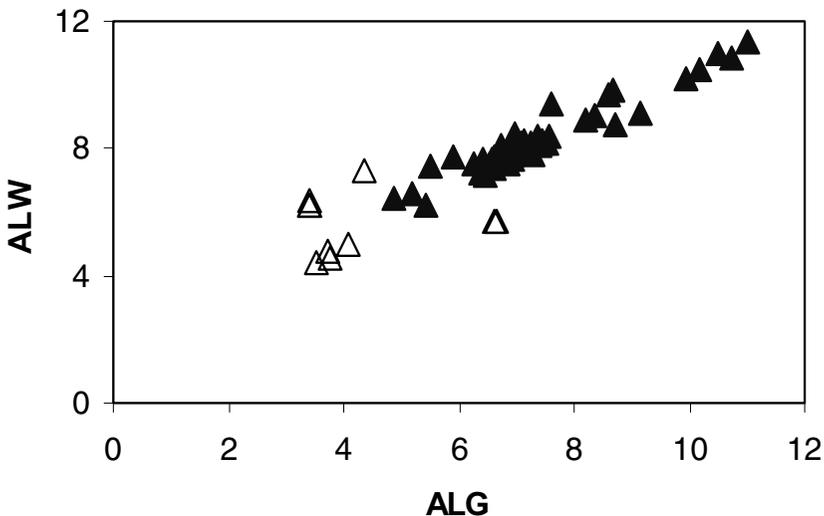


Figure 9.8. The relationship between the average path length of the entire network and that of the GSC.

removed the linear branches (the endpoint of which has only one connection and no branch point in the path) in the GSC. The resulting network is shown in Figure 9.9. Five major metabolic pathways (not necessarily complete due to omission of linear branches) can be identified in the core network of *S. pneumoniae*.

These include the glycolysis pathway, the pentosephosphate pathway, the aromatic amino acid synthesis pathway, the glycerol metabolism, and the pyrimidine synthesis pathway. There are also parts of lysine synthesis, valine synthesis, oxaloacetate anapleurotic, and Entner-Doudoroff (ED) pathways in the core network. These pathways are integrated into a network through certain metabolites such as pyruvate (PYR), 5-Phosphoribosyl diphosphate (PRPP), and D-glyceraldehyde phosphate (G3P). All of these metabolites belong to the hub metabolites in the metabolite graph without connections through currency metabolites (Ma and Zeng 2003a). As links between the different functional systems, these metabolites play a key role in metabolic regulation.

IV. FROM NETWORK TO MODULES AND FUNCTIONAL ANALYSIS

The uncovering of the bow-tie structure of metabolic networks has important implications for biotechnology and biomedicine. For example, understanding and manipulating the distribution and control of metabolic fluxes over the metabolic network are key steps in metabolic engineering of organisms and therapy of certain metabolic diseases. However, for large-scale metabolic networks the estimation of metabolic flux and control can be very difficult or even impossible. A reduction of the metabolic network is almost always necessary.

The GSCs of organisms contain a much smaller number of (albeit key) metabolites. GSCs are more feasible for analysis of flux distribution and identification of all possible elementary flux modes or extreme pathways. The distribution of metabolic fluxes is mainly controlled by regulating the flux ratio at branch-points. Most of the branch-points are in the GSC. Therefore, one may focus on the GSC when studying the flux distribution and its regulation in metabolic networks. This can largely simplify the analysis process.

For large-scale metabolic networks such as that of *E. coli*, even the GSC is still quite complex for obtaining a functional overview from the structure (as is that for *S. pneumoniae*). In this case, a top-down approach to decompose it into relatively independent functional subsets or modules is often necessary for further biological functional analysis (Bray 2003). In biochemistry, it is generally accepted that metabolic networks consist of many functionally independent metabolic pathways that are further nested to form a complex metabolic network (Hartwell et al. 1999).

Organizing the reactions into different metabolic pathways is widely used in many metabolic databases, such as KEGG and Metacyc (Karp et al. 2000; Kanehisa et al. 2004). These classically defined metabolic pathways can be regarded to a certain extent as functional modules in metabolic networks. However, it is difficult to see

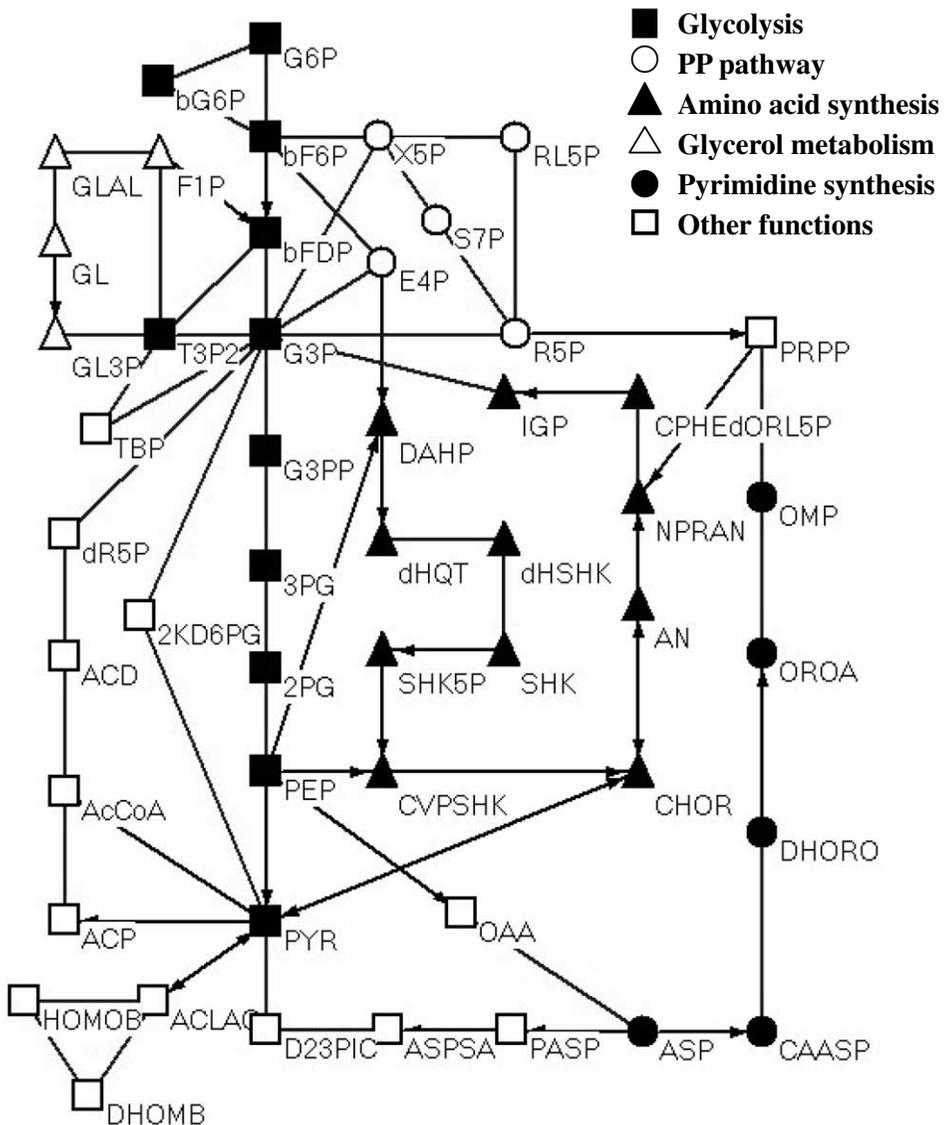


Figure 9.9. The giant strong component in the metabolic network of *Streptococcus pneumoniae*. Metabolite abbreviations: 2KD6PG (2-Dehydro-3-deoxy-6-phospho-D-gluconate); 2PG (Glycerate 2-phosphate); 3PG (Glycerate 3-phosphate); AcCoA (Acetyl-CoA); ACD (Acetaldehyde); ACLAC (2-Acetolactate); ACP (Acetyl phosphate); AN (Anthranilate); ASP (L-Aspartate); ASPSA (Aspartate semialdehyde); bF6P (beta-D-Fructose 6-phosphate); bFDP (beta-D-Fructose 1,6-bisphosphate); bG6P (beta-D-Glucose 6-phosphate); CAASP (N-Carbamoyl-ASP); CHOR (Chorismate); CPHEdORL5P (1-(2-Carboxyphenylamino)-1-deoxy-D-ribose 5-phosphate); CVPSHK (5-O-(1-Carboxyvinyl)-3-phosphoshikimate); DAHP (2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate); DHOMB ((R)-2,3-Dihydroxy-3-methylbutanoate); DHORO ((S)-Dihydroorotate); dHSHK (3-Dehydroshikimate); E4P (D-Erythrose 4-phosphate); F1P (Fructose 1-phosphate); G3P (D-Glyceraldehyde 3-phosphate); G3PP (1,3-Bisphospho-D-glycerate); G6P (D-Glucose 6-phosphate); GL (Glycerol); GL3P (Glycerol 3-phosphate); GLAL (D-Glyceraldehyde); HOMOB ((R)-3-Hydroxy-3-methyl-2-oxobutanoate); IGP (Indoleglycerol phosphate); NPRAN (N-(5-Phospho-D-ribosyl)anthranilate); OAA (Oxaloacetate); OMP (Orotidine 5'-phosphate); OROA (Orotate); PASP (Aspartate phosphate); PEP (Phosphoenolpyruvate); PRPP (5-phospho-D-ribose 1-diphosphate); PYR (Pyruvate); R5P (Ribose 5-phosphate); RL5P (Riblose 5-phosphate); S7P (D-Sedoheptulose 7-phosphate); SHK (Shikimate); SHK5P (Shikimate 5-phosphate); T3P2 (Glycerone phosphate); TBP (D-Tagatose 1,6-bisphosphate); and X5P (D-Xylulose 5-phosphate) (see color plate 7).

any modular organization from the previously described structural analysis. In contrast, the scale-free structure and the short path length seem to suggest that the metabolites in the network are highly interactive, making it difficult to identify any structurally independent modules. To resolve the contradiction between the scale-free structure and the modular organization of pathways, we proposed a method based on the global connectivity structure of metabolic networks to decompose the entire network into several subgraphs to check if these subgraphs are functionally independent modules. For network decomposition, we used reaction graph rather than metabolite graph. This allowed us to classify reactions (but not metabolites) into different modules as being in agreement with the traditional defined metabolic pathways as subsets of reactions. In a similar way, as in the analysis of metabolite graphs, the connections through currency metabolites were removed. The reaction graphs also show the power law degree distribution and the bow-tie connectivity structure. The decomposition method is based on the bow-tie structure and includes the following steps (see Ma et al. (2004b) for a detailed description): (1) calculating the path length between two reactions in GSC and using it as the distance between the two reactions, thus obtaining a distance matrix for all reaction in GSC, (2) a hierarchical classification tree is constructed from the distance matrix by using neighbor-joining or other algorithms, (3) cutting the tree at different levels to obtain modules in proper size (about 20 reactions), and (4) assign the reactions in the IN and OUT subsets of the bow-tie structure to different modules, depending on which modules they are highly connected with.

The method is applied to the decomposition of the metabolic network of *E. coli*. The decomposition result is shown in Figure 9.10. We checked the reactions in the

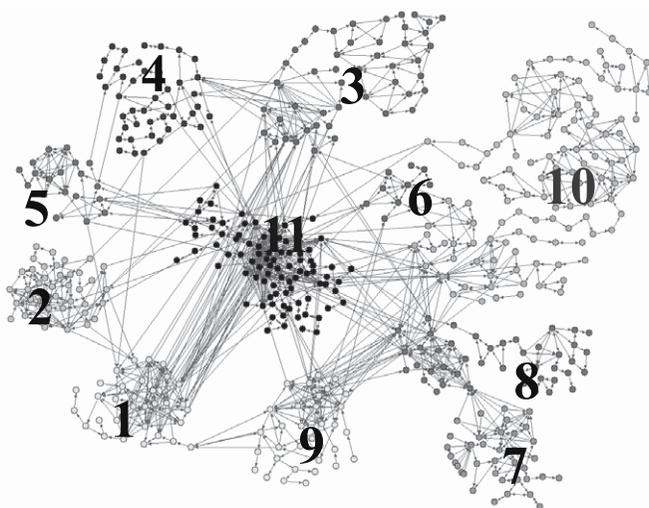


Figure 9.10. Functional modules in the metabolic network of *E. coli* obtained by network decomposition.

Table 9.1. Modules obtained by network decomposition in the genome-based metabolic network of *E. coli*.

Subset	Function	Number of reactions	Number of metabolites
1	Acetyl-CoA, Succinyl-CoA metabolism	22	53
2	Glutamate and glutamine metabolism, urea cycle, arginine, proline synthesis	38	61
3	Oxaloacetate metabolism, pyrimidine synthesis	47	61
4	Propanoyl-CoA metabolism, threonine, methionine and lysine synthesis	43	79
5	Glutathione metabolism	19	42
6	Glycerate and galactarate metabolism	10	19
7	Glucose, galactose and nucleotide sugar metabolism	37	55
8	Fructose and mannose metabolism, aromatic amino acid synthesis	44	65
9	Glycerone phosphate and glycerolipid metabolism	33	50
10	Pentose phosphate pathway, purine, folate and riboflavin synthesis	114	123
11	Pyruvate metabolism, glyoxylate metabolism, valine, leucine, isoleucine synthesis	89	134

KEGG pathway map and found that most of the reactions in a classical pathway are also in the same module. The major biological functions of the reactions in these modules are listed in Table 9.1. The results showed that the subgraphs identified from the network structure are really functional modules, indicating that the modular organization is also an inherent feature of the metabolic network. However, we also surprisingly found that three classical pathways in the central metabolism (the glycolysis pathway, pentose phosphate pathway, and citrate acid cycle) are split into parts in different modules. However, this is consistent with the organization synopsis of the *E. coli* metabolic network proposed by Gagneur et al. (2003).

One possible explanation is that the metabolites in these central pathways are used as precursors for the synthesis of different products and are thus placed in different subsets. For example, for the reactions in the TCA cycle the reaction from isocitrate to oxoglutarate is in module 2 because oxoglutarate is the precursor of the glutamate family amino acid. The reactions from malate to isocitrate are in module 4 because oxaloacetate is the precursor for aspartate family amino acids and aspartate is then used for pyrimidine synthesis. For the reactions in the pentose phosphate pathway, all erythrose 4-phosphate related reactions are in module 8 because it is one of the precursors for aromatic amino acid synthesis. All ribose 5-

phosphate related reactions are in module 10 because it is the precursor for purine synthesis.

V. CONCLUSIONS

One of the goals of systems biology is to develop theoretical models to describe and predict cellular behavior at the whole-system level. The structural and functional analysis of genome-based metabolic networks described in this chapter represents one step toward this goal. The macroscopic structure of the metabolic network (scale-free, bow-tie, modular organization), which can only be uncovered by analysis of the network as a whole, indicates certain system-level principles governing the organization of interacting cellular components (enzymes and metabolites). Although these structure properties still merely give a static picture of the metabolic network, they can serve as a basis or blueprint for analyzing the dynamic behavior of the network (e.g., information and material flows)—the next necessary and more demanding step in network analysis.

To this end, the metabolic network model needs to be further extended. In particular, transcriptional regulatory interactions should be integrated into the metabolic network. Most of the metabolic genes are regulated by one or more transcriptional factors and are activated/repressed under different environmental conditions. Therefore, by integrating the regulatory relationships one may predict which reactions (pathways) are activated or repressed under given environmental conditions. One of the challenges in this endeavor is to establish genome-scale regulatory networks. So far, our knowledge on regulatory interactions at the genome level is still limited to a few model organisms, such as *E. coli* (Ma et al. 2004a; Salgado et al. 2004) and *S. cerevisiae* (Luscombe et al. 2004).

The integration of functional genomic data (such as those from transcriptomic, proteomic, and metabolomic analyses) is also essential for functional and dynamic analysis of metabolic networks. These high-throughput technologies provide a means of measuring the expression or concentration levels of genes, proteins, and metabolites for the entire system. Combined with bioinformatics and systems biology tools, this wealth of data may allow us in the near future to reconstruct integrated metabolic and regulatory networks at different molecular levels and to understand their system-level interactions.

REFERENCES

- Albert, R., Jeong, H., and Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**:378–382.
- Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N., and Barabasi, A. L. (2004). Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**:839–843.
- Arita, M. (2003). *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Res.* **13**:2455–2466.

- Batagelj, V., and Mrvar, A. (1998). Pajek: Program for large network analysis. *Connections* **21**:47–57.
- Bray, D. (2003). Molecular networks: The top-down view. *Science* **301**:1864–1865.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the Web. *Computer Networks* **33**:309–320.
- Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**:92–96.
- Csete, M., and Doyle, J. (2004). Bow-ties, metabolism, and disease. *Trends Biotechnol.* **22**:446–450.
- Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**:125–130.
- Edwards, J. S., Covert, M., and Palsson, B. (2002). Metabolic modelling of microbes: The flux-balance approach. *Environ. Microbiol.* **4**:133–140.
- Famili, I., Forster, J., Nielsen, J., and Palsson, B. O. (2003). *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* **100**:13134–13139.
- Forster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**:244–253.
- Gagneur, J., Jackson, D. B., and Casari, G. (2003). Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* **19**:1027–1034.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucl. Acids. Res.* **31**:3784–3788.
- Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). LIGAND: Database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30**:402–404.
- Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**:C47–C52.
- Hatzimanikatis, V., Li, C., Ionita, J. A., and Broadbelt, L. J. (2004). Metabolic networks: Enzyme function and metabolite structure. *Curr. Opin. Struct. Biol.* **14**:300–306.
- Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**:186–189.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**:651–654.
- Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**:41–42.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucl. Acids. Res.* **32**:D277–D280.
- Karp, P. D. (2004). Call for an enzyme genomics initiative. *Genome Biology* **5**:401.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **28**:56–59.
- Kitano, H. (2004). Biological robustness. *Nat. Rev. Genet.* **5**:826–837.
- Klamt, S., and Stelling, J. (2002). Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.* **29**:233–236.
- Klamt, S., and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends Biotechnol.* **21**:64–69.
- Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**:308–312.

- Ma, H. W., and Zeng, A. P. (2003a). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**:270–277.
- Ma, H. W., and Zeng, A. P. (2003b). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19**:1423–1430.
- Ma, H. W., and Zeng, A. P. (2004). Phylogenetic comparison of metabolic capacities of organisms at genome level. *Molecular Phylogenetics and Evolution* **31**:204–213.
- Ma, H. W., Kumar, B., Ditges, U., Gunzer, F., Buer, J., and Zeng, A. P. (2004a). An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res.* **32**:6643–6649.
- Ma, H. W., Zhao, X. M., Yuan, Y. J., and Zeng, A. P. (2004b). Decomposition of metabolic network based on the global connectivity structure of reaction graph. *Bioinformatics* **20**:1870–1876.
- Palsson, B. O., Price, N. D., and Papin, J. A. (2003). Development of network-based pathway definitions: The need to analyze real metabolic networks. *Trends Biotechnol.* **21**:195–198.
- Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends Biotechnol.* **22**:400–405.
- Price, N. D., Papin, J. A., and Palsson, B. B. (2002). Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.* **12**:760–769.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**:1551–1555.
- Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., and Collado-Vides, J. (2004). RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**:D303–D306.
- Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002). Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**:4582–4593.
- Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: Updates and major new developments. *Nucleic Acids Res.* **32**:D431–D433.
- Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**:53–60.
- Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2002). Exploring the pathway structure of metabolism: Decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* **18**:351–361.
- Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**:190–193.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature* **410**:268–276.
- Sun, J., and Zeng, A. P. (2004). IdentiCS: Identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics* **5**:112.
- Wagner, A., and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. R. Soc. Lond. B. Biol. Sci.* **268**:1803–1810.
- Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: New insights into the fundamentals of evolution? *Bioessays* **24**:105–109.

Integrated Regulatory and Metabolic Models

Markus W. Covert

California Institute of Technology, Pasadena, California, USA

Chapter 10

ABSTRACT

This chapter describes how to reconstruct functional metabolic and transcriptional regulatory networks, as well as the modeling approaches that allow for simulation of network behavior for networks separately and for networks combined. This process is placed in the context of model-driven biological discovery, and is illustrated with a detailed case study. In this study, a genome-scale model was reconstructed and used in conjunction with experimental data to elucidate the regulatory and metabolic networks in *Escherichia coli*.

I. INTRODUCTION

A major goal of systems biology is to further our understanding of complex biological systems. Using systems biology to facilitate biological discovery may be thought of as a simple expansion of traditional biology, as shown in Figure 10.1. Traditional biology (shaded box) begins with an experimental system of interest. The “inputs” to the system are simply aspects of the system that can be controlled. Thus, the inputs may be external (such as environmental conditions) or internal, such as perturbations to the genetic makeup of the organism (gene knock-outs or knock-ins). The “outputs” to the system are aspects that are changed by the system itself and that are measurable. Outputs can also be external (such as the concentration over time of secreted by-products or biomass) or internal, such as the differential expression of genes or activity of regulatory proteins. If the experiment is

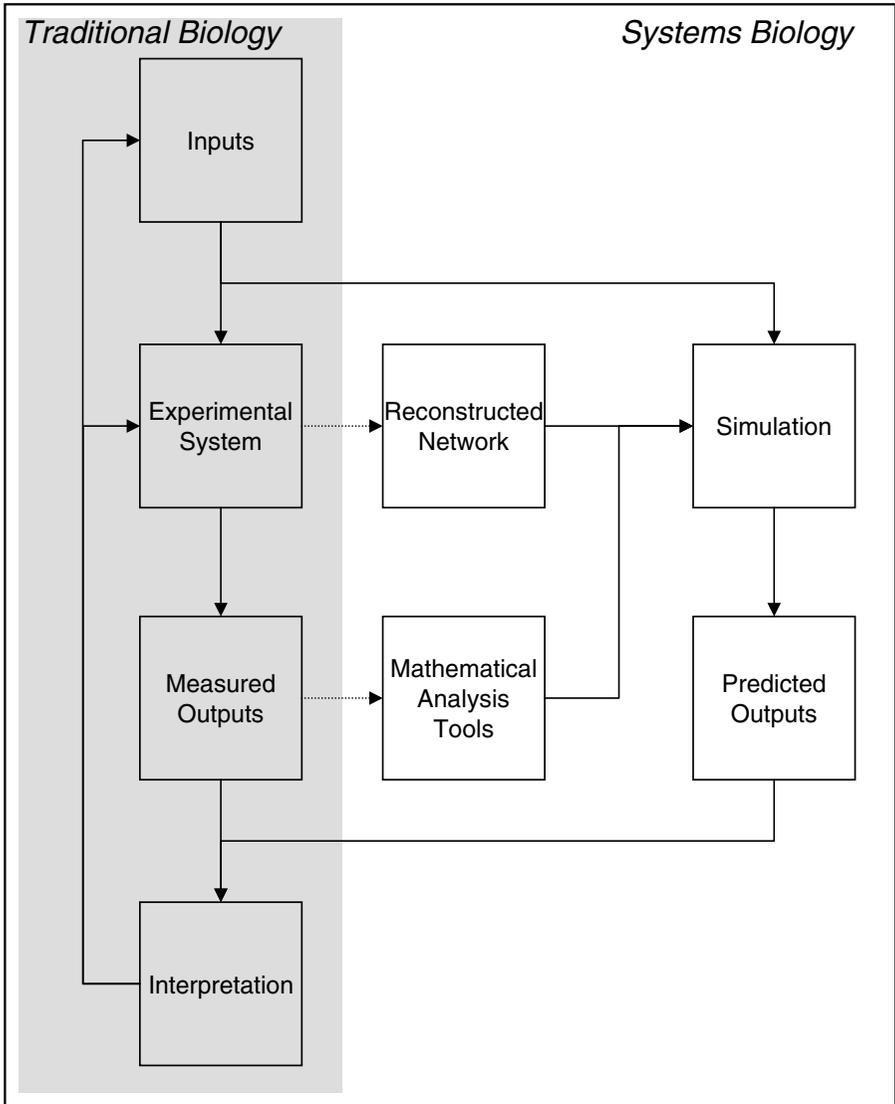


Figure 10.1. The traditional and systems biology approaches to discovery. The systems approach can greatly expedite the discovery process by incorporating the testable predictions of a mathematical model.

well-designed, the investigator can make an interpretation of the measured outputs that (1) gives new insight into the system and (2) suggests new perturbations to the inputs for a subsequent experiment.

This process has had great success over the last several decades and is the foundation for all of the biological knowledge we have. However, it now has the poten-

tial to be greatly enhanced by two major factors. First, the development of high-throughput technologies means that we are now able to vary the inputs and measure outputs many thousandfold faster than before (although arguably with different accuracy). The result is a combinatorial explosion of data that would be impossible to interpret without the aid of a computer. The development of mathematical modeling tools is the second factor, enabling a much more rapid characterization of biological systems.

How do these two factors influence traditional biology? First, experimental systems can be studied more broadly, with much the same detail. Instead of looking at one small part of the organism, we can consider an entire network. Metabolism and transcriptional regulation are currently the networks most feasible, but there is every reason to believe that others (signal transduction networks, for example) will follow. The annotated genome sequence enables us to obtain most of the components of the network, although a substantial minority of components must still be obtained from the traditional biology literature. The type of measured output should drive the choice of mathematical analysis tools, as the predictions made by a mathematical model are of much greater use if they can be directly compared to experimental data. By analyzing the network with the appropriate mathematical tools, it is possible to run simulations that predict outputs given a set of inputs.

In sum, once the inputs have been determined they are applied to the experimental system as well as to the mathematical representation. The predicted and measured outputs are obtained and compared. The reconciliation of experimental and computational results, which may also be automated, is in actuality interpretation of the experimental data on a grand scale. It can lead to the identification of many new components and interactions in the system at once.

The incorporation of these elements (high-throughput technology and mathematical modeling) with the traditional biology process is one definition of systems biology (Cowley 2004). The purpose of this chapter is to show how this integrative approach can be applied to metabolic and transcriptional regulatory networks.

II. METABOLIC NETWORKS

For several reasons, the state of metabolic network reconstruction and modeling is the more advanced. Much of the required information for network reconstruction can be obtained from the annotated genome sequence and enzyme-to-reaction databases, and several organisms are quite well characterized biochemically. Because there is a wealth of literature on this topic, we describe it fairly generally and refer to the reviews for more detail (Covert et al. 2001; Price et al. 2004; Gagneur and Casari G 2005; Patil et al. 2004).

A. Network reconstruction

Metabolic network reconstruction begins by compiling a list of all enzymes and transport proteins identified in the annotated genome sequence of an organism,

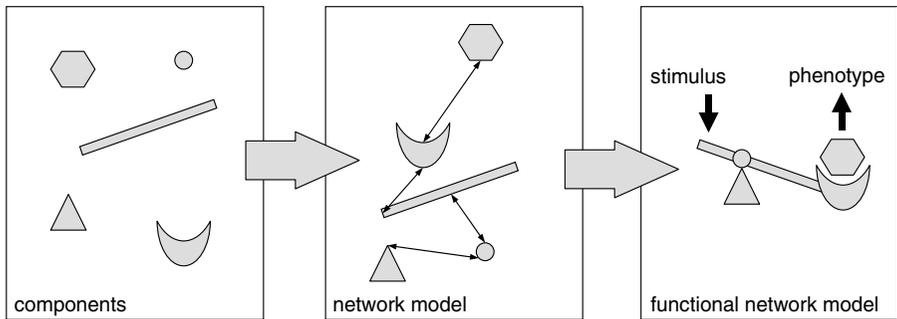


Figure 10.2. Functional network reconstructions. Metabolic and regulatory networks may be reconstructed in terms of component lists or graphs indicating some interactions, but are most useful when integrated in such a way that they actually predict behaviors that can be compared with experimental observations.

as found in a database such as the Comprehensive Microbial Resource (Peterson et al. 2001) or MetaCyc (Krieger et al. 2004). Each protein or protein complex is associated with one or more metabolic reactions or transport processes, using a database such as the ENZYME databank (Bairoch 1994). Missing pieces of the network may also be found in the biochemical literature or identified by comparison with known pathways in other organisms (Overbeek et al. 2000) to obtain a relatively complete reconstruction.

Although such reconstructions can be very useful for some types of network analysis (Jeong et al. 2000; Ma et al. 2004), for applications such as phenotype simulation a more complete network is required. Put simply, the metabolic network must be “functional”; that is, given a set of known and well-characterized behaviors of the organism the reconstruction must contain all proteins necessary to simulate these behaviors (Figure 10.2). For a vegetative cell, the network must allow production or transport of biomass components (e.g., all essential amino acids, nucleotides) given a defined medium, and must be able to take up known substrates and produce known secreted metabolites. The initial reconstruction of a functional metabolic network therefore requires a thorough integration of genomic, biochemical, and phenotypic data.

What is in a reconstruction? This depends on several factors, most importantly the type of analysis to be performed on the network. For a simple graph network analysis, all that is required is a set of nodes (e.g., metabolites) and the interactions between the nodes (e.g., reactions). To enable a metabolic flux analysis, it is also necessary to include the stoichiometry of the reactions as well as some flux information, such as a maximum oxygen or substrate uptake rate. Flux balance analysis, described in more detail later in the chapter, also requires definition of the organism’s biomass composition, in terms of how many moles of all amino acids, nucleotides, and so on are contained in one gram dry weight of the organism. For a complete kinetic description, all of the kinetic parameters would need to be

included. However, the parameters would be extremely difficult to obtain (Bailey 2001). Recently, some of the most detailed metabolic network reconstructions have been updated to include complete charge and elemental balancing, in addition to stoichiometry, biomass composition, and some maximum uptake and secretion flux rates (Reed et al. 2003; Duarte et al. 2004).

B. Analysis and simulation

Once the network is reconstructed it may be analyzed, depending on the detail of the reconstruction (as described previously). Because other chapters will discuss graph-based and detailed kinetic modeling approaches, I will focus on the analysis methods currently applicable to large-scale functional networks, under the umbrella term *constraint-based modeling*.

Constraint-based modeling itself has been reviewed thoroughly and frequently over the past several years (Covert et al. 2003; Price et al. 2004). In brief, because of the difficulty of obtaining a complete detailed description of all reaction fluxes in the metabolic network, constraint-based analysis instead focuses on limiting the ranges these flux values can have, given a set of constraints. These constraints generally include those associated with mass balance and the stoichiometry of biochemical reactions, as well as reaction reversibility and certain maximum flux rates. More recently, the constraints of energy balance have also been added (Beard et al. 2002).

In practical terms, constraint-based analysis begins with mass-balance equations for each metabolite, as shown in Equation 10.1.

$$\frac{dX}{dt} = \sum v_{\text{syn}} - \sum v_{\text{deg}} + \sum v_{\text{trans}} \quad (10.1)$$

Here, X is the metabolite concentration, and v represents reaction fluxes that synthesize (*syn*), degrade (*deg*), or transport (*trans*) metabolites into and out of the system. It is often assumed that the system is at a quasi-steady state with respect to metabolism (i.e., $dX/dt = 0$, described in more detail in material following). Incorporating this assumption and combining all of the mass balance equations yields Equation 10.2.

$$\mathbf{S}\mathbf{v} = \mathbf{0} \quad (10.2)$$

Here, \mathbf{S} is the stoichiometric matrix for the system and \mathbf{v} is a vector of all fluxes in the system. Other constraints—such as the reversibility of metabolic reactions (e.g., $v_i \geq 0$), as well as maximum enzyme/transport capacity of proteins (e.g., $v_i \leq v_{\text{max}}$)—are incorporated when known.

Once these constraints have been defined, the overall capabilities of the metabolic network may be determined using extreme pathway analysis or elementary mode analysis (Papin et al. 2004), and flux distributions that optimize network production of cellular biomass components may be determined using flux-balance analysis (Price et al. 2004). Recent years have been extremely fruitful in terms of

developing creative and insightful analysis techniques for studying metabolic networks (Price et al. 2004).

It is often assumed that such analyses are limited because one assumption crucial to all of the approaches discussed here is that the metabolic network is at a steady-state. However, as the time constants relevant to metabolic reactions are on the order of milliseconds (McAdams et al. 1998), behavior of the network may be simulated dynamically. The simulation is simply broken into several time steps just large enough that the metabolic network may be assumed to be at a quasi-steady state, and differential equations are solved to calculate the growth, uptake, and secretion of various metabolites over time (Varma et al. 1994). Such an approach is important when incorporating the highly dynamic behavior of the transcriptional regulatory network.

C. Predicted and measured outputs

Because the utility of a mathematical model depends on how directly model predictions may be compared to experimental data, it is useful to describe the experimental techniques used to study metabolism. Of particular importance are the measured outputs from such techniques, and whether they can be compared to predicted outputs. The measurable outputs for metabolic networks are growth rate, concentrations of external compounds over time, and internal metabolic fluxes. For metabolic networks, we can now assess growth rate under various environmental conditions on 96-well plates using phenotype microarrays (Bochner et al. 2001). Substrate uptake and product secretion rates can be measured using standard chromatography techniques, and high-throughput metabolomic technologies are being developed (Kell 2004). The uptake and metabolism of radiolabeled substrates may also be used to calculate internal metabolic fluxes indirectly (Sauer 2004). Current metabolic network reconstructions allow for direct comparison with all of these data using flux-balance analysis (as described previously).

III. REGULATORY NETWORKS

A. Network reconstruction

Regulatory networks differ from metabolic networks in ways that impact the network reconstruction as well as modeling approaches (Herrgard et al. 2004). First, the components are different. Whereas metabolic networks involve metabolites, enzymes, and transport proteins, regulatory networks involve regulatory proteins and the promoter regions of target genes. Second, most of the metabolic proteins are well conserved across species. Regulatory proteins may also be conserved. However, the *cis* regulatory regions are generally not conserved across species, and transcription factor binding sites are extremely difficult to find in promoter regions due to their short length, although progress is being made (Beer et al. 2004). In addi-

tion, the interactions of transcription factors at one promoter region can be extremely complex (Davidson et al. 2002), and even a single nucleotide difference in a transcription factor binding site can change the specificity of cofactor binding (Leung et al. 2004).

Accordingly, the level of characterization of regulatory networks does not approach that found in metabolic networks. Currently, detailed genome-scale regulatory networks have been reconstructed only for *Saccharomyces cerevisiae* (Lee et al. 2002; Harbison et al. 2004) and *E. coli* (Shen-Orr et al. 2002; Salgado et al. 2004). These reconstructions are qualitative, including the effect of active transcription factors on target genes (whether the factor acts as an inducer, repressor, or both). More detailed reconstructions, which would include some of the dynamics of gene expression, are extremely useful but also far more difficult to obtain (Kalir et al. 2004).

Notwithstanding these challenges to those wishing to study regulation, two high-throughput technologies have made it possible to reconstruct regulatory networks at the large scale. First, microarray analysis enables the determination of the expression profile of an entire genome in one experiment (Gardner et al. 2003). Second, it is now possible to determine with some accuracy where all of the transcription factors are binding in the genome under a given set of experimental conditions (Lee et al. 2002). These two approaches, especially when used in combination with each other or with the existing literature, are a powerful way of characterizing a regulatory network (Hartemink et al. 2002; Herrgard et al. 2003).

B. Analysis and simulation

Regulatory network modeling approaches are significantly different from metabolic network modeling approaches (McAdams et al. 1998; de Jong 2002; Tyson et al. 2003; Herrgard et al. 2004). They include Boolean logic (Thomas 1973), fuzzy logic (Lee et al. 1999), Bayesian models (Hartemink et al. 2002), kinetic models (Kremling et al. 2001; Kalir et al. 2004), and stochastic models (McAdams et al. 1998). In general, the greater the level of detail required by the modeling approach (in terms of the number of parameters) the less complex the network studied, down to the extreme simplicity of engineered regulatory networks (Hasty et al. 2002). On the other hand, the detailed models of small engineered systems have been instrumental in developing our understanding of the effect of noise on network dynamics (Elowitz et al. 2000).

For large-scale modeling, an approach that is qualitative is most advantageous, because of the qualitative nature of the existing literature (Bolouri et al. 2002). The presence of relevant stimuli, activity of regulatory proteins, and expression of target genes can all be described in terms of Boolean logic. This framework was demonstrated to be particularly useful for integrating regulatory and metabolic models, wherein the effects of regulatory events are represented as time-dependent constraints on the metabolic network (Covert et al. 2001).

C. Measured and predicted outputs

The typical outputs of comparing expression of a gene under two conditions using microarrays or quantitative real-time RT-PCR are a p -value (derived from appropriate statistical analysis of repeated expression measurements) indicating the probability that a change in expression occurred, and a ratio of expression levels or signal intensities, which assigns a quantitative magnitude of the expression shift. For comparison with genome-wide qualitative gene expression changes, a regulatory network need therefore only be expressed in terms of logical rules. More detailed models also allow comparison with the ratio data for a limited number of genes (Kremling et al. 2001; Kalir et al. 2004).

IV. EXPERIMENTAL AND COMPUTATIONAL DATA INTERPRETATION

Although there are many reasons to build models (Bailey 1998), one of current importance is to elucidate the biology of the modeled network. Specifically, models can be used to identify or indicate the presence of previously unknown components or interactions in the network. This occurs through integration and reconciliation of measured and computationally predicted experimental outcomes. The remainder of this chapter focuses in depth on the use of a combined regulatory-metabolic model in *E. coli*, which was used in coordination with high-throughput experimental studies to facilitate elucidation of the metabolic and regulatory networks (Covert et al. 2004) (Figure 10.3).

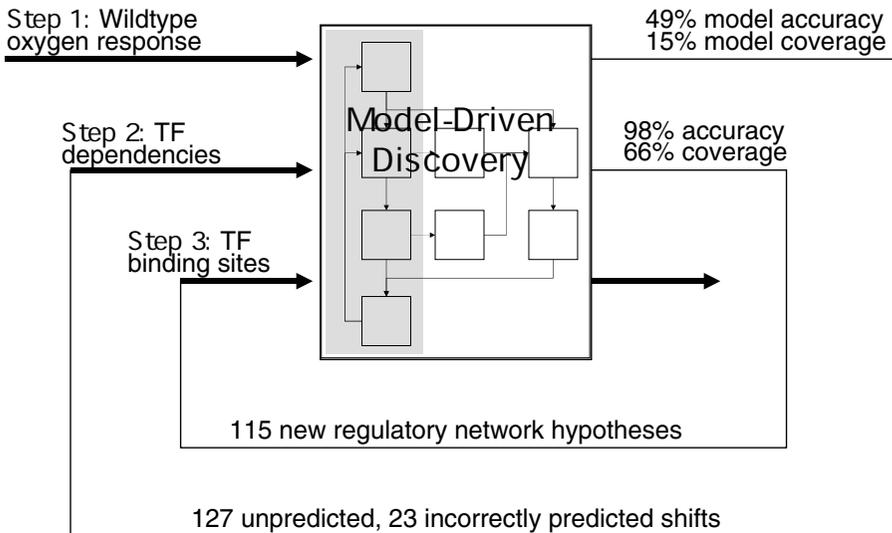


Figure 10.3. Model-driven discovery as applied to *E. coli*. High-throughput experiments and a mathematical model were integrated using the approaches described in this chapter to determine many previously unknown interactions in the transcriptional regulatory network.

The *E. coli* model accounts for the products of 1,010 genes, or roughly one-third of the annotated genes in this organism. It contains 104 regulatory proteins, which regulate the transcription of 479 metabolic genes. There are 906 genes that constitute the metabolic component of the model. The metabolic network is described and simulated using flux-balance analysis, and the regulatory model uses logic statements to describe regulatory events. The two networks interface via the constraint-based framework: regulatory events are interpreted simply by imposing time-dependent constraints on the metabolic network. Such an approach had previously been shown to result in more accurate predictive capability as well as broader scope of prediction (Covert et al. 2002).

The model is able to generate predictions of the following outputs: growth rate, substrate uptake rates, by-product secretion rates, medium concentrations of biomass and metabolites over time, internal flux rates, and shifts in gene expression. In addition, it can predict the effects of internal (i.e., deletion of one or more metabolic or regulatory genes) and external (i.e., change in medium composition, availability of oxygen, and so on) perturbations on the behavior of the system. Experimental data corresponding to all of the predictions listed here can also be obtained with relative ease using standard methods in microbial physiology and gene expression profiling.

Model predictions were compared to two large data sets for the purpose of network elucidation. The first was a large set of phenotype data available from the ASAP database (Glasner et al. 2003). Cells were seeded onto 96-well plates, with each well containing a medium designed to test one feature of microbial metabolic capability (e.g., the ability to utilize glucose as a sole carbon source) and allowed to grow overnight, upon which respiration of the cells was compared to a negative control as an indicator of growth (Bochner et al. 2001). The data that could be compared to model predictions included 110 different growth environments and 125 knockout strains of *E. coli* for a total of 13,750 outcomes.

The predicted and measured outcomes agreed in most (approximately 80%) of the cases. More interestingly, the model failures corresponded to particular environments or strains. Closer examination of the failures led to new hypotheses about *E. coli* metabolism and regulation. In all, comparison of prediction and experiment for 10 environmental conditions and eight knockout strains led to new hypotheses about regulatory interactions or uncharacterized enzymes and metabolic pathways.

As an example, one of the environmental tests was the ability of the cells to grow using thymidine as a sole carbon source. The model predicted that such growth was impossible. However, the measured data showed that each of the knockout strains was able to grow. One possible reason for the model failure is that the reconstructed metabolic network lacks a thymine-reductive pathway (including enzymes with the following E.C. numbers: 1.3.1.2 or 1.3.1.1, 3.5.2.2, and 3.5.1.6). As including this pathway would reconcile the model predictions and measured observations, one can find the most likely open reading frames to encode the pathway using sequence and phylogeny comparison tools such as MAST and MEME (Reed et al. 2003; Covert et al. 2004). In this case, the most likely open reading frames

(ORFs) for the thymine-reductive pathway enzymes are b2106 for 1.3.1.2 and b2873 or b0512 for 3.5.2.2. Such hypotheses have been verified in past metabolic network studies (Covert et al. 2001).

The second set of data was a collection of gene expression profiles generated as part of the study. Based on an earlier study (Herrgard 2003) the aerobic-anaerobic shift was targeted as a portion of the network with an intermediate level of characterization. The gene expression profile was obtained for *E. coli* during exponential growth on M9 glucose minimal medium under aerobic and anaerobic conditions. The model was used to predict the differential gene expression between the profiles, as well as growth rates and the like. In this case, the comparison between model predictions and experimental outcomes involves two measures: the accuracy (where a shift was predicted, it was also observed) and coverage (where a shift was observed, it was also predicted) of model predictions. For the first version of the model, the accuracy was about 49% and the coverage was only about 15%. These measurements indicate first that the regulatory network is much less characterized than the metabolic network, and second that the aerobic-anaerobic part of the network in particular requires more scrutiny to be fully understood.

The discrepancies between experiment and model were examined in more detail by determining the transcription factor dependencies of the differential expression observed in the wild type. This was accomplished via a perturbation analysis (Ideker et al. 2001) (Figure 10.4). Strains in regulatory proteins involved in the molecular response to oxygen were constructed, and their gene expression profiles under conditions identical to the wild type were determined. Using analysis of variance enabled determination of whether a shift in expression observed in the wild type was abolished in the knockout strain. This led directly to description of a logical rule.

For example, the *kgtP* gene (b2587) was listed without a regulatory rule in the original model. However, the microarray data indicated a significant shift with a log₂ ratio of 2.05 between the aerobic and anaerobic conditions. The perturbation studies indicated that the differential expression observed in the wild type was abolished in the $\Delta arcA$ and the $\Delta arcA \Delta fnr$ knockout strains. As a result, the rule was rewritten as $kgtP = \text{IF NOT (ArcA) (ArcA, Fnr, and NarL are regulatory proteins that also have rules that dictate their activity)}$. For the *fdnI* gene (b1476), a rule already existed: $fdnI = \text{Fnr OR NarL}$. However, no differential expression was observed. The rule became $fdnI = \text{NarL}$. In several cases, the only change made to resolve the model predictions and observations were in the interactions between regulatory proteins (e.g., changing an AND to an OR, and vice versa). This is an important observation, as the regulatory effects of most regulatory proteins to date have been tested singly and not in combination.

This analysis led to a greatly improved network model. The second-version *E. coli* model predicted 67% of the 151 observed expression shifts (coverage), with a predictive accuracy of 98%. More importantly, reconciliation of the model and the data led to many new hypotheses about the regulatory network in *E. coli* that are readily testable. Finally, the new model was compared to the phenotype microarray study

Bnum	Gene	L2R	Ar	Ap	F	O	S	A/F	Rule	Addition
b0033	<i>carB</i>	0.63				X			OxyR	Oxygen
b0034	<i>calF</i>	-1.37						X	ArcA and Fnr	Oxygen
b0068	<i>sfuA</i>	0.93							Oxygen	Oxygen
b0113	<i>pdhR</i>	0.35	X		X			X	Not (ArcA and Fnr)	Oxygen
b0114	<i>aceE</i>	0.48	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0115	<i>aceF</i>	0.48	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0116	<i>lpdA</i>	1.32	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0118	<i>acnB</i>	2.63	X					X	Not (ArcA)	Oxygen
b0313	<i>betI</i>	1.98	X					X	Not (ArcA)	Oxygen
b0336	<i>codB</i>	0.43				X			OxyR	Oxygen
b0401	<i>brnQ</i>	-0.65							Not (Oxygen)	Oxygen
b0564	<i>appY</i>	-1.87			X				Not (ArcA) and Fnr	Oxygen
b0653	<i>glkK</i>	0.73	X		X			X	Not (ArcA and Fnr)	Oxygen
b0683	<i>fur</i>	0.99						X	Not (ArcA or Fnr)	Oxygen
b0721	<i>sdhC</i>	4.70	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0722	<i>sdhD</i>	4.63	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0723	<i>sdhA</i>	3.01	X	X	X			X	Not (ArcA and Fnr)	Oxygen
b0726	<i>sucA</i>	2.17	X					X	Not (ArcA)	Oxygen
b0727	<i>sucB</i>	2.07	X					X	Not (ArcA)	Oxygen
b0733	<i>cydA</i>	-0.79							Not (Oxygen)	Oxygen
b0734	<i>cydB</i>	-0.66							Not (Oxygen)	Oxygen
b0755	<i>gpmA</i>	0.84	X		X			X	Not (ArcA and Fnr)	Oxygen
b0776	<i>bioF</i>	0.48							Oxygen	Oxygen
b0778	<i>bioD</i>	0.43							Oxygen	Oxygen
b0854	<i>poIF</i>	0.83	X		X			X	Not (ArcA and Fnr)	Oxygen
b0864	<i>artP</i>	-0.57							Not (Oxygen)	Oxygen
b0993	<i>torS</i>	-0.97							Not (Oxygen)	Oxygen
b1033	<i>ycdW</i>	0.42				X			Not (ArcA or Fnr)	Oxygen
b1221	<i>narL</i>	0.56				X			Not (ArcA or Fnr)	Oxygen
b1241	<i>adhE</i>	-1.44							Not (Oxygen)	Oxygen
b1323	<i>tyrR</i>	-0.62							Not (Oxygen)	Oxygen
b1531	<i>marA</i>	0.90	X	X	X			X	Not (ArcA and Fnr) or OxyR	Oxygen
b1656	<i>sodB</i>	-0.20							Not (Oxygen)	Oxygen
b1676	<i>pykF</i>	-0.47							Not (Oxygen)	Oxygen
b1702	<i>pps</i>	0.68							Oxygen	Oxygen
b1779	<i>gapA</i>	-0.18							Not (Oxygen)	Oxygen
b1827	<i>kdgR</i>	-0.47						X	ArcA and Fnr	Oxygen
b1991	<i>cobT</i>	-0.27			X				Fnr	Oxygen
b1993	<i>cobU</i>	-0.17			X				Fnr	Oxygen
b2040	<i>rfbD</i>	0.16							Oxygen	Oxygen
b2129	<i>yehX</i>	0.35	X		X			X	Not (ArcA and Fnr)	Oxygen
b2296	<i>ackA</i>	-1.49						X	ArcA and Fnr	Oxygen
b2308	<i>hisQ</i>	0.26	X		X			X	Not (ArcA and Fnr)	Oxygen
b2309	<i>hisJ</i>	0.44	X		X	X		X	Not (ArcA and Fnr) or OxyR	Oxygen
b2344	<i>fadL</i>	0.98	X					X	Not (ArcA)	Oxygen

Legend	
L2R > +1.0	
1 > L2R > 0.5	
+0.5 > L2R > -0.5	
-0.5 > L2R > -1.0	
-1 > L2R	

Figure 10.4. Determining new regulatory rules using the perturbation approach. A list of genes for which the computational model failed to predict observed differential expression (false negatives). The observed aerobic-anaerobic log2 ratio for the wild-type cells (L2R) is shown numerically and color coded, as explained in the legend. The observed wild-type differential expression was abolished in certain transcription factor knockout strains (Ar = $\Delta arcA$, Ap = $\Delta appY$, F = Δfnr , O = $\Delta oxyR$, S = $\Delta soxS$, A/F = $\Delta arcA \Delta fnr$), as indicated by an X. These transcription factor dependencies were used to determine new regulatory rules, as shown. Note that certain transcription factors, such as OxyR, are generally active in the presence of oxygen, whereas others (such as ArcA and Fnr) are active in the absence of oxygen.

described previously, with slight improvement to the predictive capabilities there, and is therefore completely consistent with regard to all of the other available data.

V. CONCLUSIONS

This chapter shows how model-building fits in the context of experimental discovery in terms of metabolism and transcriptional regulation, using a model of *E. coli* as an example. How well this approach can be more broadly applied to organisms and processes more complex and much less understood remains to be seen.

Protein chips to measure outputs of cell signaling processes (Hall et al. 2004) and methods for simulating signaling networks at the large scale (Papin et al. 2005) are also being developed. It can be expected, however, that the success of such efforts will depend on the ability of models to generate predictions that can be directly compared to experimental measurements at a large scale. As can be seen from this case study, such models will have the potential to greatly facilitate biological discovery.

ACKNOWLEDGMENTS

The author is a Robert Black Fellow supported by the Damon Runyon Cancer Research Foundation (DR6-#1835-04), and would also like to thank Markus Herrgard for critical reading of the manuscript.

RECOMMENDED RESOURCES

- Davidson, E. (2001). *Genomic Regulatory Systems*. San Diego: Academic Press.
- Ptashne, M., and Gann, A. (2001). *Genes and Signals*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
- Stephanopoulos, G., Aristidou, A., and Nielsen, J. (1998). *Metabolic Engineering*. San Diego: Academic Press.

REFERENCES

- Bailey, J. E. (1998). Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnol. Prog.* **14**:8–20.
- Bailey, J. E. (2001). Complex biology with no parameters. *Nat. Biotechnol.* **19**:503–504.
- Bairoch, A. (1994). The ENZYME data bank. *Nucleic. Acids. Res.* **22**:3626–3627.
- Beard, D. A., Liang, S. D., and Qian, H. (2002). Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83**:79–86.
- Beer, M. A., and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell* **117**:185–198.
- Bochner, B. R., Gadzinski, P., and Panomitros, E. (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res.* **11**:1246–1255.
- Bolouri, H., and Davidson, E. H. (2002). Modeling transcriptional regulatory networks. *Bioessays* **24**:1118–1129.
- Covert, M. W., and Palsson, B. O. (2002). Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* **277**:28058–28064.
- Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E., and Palsson, B. O. (2001). Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* **26**:179–186.
- Covert, M. W., Schilling, C. H., and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *J. Theo. Biol.* **213**:73–88.

- Covert, M. W., Famili, I., and Palsson, B. O. (2003). Identifying constraints that govern cell behavior: A key to converting conceptual to computational models in biology? *Biotechnol. Bioeng.* **84**:763–772.
- Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**:92–96.
- Cowley, A. W. Jr. (2004). The elusive field of systems biology. *Physiol. Genomics* **16**:285–286.
- Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., et al. (2002). A genomic regulatory network for development. *Science* **295**:1669–1678.
- de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **9**:67–103.
- Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004). Reconstruction and validation of *Saccharomyces cerevisiae* iND750: A fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**:1298–12309.
- Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* **403**:335–338.
- Gagneur, J., and Casari, G. (2005). From molecular networks to qualitative cell behavior. *FEBS Lett.* **579**:1867–1871.
- Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**:102–105.
- Glasner, J. D., Liss, P., Plunkett, G. III, et al. (2003). ASAP: A systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31**:147–151.
- Hall, D. A., Zhu, H., Zhu, X., Royce, T., Gerstein, M., and Snyder, M. (2004). Regulation of gene expression by a metabolic enzyme. *Science* **306**:482–484.
- Harbison, C. T., Gordon, D. B., Lee, T. I., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**:99–104.
- Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* 437–449.
- Hasty, J., McMillen, D., and Collins, J. J. (2002). Engineered gene circuits. *Nature* **420**:224–230.
- Herrgard, M. J., Covert, M. W., and Palsson, B. O. (2003). Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.* **13**:2423–2434.
- Herrgard, M. J., Covert, M. W., and Palsson, B. O. (2004). Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* **15**:70–77.
- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**:929–934.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**:651–654.
- Kalir, S., and Alon, U. (2004). Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117**:713–720.
- Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7**:296–307.
- Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J. W., and Gilles, E. D. (2001). The organization of metabolic reaction networks: Application for diauxic growth on glucose and lactose. *Metab. Eng.* **3**:362–379.

- Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y., and Karp, P. D. (2004). MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**:D438–D442.
- Lee, B., Yen, J., Yang, L., and Liao, J. C. (1999). Incorporating qualitative knowledge in enzyme kinetic models using fuzzy logic. *Biotechnol. Bioeng.* **62**:722–729.
- Lee, T. I., Rinaldi, N. J., Robert, F., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**:799–804.
- Leung, T. H., Hoffmann, A., and Baltimore, D. (2004). One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell* **118**:453–464.
- Ma, H. W., Zhao, X. M., Yuan, Y. J., and Zeng, A. P. (2004). Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* **20**:1870–1876.
- McAdams, H. H., and Arkin, A. (1998). Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.* **27**:199–224.
- Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E. Jr., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. (2000). WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**:123–125.
- Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends Biotechnol.* **22**:400–405.
- Patil, K. R., Akesson, M., and Nielsen, J. (2004). Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotechnol.* **15**:64–69.
- Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001). The comprehensive microbial resource. *Nucleic Acids Res.* **29**:123–125.
- Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* **6**:99–111.
- Price, N. D., Reed, J. L., and Palsson, B. O. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**:886–897.
- Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**:R54.
- Salgado, H., Gama-Castro, S., Martinez-Antonio, A., et al. (2004). RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**:D303–D306.
- Sauer, U. (2004). High-throughput phenomics: Experimental methods for mapping fluxomes. *Curr. Opin. Biotechnol.* **15**:58–63.
- Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**:64–68.
- Thomas, R. (1973). Boolean formalization of genetic control circuits. *J. Theo. Biol.* **42**:563–585.
- Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell. Biol.* **15**:221–231.
- Varma, A., and Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**:3724–3731.

Gene Networks: Estimation, Modeling, and Simulation

**Seiya Imoto, Hiroshi Matsuno*,
and Satoru Miyano**

*Human Genome Center, Institute of Medical Science,
The University of Tokyo, Tokyo, Japan, and *Faculty of
Science, Yamaguchi University, Yamaguchi, Japan*

Chapter 11

I. INTRODUCTION

Advances in measurement technology have enabled us to obtain genome-wide biological data production ranging from DNA sequences to data from developmental biology. The computational developmental stages to bridge this infra-data and the understanding of life from a systems perspective are represented in Figure 11.1, together with the requisite milestones. In this post-genomic research direction, gene networks will play a central role in the first stage of development. In particular, computational methods for estimating, modeling, and simulating biological systems are becoming more important. Here we present our computational strategy by giving an overview of our recent contributions in computational systems biology.

The first step is “how to create gene network information” from data. For this direction, we have developed computational methods for estimating gene networks from microarray gene expression data obtained from various perturbations such as gene disruptions, shocks, and so on. One of the most promising methods is the Bayesian network model, in which genes are regarded as random variables. The discrete Bayesian network model was first applied to gene network modeling by Friedman et al. (2000), wherein gene expression levels are categorized into +1, 0, and -1.

Inspired by this strategy, we developed methods that can process continuous numerical data and automatically detect linear and even nonlinear relationships between genes (Imoto et al. 2002, 2003). We employ a nonparametric regression

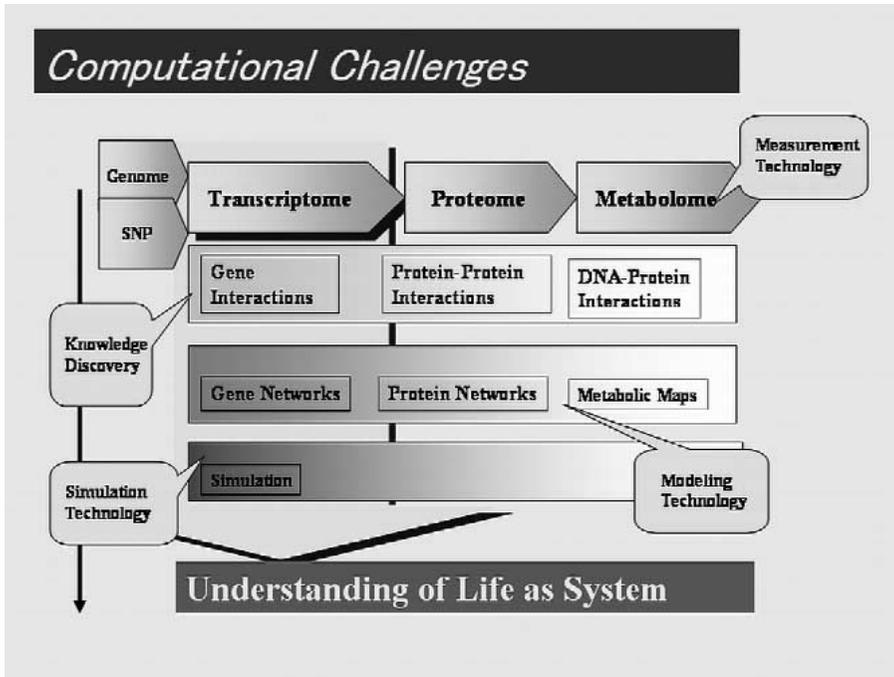


Figure 11.1. Genome-wide data and computational issues toward the understanding of life.

for capturing nonlinear relationships between genes and derive a new criterion called BNRC (Bayesian network and Nonparametric Regression Criterion) for choosing the best networks in general situations. To resolve the acyclicity restriction of the Bayesian network model, a dynamic Bayesian network with nonparametric regression for time series gene expression data has also been devised (Kim et al. 2003, 2004).

Naturally, the sole use of microarray data has limitations for gene network estimation. For improving the biological accuracy of estimated gene networks, we have created a general framework by extending this method so that it can employ other genome-wide biological information such as sequence information on promoter regions, protein-protein interactions, protein-DNA interactions, localization information, subcellular localization, and literature. Computational experiments were conducted with yeast data. These show that cascades of gene regulations were effectively extracted from the data (Tamada et al. 2003; Imoto et al. 2004; Nariai et al. 2004).

The problem of finding an optimal Bayesian network is known to be NP-hard. The brute force method employing all computing resources in the world would even require time exceeding the lifetime of the solar system for finding an optimal Bayesian network of 30 genes from 100 microarray data sets. Our approach has

made it possible to find optimal and near-optimal Bayesian networks with respect to the BNRC score in a reasonable time, and has provided evidence of the biological rationality in this computational approach (Ott and Miyano 2003; Ott et al. 2004, 2005).

The second step is “how to model and simulate gene networks” with data and biological knowledge. An important challenge is creation of a software platform with which scientists in systems biology can model and simulate dynamic causal interactions and processes in the cell, such as gene regulation, metabolic pathways, and signal transduction cascades. There have been pioneering attempts and an accumulation of knowledge in this area; for example, simulation tools (Gepasi, E-Cell, BioSPICE) and pathway databases (KEGG, BioCyc).

We have also developed a software tool (the Genomic Object Net) for pathway modeling and simulation (Matsuno et al. 2001, 2003; Doi et al. 2003; Nagasaki et al. 2003). As its architecture, we defined a notion called the Hybrid Functional Petri Net—with an extension (HFPNe)—which is a graphical programming language for describing concurrent processes. We show how computational systems biology can be explored with computational modeling and simulation through an example of a gene network for mammalian circadian rhythms (Matsuno et al. 2005).

II. GENE NETWORK ESTIMATION FROM MICROARRAY GENE EXPRESSION DATA

A. Bayesian networks and nonparametric regression

In this section, we introduce Bayesian network and nonparametric regression for estimating gene networks from microarray gene expression data.

1. Bayesian networks

A *Bayesian network* is a mathematical model for representing causal relationships among random variables by using conditional probabilities. In the context of a Bayesian network, we assume that there is a directed acyclic graph (DAG), denoted by G , as a relationship among random variables. In the gene network estimation based on Bayesian networks, a gene is regarded as a random variable and shown as a node. Let X_i ($i = 1, \dots, p$) be a discrete random variable that takes a value from $\{u_1, \dots, u_m\}$. If there is a directed edge e_{ij} from X_i to X_j , we call X_i a parent of X_j .

Further, we define $Pa(X_i) \subset \{X_1, \dots, X_p\}$ as the set of parents of X_i in G . In the DAG G , the random variable X_i only depends on its direct parents $Pa(X_i)$ and is independent of other variables (i.e., this offers the first-order *Markov property* to the relationship among variables described by G). Using the DAG G and its Markov property, the joint probability of all random variables can be decomposed as the product of conditional probabilities:

$$P(X_1, \dots, X_p) = \prod_{j=1}^p P(X_j | Pa(X_j)). \quad (11.1)$$

Because X_j is a discrete variable, the probabilities $\theta_{kjl} = P(X_j = u_k | Pa(X_j) = \mathbf{u}_{jl})$ ($j = 1, \dots, p$; $k = 1, \dots, m$; $l = 1, \dots, m^{|Pa(X_j)|}$) are parameters, where \mathbf{u}_{jl} is the l -th entry of the state table of parents of X_j and $|Pa(X_j)|$ is the number of parents of X_j . For example, for $|Pa(X_j)| = 2$ we have $\mathbf{u}_{j1} = (u_1, u_1)$, $\mathbf{u}_{j2} = (u_1, u_2)$, and so on. In this case, we can assume that $X_j | Pa(X_j) = \mathbf{u}_{jl}$ follows the multinomial distribution with probabilities $\theta_{j1l}, \dots, \theta_{jml}$ (Friedman and Goldszmidt 1998).

The conditional probabilities $P(X_j | Pa(X_j))$ describe the parent-child relationships and can be viewed as an extension of the deterministic models, such as Boolean networks (Somogyi and Sniegoski 1996). If we know the true structure of G a priori, from Equation 11.1 we can construct the joint probability function by estimating each conditional probability. However, in the gene network estimation the true G is not known and we have to estimate based on the observed data. This problem can be considered a statistical model selection problem. We describe a graph selection criterion in Section II.A.3.

Because gene expression data take continuous variables, some discretization methods are required for using the Bayesian networks based on the discrete random variables described previously. However, the discretization leads to information loss, and the number of categories and the threshold values are parameters to be optimized. Hence, a modification of Bayesian networks in order to handle continuous variables is an important problem in the gene network estimation problem. A possible solution of this problem is given by using the nonparametric regression introduced in the next section.

2. Introduction of nonparametric regression

Suppose we have n sets of data $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of p -dimensional random variable vector $\mathbf{X} = (X_1, \dots, X_p)^t$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$ corresponds to the vector of p gene expression values measured by the i -th microarray. Here, \mathbf{a}^t represents the transpose of \mathbf{a} . Using the data \mathbf{X}_n and densities instead of the probabilistic measure, we can rewrite Equation 11.1 as

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n | \boldsymbol{\theta}, G) = \prod_{i=1}^n \prod_{j=1}^p f_j(x_{ij} | \mathbf{p}_{ij}, \boldsymbol{\theta}_j), \quad (11.2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \dots, \boldsymbol{\theta}_p^t)^t$ is the parameter vector and \mathbf{p}_{ij} is the expression value vector of parents of X_j measured by i -th microarray. The construction of the conditional probability $f_j(x_{ij} | \mathbf{p}_{ij}, \boldsymbol{\theta}_j)$ is equivalent to the problem of fitting the regression model to the data $\{(x_{ij}, \mathbf{p}_{ij}); i = 1, \dots, n\}$ by $x_{ij} = m_j(\mathbf{p}_{ij}) + \varepsilon_{ij}$, where $m_j(\cdot)$ is a smooth function from $\mathbb{R}^{|Pa(X_j)|}$ to \mathbb{R} and ε_{ij} ($i = 1, \dots, n$) are independently and normally distributed with mean 0 and variance σ_j^2 .

If we set the function $m_j(\cdot)$ by $m_j(\mathbf{p}_{ij}) = \beta_o + \boldsymbol{\beta} \mathbf{p}_{ij}$, where β_o and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{|Pa(X_j)|})^t$ are parameters, we have a linear regression model to capture the relationship between x_{ij} and \mathbf{p}_{ij} (Friedman et al. 2000). However, this model assumes that the relationships between variables are linear, and it is unsuitable for extracting effective information from complex phenomena. To capture even nonlinear dependencies, Imoto et al. (2003) proposed the use of the *nonparametric additive regression model* (Hastie and Tibshirani 1990) of the form

$$x_{ij} = m_{j,1}(p_{i,1}^{(j)}) + \dots + m_{j,|Pa(X_j)}(p_{i,|Pa(X_j)}^{(j)}) + \varepsilon_{ij}, \tag{11.3}$$

where $m_{j,k}(\cdot)$ ($k = 1, \dots, |Pa(X_j)|$) are smooth functions from \mathbb{R} to \mathbb{R} and $\mathbf{p}_{ij} = (p_{i,1}^{(j)}, \dots, p_{i,|Pa(X_j)}^{(j)})^t$. We construct $m_{j,k}(\cdot)$ by the basis function expansion method with B -splines

(de Boor 1978; Imoto and Konishi 2003): $m_{j,k}(p) = \sum_{s=1}^{M_{jk}} \gamma_{sk}^{(j)} b_{sk}^{(j)}(p)$, where $\gamma_{sk}^{(j)}$ ($s = 1, \dots, M_{jk}$) are parameters, $\{b_{1k}^{(j)}(\cdot), \dots, b_{M_{jk}k}^{(j)}(\cdot)\}$ is the prescribed set of B -splines, and M_{jk} is the number of B -splines. Figure 11.2 shows an example of B -splines ($M_{jk} = 6$) of degree 3. t_d ($d = 1, \dots, 10$) are called knots. By using nonparametric regression with B -splines, we can capture even nonlinear dependencies.

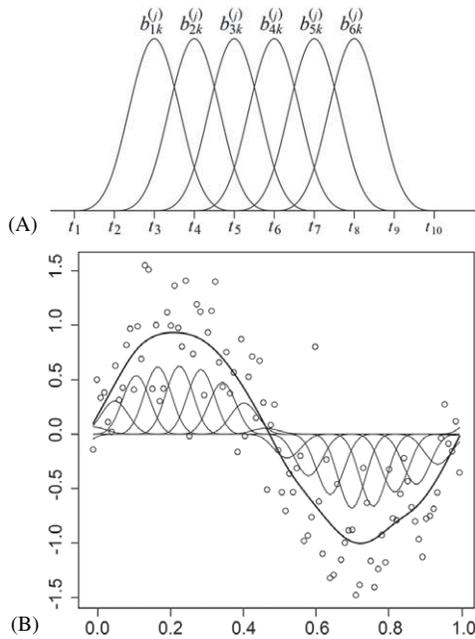


Figure 11.2. Example of B -splines. (A) Example of 6 B -splines of degree 3. The knots are equally spaced. (B) The fitted curve to simulated data: The thin curves are B -splines that are weighted by coefficients, and the thick curve is the smoothed estimate obtained by the linear combination of the weighted B -splines.

3. Bayesian networks for modeling gene networks

In this section, we describe a method for estimating gene networks from gene expression data using Bayesian networks and nonparametric regression. By combining Equations 11.2 and 11.3, we have a Bayesian network model with B-spline nonparametric regression of the form

$$f(\mathbf{X}_n | \boldsymbol{\theta}, G) = \prod_{i=1}^n \prod_{j=1}^p \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp \left\{ -\frac{\left(x_{ij} - \sum_k \sum_s \gamma_{sk}^{(j)} b_{sk}^{(j)}(\rho_{ik}^{(j)}) \right)^2}{2\sigma_j^2} \right\}. \quad (11.4)$$

Once we set a graph, the statistical model based on Equation 11.4 can be estimated by a suitable procedure. However, the problem that still remains to be solved is how we can choose the optimal graph, which gives the best approximation of the system underlying the data. We construct a criterion for evaluating a graph based on our model from Bayes' approach that is the maximization of the posterior probability of the graph.

The *posterior probability* of the graph $P(G|\mathbf{X}_n)$ is written as $P(G|\mathbf{X}_n) = p(\mathbf{X}_n|G)P(G) / p(\mathbf{X}_n) \propto p(\mathbf{X}_n|G)P(G)$, where $P(G)$ is the *prior probability* of the graph and $p(\mathbf{X}_n)$ is the *normalizing constant* and not related to the graph selection. The likelihood $p(\mathbf{X}_n|G)$ is obtained by marginalizing the joint density $p(\mathbf{X}_n, \boldsymbol{\theta}|G)$ against $\boldsymbol{\theta}$ and given by

$$p(\mathbf{X}_n|G) = \int f(\mathbf{X}_n, \boldsymbol{\theta}|G) d\boldsymbol{\theta} = \int f(\mathbf{X}_n | \boldsymbol{\theta}, G) p(\boldsymbol{\theta} | \boldsymbol{\lambda}, G) d\boldsymbol{\theta}, \quad (11.5)$$

where $p(\boldsymbol{\theta} | \boldsymbol{\lambda}, G)$ is the prior distribution on the parameter $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is the hyperparameter vector. Under the Bayesian approach, we can choose the optimal graph such that $P(G|\mathbf{X}_n)$ is the maximum. A crucial problem for constructing a criterion based on the posterior probability of the graph is the computation of the high-dimensional integration in Equation 11.5. For $\log p(\boldsymbol{\theta} | \boldsymbol{\lambda}, G) = O(n)$, the Laplace approximation for integrals (Davison 1986; Tinerey and Kadane 1986; Konishi et al. 2004) gives the analytical solution

$$\int f(\mathbf{X}_n | \boldsymbol{\theta}, G) p(\boldsymbol{\theta} | \boldsymbol{\lambda}, G) d\boldsymbol{\theta} = \frac{(2\pi/n)^{r/2}}{|J_{\lambda}(\hat{\boldsymbol{\theta}}|\mathbf{X}_n)|^{1/2}} \exp\{nl_{\lambda}(\hat{\boldsymbol{\theta}}|\mathbf{X}_n)\} \{1 + O_p(n^{-1})\}, \quad (11.6)$$

where $l_{\lambda}(\boldsymbol{\theta} | \mathbf{X}_n) = \{\log f(\mathbf{X}_n | \boldsymbol{\theta}, G) + \log p(\boldsymbol{\theta} | \boldsymbol{\lambda}, G)\} / n$, $J_{\lambda}(\boldsymbol{\theta} | \mathbf{X}_n) = -\partial^2 l_{\lambda}(\boldsymbol{\theta} | \mathbf{X}_n) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t$, r is the dimension of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is the mode of $l_{\lambda}(\boldsymbol{\theta} | \mathbf{X}_n)$. Hence, by taking minus twice logarithm of $P(G|\mathbf{X}_n)$ and substituting Equations 11.5 and 11.6 into $P(G|\mathbf{X}_n)$, Imoto et al. (2002) derived a criterion named BNRC (*Bayesian network and nonparametric regression criterion*) for choosing the optimal graph, represented as

$$\text{BNRC}(G) = -2 \log P(G) - r \log(2\pi/n) + \log |J_{\lambda}(\hat{\boldsymbol{\theta}}|\mathbf{X}_n)| - 2nl_{\lambda}(\hat{\boldsymbol{\theta}}|\mathbf{X}_n). \quad (11.7)$$

The optimal graph \hat{G} is chosen such that the criterion of Equation 11.7 is minimal. Imoto et al. (2003) also extended to results of their 2002 work to handle the *nonparametric heteroscedastic regression*. In practice, the value of BNRC(G) defined

in Equation 11.7 can be computed by the sum of the local scores, $\text{BNRC}(G) = \sum_{j=1}^p \text{BNRC}_j$, where BNRC_j is defined by the approximation of

$$-2 \log P_j(G) \int \prod_{i=1}^n f_j(x_{ij} | p_{ij}, \theta_j) p_j(\theta_j | \lambda_j) d\theta_j$$

obtained by the Laplace approximation. Here, we assume $p(\theta | \lambda, G) = \prod_{j=1}^p p(\theta_j | \lambda_j)$, and $P_j(G)$ is called the prior probability for the j -th local structure defined by the j -th variable and its direct parents. Note that $P(G) = \prod_{j=1}^p P_j(G)$ holds.

In the Bayesian network literature (Chickering 1996; Ott 2004), it is shown that determining the optimal network is an NP-hard problem. When we focus on gene networks with a small number of genes such as 30 or 40, we can find the optimal graph structure by using a suitable algorithm (Ott et al. 2004). However, for larger numbers of genes we employ a heuristic strategy such as a greedy hill-climbing algorithm to learn graph structure. The details of model learning are described in Section III.C.

III. ADVANCED METHODS FOR GENE NETWORK ESTIMATION

A. Multi-source biological information for estimating gene networks

In addition to microarray data, there are several types of information useful for estimating gene networks. In this section, we describe methods of combining gene expression data and other biological information (such as binding site information and protein-protein interaction data) to estimate gene networks.

1. General framework

The main drawback for the gene network construction from microarray data is that whereas the gene network contains a large number of genes the information contained in gene expression data is limited by the number of microarrays, their quality, the experimental design, noise, and measurement errors. Therefore, estimated gene networks contain some incorrect gene regulations, which cannot be evaluated from a biological viewpoint. In particular, it is difficult to determine the direction of gene regulation using gene expression data only. Hence, the use of biological knowledge—including protein-protein and protein-DNA interactions, sequences of the binding site of the genes controlled by transcription regulators, literature and so on—is considered a key component of microarray data analysis (Hartemink et al. 2002; Imoto et al. 2004).

Imoto et al. (2004) provided a general framework for combining microarray data and biological knowledge aimed at estimating a gene network by using a Bayesian network model. The criterion $\text{BNRC}(G)$ of Equation 11.7 contains two quantities: the prior probability $P(G)$ of the graph and the marginal likelihood of the data

$p(X_n|G)$. The marginal likelihood shows the fitness of the model to the gene expression data.

The biological knowledge can then be used as the prior probability of the graph. Suppose that the biological knowledge is represented as the matrix $\mathbf{A} = (a_{ij})$, where if we know gene_{*i*} regulates gene_{*j*} we set $a_{ij} = 1$, and otherwise $a_{ij} = 2$. Using the information of \mathbf{A} , we put a value $\zeta_{a_{ij}}$ on the edge e_{ij} . Note that $\zeta_1 < \zeta_2$ holds. The prior probability of the graph G can be expressed as

$$P(G) = \frac{1}{Z} \exp\left(-\sum_{i,j:e_{ij} \in G} \zeta_{a_{ij}}\right), \quad (11.8)$$

where Z is the normalizing constant. In Imoto et al. (2004), ζ_1 and ζ_2 are optimized by the proposed criterion. This prior probability puts a higher probability to a graph that is consistent with the information in \mathbf{A} .

2. Promoter regions

The regulation of genes is known to be realized by transcription factors (TFs), which are important subsets of proteins that transcribe mRNAs from DNAs. Genes a specific TF regulates contain a binding consensus motif called the transcription factor binding site, located in the upstream regions of the genes. Tamada et al. (2003) provided a statistical method for estimating gene networks and detecting promoter elements simultaneously. Suppose that a gene g in the network is a transcription factor.

If the children of g are directly regulated by g , they may share a consensus motif in their upstream DNA sequences. By detecting a consensus motif from a set of genes selected based on the structure of the network, we can correct the network by repairing misdirected edges and/or adding direct edges from g , based on the existence of the motif. The algorithm for simultaneous estimation of a gene network and detection of binding site is as follows.

Algorithm for Simultaneous Estimation of a Gene Network and Detection of Binding Site

- Step 1: Estimate a gene network from microarray data alone using a Bayesian network model.
- Step 2: For each gene g , let D_g be the set of child and grandchild genes of g . Consider genes with $|D_g| \geq 4$ TFs, and search for motifs in D_g .
- Step 3: For each TF, based on the result of the motif detection:
 - A: If a parent of the TF contains the motif, we reverse the edge and make it a direct child.
 - B: If a grandchild of the TF contains the motif, we add an edge and make it a direct child.
 We also embed this information into Equation 11.8.
- Step 4: Estimate a gene network again, along with the motif information.
- Step 5: Repeat steps 2 through 4 until the network does not change.

For the motif detection method used in step 2, Tamada et al. (2003) used a method called *string pattern regression* (Bannai et al. 2002), which employs the *sub-string pattern class* as the motif model.

3. Protein-protein interactions

Nariai et al. (2004) proposed the use of protein-protein interaction data for refining gene networks estimated by microarray gene expression data. When a gene is regulated by a protein complex, it is natural that a protein complex is considered a direct parent. Therefore, Nariai et al. (2004) proposed the use of virtual nodes corresponding to protein complexes in the Bayesian networks. The virtual nodes corresponding to protein complexes are created by principal component analysis, and the proposed criterion can be used to decide whether we make a protein complex.

The information of the protein-protein interaction data can be converted into the prior probability of the graph. If gene_{*i*} and gene_{*j*} show the protein-protein interaction, we set $a_{ij} = a_{ji} = 1$ in **A**. Figure 11.3 shows part of the results of Nariai et al. (2004). This method enables us not only to refine gene networks but to find unknown protein complexes.

B. Dynamic Bayesian networks

A shortcoming of the Bayesian network is that this model cannot construct cyclic networks, whereas a real gene regulation mechanism has cyclic regulations. The

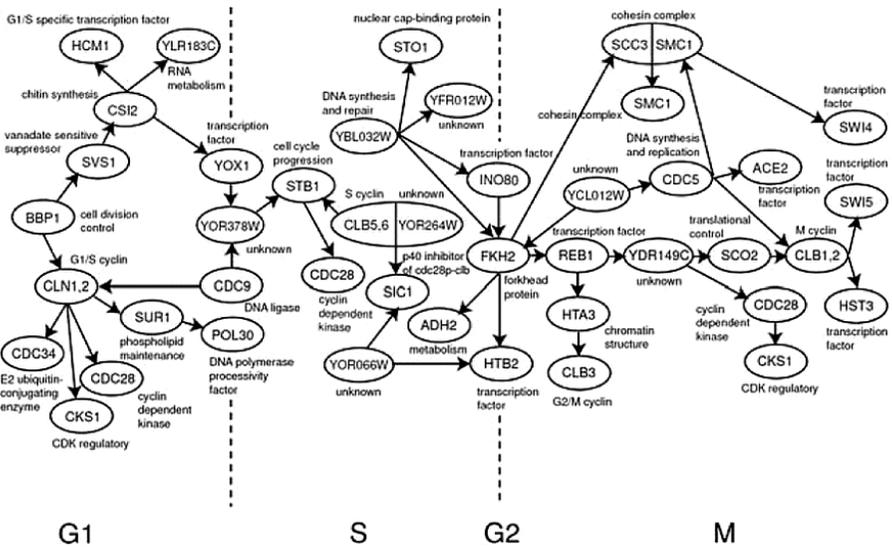


Figure 11.3. Cell cycle gene network estimated by using “phase” information together with microarray data and protein-protein interactions. The ellipses that have more than two genes are estimated protein complexes.

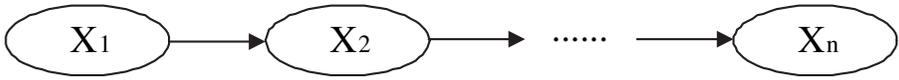


Figure 11.4. Time dynamics in the dynamic Bayesian networks.

use of *dynamic Bayesian networks* has been proposed for constructing a gene network with cyclic regulations. In the context of the dynamic Bayesian network, we consider time series data; that is, the t -th microarray data x_t corresponds to the states of p genes at time t ($t = 1, \dots, T$). Note that x_t is considered an observation of the p -dimensional random vector \mathbf{X}_t . As for the time dependency, we consider the first-order Markov relation represented in Figure 11.4.

Under this condition, the joint probability can be decomposed as

$$P(\mathbf{X}_1, \dots, \mathbf{X}_T) = P(\mathbf{X}_1)P(\mathbf{X}_2 | \mathbf{X}_1) \dots P(\mathbf{X}_T | \mathbf{X}_{T-1}). \quad (11.9)$$

The gene regulations can be modeled through the construction of $P(\mathbf{X}_t | \mathbf{X}_{t-1})$ for $t = 2, \dots, T$. The network structure is assumed to be stable through all time points. The conditional probability $P(\mathbf{X}_t | \mathbf{X}_{t-1})$ can also be decomposed into the product of conditional probabilities of each gene (given its parents) as

$$P(\mathbf{X}_t | \mathbf{X}_{t-1}) = \prod_{j=1}^p P(X_{tj} | Pa(X_j)_{t-1}), \quad (11.10)$$

where $Pa(X_j)_{t-1}$ is the set of random variables corresponding to the parent genes of the j -th gene at time $t - 1$. By combining Equations 11.9 and 11.10, we have the decomposition

$$P(\mathbf{X}_1, \dots, \mathbf{X}_T) = P(\mathbf{X}_1) \prod_{t=2}^T \prod_{j=1}^p P(X_{tj} | Pa(X_j)_{t-1}). \quad (11.11)$$

From Equation 11.11, the extension of the dynamic Bayesian networks to handle continuous variables, the detection of nonlinear relationships by using nonparametric regression, and the construction of a graph selection criterion based on the Bayesian approach can be done in the same way as the Bayesian networks described in Section II.A. The details of the combination of the dynamic Bayesian networks with the nonparametric regression are described by Kim et al. (2003, 2004).

C. Searching optimal Bayesian networks

Finding optimal Bayesian networks is computationally difficult. Potentially, we need to search the space of directed acyclic graphs of n vertices whose size c_n is approximately (Robinson 1973)

$$c_n = \frac{n! \cdot 2^{\binom{n}{2}}}{r \cdot z^n}; r \approx 0.57436; z \approx 1.4881.$$

From this formula we can see that there are roughly $2.34 \cdot 10^{72}$ networks with 20 vertices and $2.71 \cdot 10^{158}$ for 30 vertices. This complexity does not allow us any brute-force approach, even with a supercomputer system. Furthermore, without obtaining the optimal Bayesian networks we cannot conclusively determine that the Bayesian network model can really extract biologically meaningful regulatory information from microarray gene expression data. Thus, we face two issues. The first issue is how to cope with this complexity, and the second is the search for optimal Bayesian networks and their biological evaluation.

1. Greedy heuristics for searching Bayesian networks

Heuristic approaches have been applied to this search problem such as greedy algorithms (Heckerman et al. 1995; Friedman et al. 2000; Imoto et al. 2002), simulated annealing (Hartemink et al. 2002), and genetic algorithms (van Someren et al. 2002). The greedy hill-climbing algorithm due to Heckerman et al. (1995) is presented in the following as a typical example, where n is the number of repeats.

The greedy algorithm assumes a score function for solutions. It starts from some initial solution and successively improves the solution by selecting the modification from the space of possible modifications that yields the best score. When no improvement is found, the algorithm terminates with the current best solution. Some ideas should be employed for the choice of the initial solution and for the choice of the space of possible modifications. Biologically reasonable locally optimal Bayesian networks of several hundred genes have been reported (Imoto et al. 2002, 2003; Tamada et al. 2003; Nariai et al. 2004).

Greedy Hill Climbing (Heckerman et al. 1995)

- Step 1: Initialize the network as the empty network.
- Step 2: Randomly select a permutation $\pi: \{1, \dots, |X|\} \rightarrow X$.
- Step 3: For all $i = 1, \dots, |X|$, do the following two steps:
 - A: Compute the changes of the score when adding a new parent for $\pi(i)$ or removing or reversing the edge of a parent gene of $\pi(i)$.
 - B: Select the modification among the modifications that improve the score most without violating the acyclicity condition.
- Step 4: Repeat step 3 until the score does not improve.
- Step 5: Repeat steps 1 through 4 for n times and return the best solution found in these iterations.

2. Search algorithm for optimal Bayesian networks

A BNRC score can be decomposed to the additive form $\text{BNRC}(G) = \sum_{j=1}^p \text{BNRC}_j$, as discussed in Section II.A.2. We will formulate this optimization problem in an abstract way: For a finite set X (of genes), we call a function $s: X \times 2^X \rightarrow \mathbf{R}$ a *score function* for X . Then, for a DAG $G = (X, E)$ we define the score of X by $\text{score}(G) = \sum_{g \in X} s(g, \text{Pa}(g))$. This corresponds to Equation 11.7 and its decomposition

as previously. The problem is to find the best network $G = (X, E)$ that attains the optimal score.

In the case of the BNRC score, the problem is defined as a minimization problem. Furthermore, it is noted by Ott (2004) that the case for the MDL score (Friedman and Goldszmidt 1998) is also formulated as a minimization problem, whereas the case for the BDE score (Cooper and Herskovits 1992; Heckerman et al. 1995; Friedman and Goldszmidt 1998) is defined as a maximization problem.

Ott et al. (2004) have devised an algorithm that can find optimal Bayesian networks of size up to 35 if a supercomputer such as the SUN FIRE 15-K (with 96 CPUs of 900 MHz each) is used. The algorithm decomposes the search space into subspaces and employs the dynamic programming technique for finding the right subspace as well as for determining the optimal solution in the subspace.

To describe the algorithm, several notations require introduction. For a gene g in X and a subset $A \subseteq X$, $F(g, A) = \min_{B \subseteq A} s(g, B)$ gives the optimal choice of parents for g if the parents are selected from A . An order on a subset $A \subseteq X$ is given as a permutation $\pi: \{1, \dots, |A|\} \rightarrow A$. We denote by Π^A the set of all permutations on A . We denote the subnetwork of $G = (X, E)$ restricted to A by $G(A) = (A, E(A))$. For a permutation $\pi \in \Pi^A$, we say that $G(A)$ is π -linear if $\pi^{-1}(g) < \pi^{-1}(h)$ holds for all $(g, h) \in E(A)$. The idea of the algorithm is to decompose the set of all DAGs on A into subsets of π -linear DAGs for all $\pi \in \Pi^A$.

Then we divide the problem into (1) finding the subspace of the search space that contains the optimal network and (2) finding the optimal network within the selected subspace. We denote $Q^A(\pi) = \sum_{g \in A} F(g, \{h \in A | \pi^{-1}(h) < \pi^{-1}(g)\})$. Then we find the best π -linear network for any given permutation by F and Q . The optimal network can be found by finding the optimal permutation that yields the global minimum, which is given by $M(A) = \arg \min_{\pi \in \Pi^A} Q^A(\pi)$. Then the entire algorithm is described as follows.

Algorithm for Finding Optimal Bayesian Network (Ott et al. 2004)

- Step 1: Compute $F(g, \emptyset) = s(g, \emptyset)$ for all $g \in X$.
- Step 2: For all $g \in X$ and all $A \subseteq X - \{g\}$ with $A \neq \emptyset$ compute $F(g, A)$ as $\min_{a \in A} \{s(g, A), \min F(g, A - \{a\})\}$.
- Step 3: Set $M(\emptyset) = \emptyset$.
- Step 4: For all $A \subseteq X$ with $A \neq \emptyset$ execute the following steps.
- A: Compute $g^* = \arg \min_{g \in A} (F(g, A - \{g\}) + Q^{A - \{g\}}(M(A - \{g\})))$.
- B: For all $1 \leq i < |A|$, set $M(A)(i) = M(A - \{g^*\})(i)$ and $M(A)(|A|) = g^*$.
- Step 5: Return $Q^*(M(X))$.

Theorem (Ott et al. 2004): Optimal Bayesian networks can be found using $\left(\frac{|X|}{2} + 1\right) \cdot 2^{|X|}$ dynamic programming steps, where X is a set of genes.

A rigorous proof is required to show the correctness of this algorithm (Ott et al. 2004). Furthermore, with some biologically reasonable constraints on the networks we can obtain a much faster algorithm (Ott and Miyano 2003). By computing optimal Bayesian networks of small size and evaluating them, it is reported that optimal Bayesian networks are not necessarily biologically optimal. However, by combining optimal to near-optimal Bayesian networks thoroughly we can extract biologically more accurate information from microarray gene expression data (Ott and Miyano 2003; Ott et al. 2004, 2005).

IV. PETRI-NET-BASED MODELING OF GENE NETWORKS

A. Hybrid functional Petri nets for modeling gene networks

1. Hybrid functional Petri net

Petri net is a graphical programming language for modeling concurrent systems. It has been mainly used to model artificial systems such as manufacturing systems and communication protocols. From the first attempt by Reddy et al. (1993), several types of Petri nets—including the stochastic Petri net (Goss et al. 1998) and the colored Petri net (Genrich et al. 2001)—have been employed to model biological pathways. On the other hand, biological pathways can be observed as hybrid systems. For example, protein concentration dynamics behave continuously when coupled with discrete switches. Protein production is switched on or off, depending on the expression levels of other genes (i.e., the presence or absence of other proteins in sufficient concentration).

Based on this observation, we proposed the *hybrid functional Petri net* (HFPN) (Matsuno et al. 2003) and its extension, called the *hybrid functional Petri net with extension* (HFPNe) (Nagasaki et al. 2003, 2004, 2005) for modeling biological pathways. We also developed the HFPNe-based simulation software called the Genomic Object Net (GON).

With GON, we have modeled and simulated many biological pathways, including the gene switch mechanism of lambda phage (Matsuno et al. 2000), the signal transduction pathway for apoptosis induced by the protein Fas (Matsuno et al. 2003), the glycolytic pathway in *E. coli* with the lac operon gene regulatory mechanism (Doi et al. 2004), alternative splicing, frame shifting, and Huntington's disease model (Nagasaki et al. 2004).

Because the GON incorporates a biology-oriented GUI, modeling of very complex biological processes with HFPNe can be performed in a simply way. In that the purpose of this chapter is to show that the notion of the Peri net has a high affinity to biological process modeling, we deal only with HFPN and will not expand on HFPNe. For further details, see Nagasaki et al. (2005).

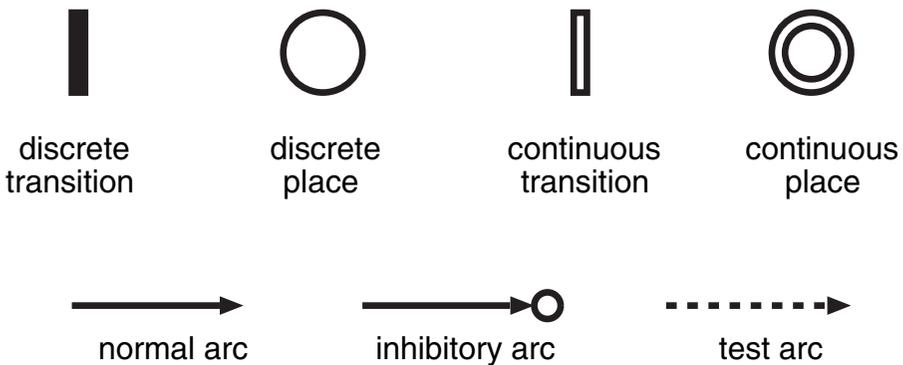
Generally, biological molecular interactions are explained with pictures representing molecules (e.g., genes, mRNAs, proteins, and protein complexes), and with arrows representing interactions of these molecules such as activation and repression. To model these interactions mathematically, differential equations have been

commonly used. However, in this modeling process we have to make redundant efforts to reconstruct a system of differential equations from the biological interaction map.

Modeling with HFPN allows us to construct a computational model for simulation without making such a redundant effort. That is, an HFPN model is directly constructed from the map of a biological pathway. Thereafter, parameters of reactions such as the transcription speeds of genes and degradation rates of proteins are tuned so that input/output concentration behaviors of substances such as mRNAs and proteins are matched to biological facts obtained from experiments or found in the literature. Because the HFPN-based modeling method follows the graphical pictures of biological pathways, the constructed HFPN model can be readily understood without getting into mathematical consideration.

The *Petri net* (Reisig 1985) is a network consisting of *place*, *transition*, *arc*, and *token*. A place can hold tokens as its content. A transition has arcs coming from places and arcs going out from the transition to some places. A transition with these arcs defines a firing rule in terms of the content of the places where the arcs are attached. *Hybrid Petri net* (HPN) (Alla and David 1998) has two types of places (*discrete place* and *continuous place*) and two types of transitions (*discrete transition* and *continuous transition*). A discrete place and a discrete transition are the same concepts as used in the traditional discrete Petri net.

A continuous place can hold a nonnegative real number as its content. A continuous transition fires continuously at the speed of a parameter assigned at the continuous transition. The graphical notations of a discrete transition, a discrete place, a continuous transition, and a continuous place are shown in Figure 11.5, together with three types of arcs. A specific value w is assigned to each arc as a weight. When a *normal arc* is attached to a discrete/continuous transition, w tokens are transferred through the normal arc, as normal arcs coming from places or going out to places.



An *inhibitory arc* with weight w enables the transition to fire only if the content of the place at the source of the arc is less than or equal to w . For example, an inhibitory arc can be used to represent repressive activity in gene regulation. A *test arc* does not consume any content of the place at the source of the arc by firing. For example, a test arc can be used to represent enzyme activity, in that the enzyme itself is not consumed.

The *hybrid dynamic net* (HDN) (Drath 1998) has a structure similar to that of the HPN, using the same types of places and transitions as the HPN. The main difference between HPN and HDN is the firing rule of continuous transition. As we can see from the previous explanation of HPN, for a continuous transition of HPN the different amounts of tokens can be flowed through the two types of arcs (i.e., coming from/going out the continuous transition). In contrast, the definition of HDN does not allow transferring different amounts through these two types of arcs. However, HDN has the following firing feature of continuous transition (which HPN does not have): The speed of continuous transition of HDN can be given as a function of values in the places.

From the previous discussion, we can see that HPN and HDN have their own feature for the firing mechanism of continuous transition. As a matter of fact, both of these features are essentially required for modeling common biological reactions. This motivated us to define the notion of the *hybrid functional Petri net* (HFPN) (Matsuno et al. 2003), which includes HPN and HDN. Moreover, HFPN has the third feature for arcs; that is, a function of values of the places can be assigned to any arc. This feature was originated from the *functional Petri net* (Hofestädt and Thelen 1998), which was introduced in order to realize the calculation of dynamic biological catalytic process on Petri-net-based biological pathway modeling. The formal definition of the HFPN is given by Nagasaki et al. (2004, 2005).

2. A model of operon with HFPN

Figure 11.6 shows a hybrid Petri net model of an operon with two genes. Discrete place S_1 (S_2), discrete transition TR_1 (TR_2), continuous places R_1 (R_2) and P_1 (P_2), and continuous transitions TP_1 (TP_2), DR_1 (DR_2), and DP_1 (DP_2) constitute the first gene (the second gene) in the operon. Discrete place F_1 is used to represent transaction of transcription from the first gene to the second gene. At discrete transition T_{R1} (T_{R2}), the parameter that reflects time for transcription of the first gene (the second gene) is assigned, and at discrete transition T_{12} time for RNA polymerase to traverse the gap between the first and the second genes is assigned. For continuous transitions, parameters at T_{P1} (T_{P2}) represent translation rate of the first gene (the second gene) and parameters at D_{R1} and D_{P1} (D_{R2} and D_{P2}) represent degradation rates of mRNA and protein of the first (the second) gene, respectively.

Initially, only discrete place S_1 has one token. This reflects the situation in which RNA polymerase binds at the promoter of the operon. Just after the transcription of the first gene (the second gene) is finished, the amount of continuous place R_1 (R_2) increases by the weight assigned at the arc from the transition T_{R1} (T_{R2}) to

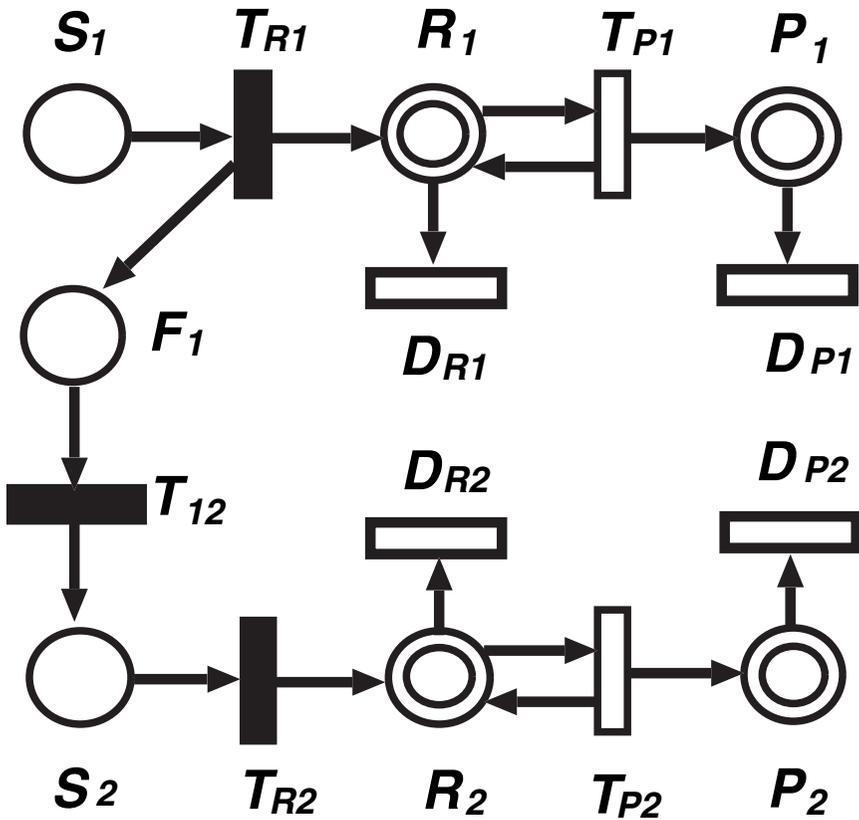


Figure 11.6. HFPN model of operon constituted by two genes.

the place R_1 (R_2). In the speed of parameters at continuous transition T_{P1} (T_{P2}), the amount of continuous place P_1 (P_2) is increased by the weight at the arc from the transition T_{P1} (T_{P2}), which reflects the translation rate of the first gene (the second gene).

Note that in order to represent the fact that mRNA is not consumed by translation two arcs are described in both directions between the place R_1 (R_2) and the transition T_{P1} (T_{P2}). This can also be represented by one test arc from R_1 (R_2) to T_{P1} (T_{P2}). Continuous transitions D_{R1} and D_{P1} (D_{R2} and D_{P2}) without outgoing arcs are used to represent degradation of mRNA and proteins of the first gene (the second gene).

B. Modeling a gene network for circadian rhythms

This section presents an example of modeling and simulation analysis by following Matsuno et al. (2005) and shows a computational strategy with a simulation tool for systems biology.

1. Mammalian circadian genetic control mechanism

Molecular clocks reside within suprachiasmatic nucleus cells. Each molecular circadian clock is a negative feedback loop of the gene transcription and its translation into protein. The loop includes several genes and their protein products. In the case of mammals, three period genes (*Per1*, *Per2*, and *Per3*) and two Cryptochrome genes (*Cry1* and *Cry2*) constitute the negative limb, whereas *Clock* and *Bmal1* (*Bmal*) genes constitute the positive limb of the feedback loop in the molecular circadian clock. To simplify the model and gain the insight of each interaction path, we deal with two groups of genes—*Per1*, *Per2*, and *Per3* genes and *Cry1* and *Cry2* genes—as *Per* and *Cry*, respectively.

The mammalian circadian genetic control mechanism consists of two interlocked negative feedback loops. PER and CRY proteins collaborate in the regulation of their own expression, assembling in PER/CRY complexes that permit nuclear translocation and inactivation of *Per* and *Cry* transcription in a cycling negative feedback loop. At the same time, the PER/CRY complex inactivates the expression of the *Rev-Erb* gene. Proteins of *Bmal* and *Clock* form heterodimers that activate *Per*, *Cry*, and *Rev-Erb* transcriptions. The *Bmal* gene is inactivated by the REV-ERB protein in the nucleus. Except for the gene *Clock*, the genes are rhythmically expressed in about 24 hours according to these molecular interactions of genes and their products.

2. HFPN model

In the present model, *Per* and *Cry* genes and their protein products constitute the first major circadian feedback loop. The second loop consists of the *Clock* and *Bmal* genes and their protein products. These two pathways are connected by the interaction, including *Rev-Erb* and its product. Expression of *Rev-Erb* was accelerated by the PER/CRY dimmers, and the REV-ERB protein suppresses transcription of the *Bmal* gene. Figure 11.7 is an HFPN model of the mammalian circadian gene mechanism.

In the HFPN, symbols of places and transitions were changed to pictures depicted according to the corresponding biological reactions. These changes are meaningless in a mathematical sense, but meaningful in a biological sense. With only these changes of Petri net symbols to biological pictures, we can make the entire biological pathway described in the Petri net more biologically intuitive.

This HFPN model was described according to the following simple rules. To each substance such as mRNA and protein, a continuous place is assigned. To each transition, a function of the style $mX/10$ is assigned, which defines the speed of the corresponding reaction. For example, the translation speed of the PER protein is controlled by the formula $m1/5$, where $m1$ is the concentration of *Per* mRNA. This reflects the biological observation that the reaction speed of transcription is changed depending on the concentration of *Per* mRNA.

Complex forming rate is given as a formula of the style $mX*mY/10$. For example, the formula $m2*m4/10$ is assigned to the continuous transition as the complex

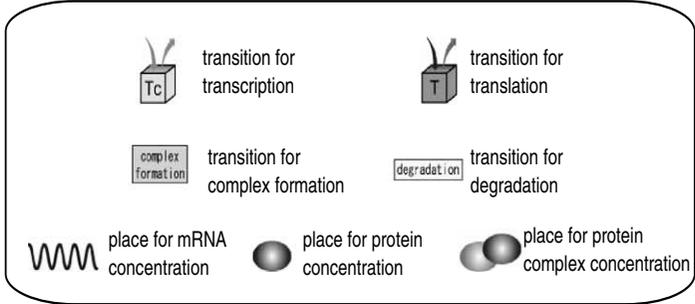
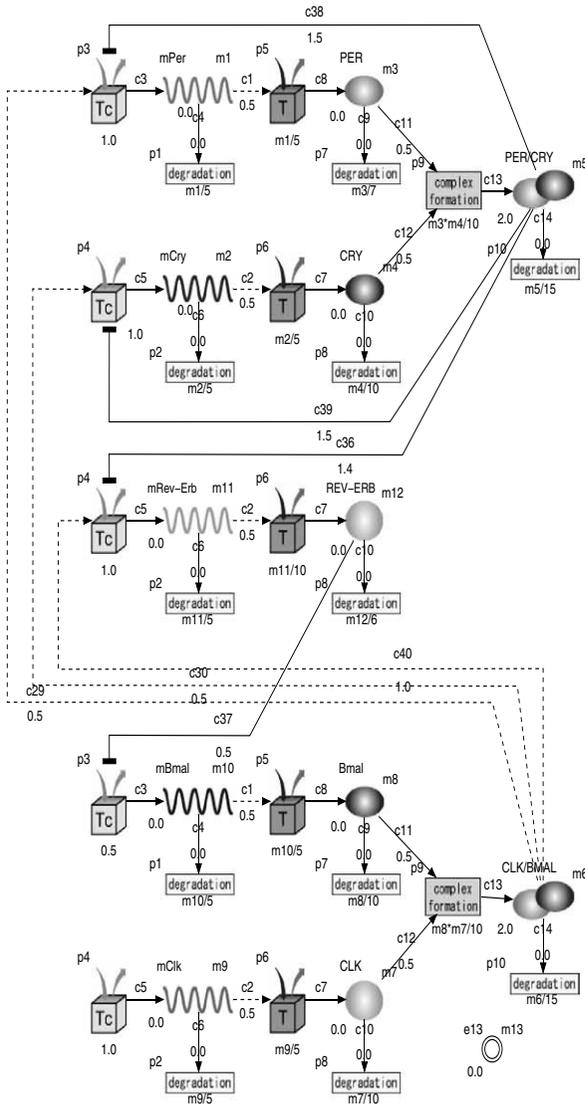


Figure 11.7. HFPN model of circadian gene regulatory mechanism in mammals. Places and transitions of HFPN have been changed to pictures depicted according to biological reactions.

forming rate of the proteins PER (m_2) and CRY (m_4). Continuous transitions without outgoing arcs are used for representing natural degradation rates of mRNAs, proteins, and protein complexes.

After constructing an HFPN of the biological mechanism to be modeled, parameters of transition speed and initial values of places have to be tuned based on the biological knowledge and/or the facts described in biological literature. In general, many trial-and-error processes are required until appropriate parameters for simulation are determined. Because GON provides the GUI specially designed for biological modeling, we can perform these processes very easily and smoothly.

3. Inconsistency discovered by simulation

We carried out simulations of the HFPN model shown in Figure 11.7 with GON. This model produces periodic oscillations of mRNA and protein concentrations, as shown in Figure 11.8. We made some modifications on this HFPN model for checking mutant behaviors, including *Per* gene disruption (by removing the normal arc going into the place PER) and preventing the *Cry* gene from transcription (by removing the test arc going into the transition attached to the place m_{Cry}). The resulting behavior of these modifications corresponded well to the facts in the biological literature (Reppert et al. 2001; Sehgal 2004). However, at the same time we found the following inconsistency with the biological observation of Figure 11.8.

- In Figure 11.8, the *Bmal* mRNA peaks at almost the same time as the peaks of *Cry* and *Per* mRNAs. However, it is biologically known that the peak of *Bmal* mRNA is located at almost the mid point of two successive peaks of *Cry* or *Per* mRNA.

4. A new interaction resolves the inconsistency

Circadian clock mechanisms have been examined in many living organisms, such as cyanobacteria, the fruit fly, and the mouse (Sassone 2003; Sehgal 2004). Many

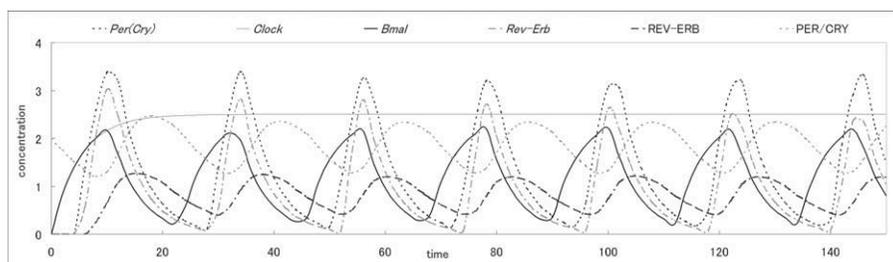


Figure 11.8. Simulation result of the HFPN model of Figure 11.7, where the dark solid line is *Bmal* mRNA, the dark dotted line is *Cry* and *Per* mRNAs, the pale dot-dash line is *Rev-Erb* mRNA, the pale dotted line is PER/CRY complex, and the dark dot-dash line is the REV-ERB protein.

investigations have been made of the fruit fly (*Drosophila melanogaster*), and it is known that it has a similar circadian gene regulatory mechanism to that of the mouse. Then, in order to fix the inconsistency pointed out in the previous section we compared these two circadian mechanisms. Consequently, we noticed that a path in the *Drosophila* circadian mechanism has not been identified in that of the mouse.

- PER/TIM complex activates the gene *dClock*, where TIM (timeless) is a protein of *Drosophila* that works in place of CRY, and *dClock* is a gene of *Drosophila* that corresponds to *Bmal*.

We conducted simulation again on the modified HFPN model in which the previously cited hypothetical path was incorporated by adding the test arc from the place PER/CRY to the transition p3 from which the normal arc connects to the place *mBmal*. Figure 11.9 shows the result of simulation. This figure shows that the inconsistency is resolved by introducing this hypothetical path. Recall that in the original model the transcription switch of gene *Bmal* was controlled only by inhibition from the REV-ERB protein.

In contrast, in the new model this transcription is controlled not only by inhibition from the REV-ERB but also by the activation from the PER/CRY protein complex. This activation from the PER/CRY complex allows the *Bmal* transcription to be off at some point during the decrease in the PER/CRY complex concentration. In summary, the simultaneous operation of two reactions “inhibition from REV-ERB” and “activation from PER/CRY” on the gene *Bmal* enables the *Bmal* mRNA peak to locate at the mid point of two successive *Cry* (*Per*) mRNA peaks.

C. Remarks

GON is a biosimulation tool developed by inheriting the tradition of research on Petri nets. Many Petri net tools have been developed by researchers in concurrent technology (Petri net tools). These Petri net tools so far developed generally have user-friendly GUIs that allow us to describe complex concurrent systems very easily

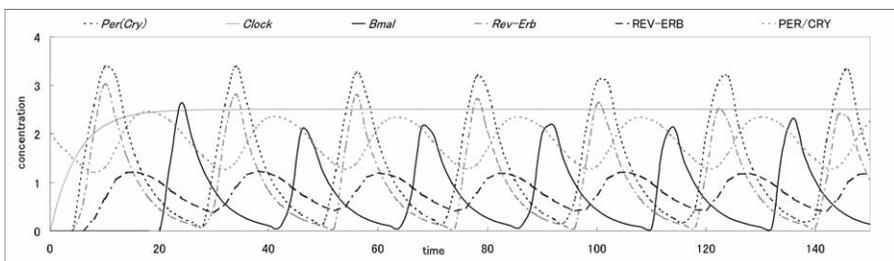


Figure 11.9. Simulation results of the modified HFPN in which the new hypothetical interaction is added. Lines have the same meanings as the lines in Figure 11.8.

and smoothly. GON inherits this feature of the Petri net, enabling us to describe and manipulate biological pathways naturally (even for biologists who are not familiar with mathematical description and programming language). GON has been commercialized as Cell Illustrator from Gene Networks International (GNI).

In this section, we explained how gene networks can be described by HFPN (with the example of the circadian genetic control mechanism in mammals) and demonstrated that computer simulations make it possible to observe behaviors of gene networks more systematically, being able to suggest new regulatory interactions that have not been found with only viewing the gene network as a map.

V. CONCLUSIONS

In the understanding of complex biological systems, computational methods, software tools, and biological databases should be extensively developed and employed. This chapter presented two approaches to understanding biological systems and described a method and a software tool developed by our research group.

We devised a Bayesian network model with nonparametric regression to extract gene network information from microarray data, and developed a series of computational methods based on this approach. It should be briefly mentioned that there are other gene network models and analysis methods. The simplest model is the Boolean network model (Somogyi and Sniegoski 1996; Liang et al. 1998; Akutsu et al. 1999, 2003). This model is suited for modeling qualitative relations between genes, and it allows mathematical and algorithmic analyses.

Another important mathematical model is based on ordinary differential equations. For example, Chen et al. (1999) considered modeling of both mRNA and protein concentrations by using a system of linear differential equations. We also devised a method to infer a gene network in terms of a linear system of differential equations from time series gene expression data (de Hoon et al. 2003).

We developed a software tool, based on the Petri net, for modeling and simulating gene networks. With this software tool, various models have been constructed. The model's utility has been demonstrated in practice. The strategy presented in Section IV will be an important key to systems biology. Furthermore, with this software tool it is possible to develop various databases of dynamic pathway models. These dynamic pathway models can then be simulated on computers.

Systems biology is anticipated to produce practical benefits such as biomedical applications, solutions for environmental problems, and so on. As an example of this, we have succeeded in discovering a drug target gene by analyzing gene networks constructed from gene expression profile data. This gene expression profile data was based on gene disruptions and drug doses (Imoto et al. 2003; Savoie et al. 2003). This example suggests that systems biology will lead to a new paradigm for target selection by employing computational modeling of gene networks.

ACKNOWLEDGMENTS

The authors would like to thank our colleagues and collaborators Hideo Bannai, Michiel de Hoon, Atsushi Doi, Takao Goto, Sunyong Kim, Satoru Kuhara, Masao Nagasaki, Naoki Nariai, Sascha Ott, Christopher J. Savoie, Yoshinori Tamada, and Kousuke Tashiro. Especially, Hiroshi Matsuno would like to thank Professor Shin-Ichi T. Inouye, who guided him to the construction of the circadian genetic control mechanism in mammals and gave him many useful and suggestive comments for simulations.

RELATED INTERNET RESOURCES

BioCyc: www.biocyc.org/

BioSPICE: www.biospice.org/

E-Cell: www.e-cell.org/

Genomic Object Net: www.GenomicObject.Net/

Gepasi: www.gepasi.org/

GNI: www.gene-networks.com/

KEGG: www.genome.ad.jp/

Petri net tools: www.daimi.au.dk/PetriNets/

REFERENCES

- Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S. (2003). Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model. *Theoretical Computer Science*, **298**(1):235–251.
- Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pacific Symp. Biocomput.*, **4**:17–28.
- Alla, H., and David, R. (1998). Continuous and hybrid Petri nets. *J. Circ. Syst. Comp.*, **8**:159–188.
- Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., and Miyano, S. (2002). A string pattern regression algorithm and its application to pattern discovery in long introns. *Genome Informatics*, **13**:3–11.
- Chen, T., He, H. L., and Church, G. M. (1999). Modeling gene expression with differential equations. *Pacific Symp. Biocomput.*, **4**:29–40.
- Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In "Learning from Data: Artificial Intelligence and Statistics V" (D. Fisher and H.-J. Lenz, eds.), pp. 121–130. Springer-Verlag.
- Cooper, G. F., and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**:309–347.
- Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, **73**:323–332.
- De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, Berlin.

- De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S. (2003). Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pacific Symp. Biocomput.*, **8**:17–28.
- Doi, A., Fujita, S., Matsuno, H., Nagasaki, M., and Miyano, S. (2004). Constructing biological pathway models with hybrid functional Petri nets. *In Silico Biology*, **4**(3):271–291.
- Friedman, N., and Goldszmidt, M. (1998). Learning Bayesian networks with local structure. *In "Learning in Graphical Models"* (M. I. Jordan, ed.), pp. 421–459. Kluwer Academic Publishers.
- Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**:601–620.
- Genrich, H., Kuffner, R., and Voss, K. (2001). Executable Petri net models for the analysis of metabolic pathways. *International J. Software Tools for Technology Transfer*, **3**(4):394–404.
- Goss, P. J. E., and Peccoud, J. (1998). Quantitative modeling of stochastic systems in molecular biology by using Stochastic Petri nets. *Proc. Natl. Acad. Sci. USA*, **95**:6750–6755.
- Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symp. Biocomput.*, **7**:437–449.
- Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**:197–243.
- Hofestädt, R., and Thelen, S. (1998). Quantitative modeling of biochemical networks. *In Silico Biology*, **1**:39–53.
- Imoto, S., Goto, T., and Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pacific Symp. Biocomput.*, **7**:175–186.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J. Bioinform. Comp. Biol.*, **2**:77–98.
- Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comp. Biol.*, **1**:231–252.
- Imoto, S., and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Ann. Inst. Statist. Math.*, **55**:671–687.
- Imoto, S., Savoie, C. J., Aburatani, S., Kim, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Use of gene networks for identifying and validating drug targets. *J. Bioinform. Comp. Biol.*, **1**(3):459–474.
- Kim, S., Imoto, S., and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.*, **4**:228–235.
- Kim, S., Imoto, S., and Miyano, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, **75**:57–65.
- Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**:27–43.
- Liang, S., Fuhrman, S., and Somogyi, R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp. Biocomput.*, **3**:18–29.
- Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomput.*, **5**:341–352.
- Matsuno, H., Inouye, S. T., Okitsu, Y., Fujii, Y., and Miyano, S. (2005). A new regulatory interaction suggested by simulations for circadian genetic control mechanism in mammals. *In*

- "Proc. 3rd Asia-Pacific Conf. Bioinformatics" (Y. P. Chen and L. Wong, eds.), pp. 171–180. Imperial College Press.
- Matsuno, H., Tanaka, Y., Aoshima, H., Doi, A., Matsui, M., and Miyano, S. (2003). Biopathways representation and simulation on hybrid functional Petri net. *In Silico Biology*, **3**(3):389–404.
- Nagasaki, M., Doi, A., Matsuno, H., and Miyano, S. (2003). Genomic Object Net I: a platform for modeling and simulating biopathways. *Applied Bioinformatics*, **2**(3):181–184.
- Nagasaki, M., Doi, A., Matsuno, H., and Miyano, S. (2004). A versatile Petri net based architecture for modeling and simulation of complex biological processes. *Genome Informatics*, **15**(1):180–197.
- Nagasaki, M., Doi, A., Matsuno, H., and Miyano, S. (2005). Computational modeling of biological processes with Petri net based architecture. In "Bioinformatics Technologies" (Y. P. Chen, ed.), pp. 179–242. Springer-Verlag.
- Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2004). Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symp. Biocomput.*, **9**:336–347.
- Ott, S., and Miyano, S. (2003). Finding optimal gene networks using biological constraints. *Genome Informatics*, **14**:124–133.
- Ott, S., Imoto, S., and Miyano, S. (2004). Finding optimal models for small gene networks. *Pacific Symp. Biocomput.*, **9**:557–567.
- Ott, S. (2004). *Finding Optimal Models for Gene Networks*. Ph.D. Thesis, Department of Computer Science, University of Tokyo.
- Ott, S., Hansen, A., Kim, S.-Y., and Miyano, S. (2005). Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution. *Bioinformatics*, in press.
- Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993). Petri net representations in metabolic pathways. In "Proc. First International Conference on Intelligent Systems for Molecular Biology (ISMB '93)", pp. 328–336. AAAI Press.
- Reisig, W. (1985). *Petri Nets*. Springer-Verlag, Berlin.
- Reppert, S. M., and Weaver, D. R. (2001). Molecular analysis of mammalian circadian rhythms. *Annual Review of Physiology*, **63**:647–676.
- Robinson, R. W. (1973). Counting labeled acyclic digraphs. In "New Directions in the Theory of Graphs" (F. Harary, ed.), pp. 239–273. Academic Press, New York.
- Sassone-Corsi, P. (2003). Novartis Foundation Symposium 253. Molecular Clocks and Light Signaling. John Wiley and Sons, Hoboken, NJ.
- Sehgal, A. (2004). *Molecular Biology of Circadian Rhythms*. John Wiley and Sons, Hoboken, NJ.
- Somogyi, R., and Sniegowski, C. A. (1996). Modeling the complexity of genetic networks: Understanding multigene and pleiotropic regulation. *Complexity*, **1**:45–63.
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19** Suppl. 2, ii227–ii236.
- Tinerey, L., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.*, **81**:82–86.
- Van Someren, E. P., Wessels, L. F. A., Backer, E., and Reinders, M. J. T. (2002). Genetic network modeling. *Pharmacogenomics*, **3**:507–525.

A Discrete Approach to Top-Down Modeling of Biochemical Networks

Reinhard Laubenbacher and Pedro Mendes

Virginia Bioinformatics Institute, Blacksburg, Virginia

Chapter 12

ABSTRACT

Mathematical and statistical network modeling is an important step toward uncovering the organizational principles and dynamic behavior of biological networks. This chapter focuses on methods of constructing discrete dynamic models of biochemical networks from high-throughput experimental data sets, also sometimes referred to as top-down modeling or reverse-engineering. Time-discrete dynamical systems models have long been used in biology, particularly in population dynamics. The models we mainly focus on here are also assumed to have a finite set of possible states for each variable. That is, the modeling framework discussed in this chapter is that of time-discrete dynamical systems over a finite state set.

After a brief survey of Boolean network and multi-state models, we discuss a modeling method using tools from computer algebra and the theory of Groebner bases. The method provides a compact description of the entire space of possible models and chooses from that space a model that is minimal in the sense that it contains no components that vanish on the data set used to construct the model. We also discuss the requirements of a mathematical program for the identification of biological systems.

I. INTRODUCTION

“All processes in organisms, from the interaction of molecules to the complex functions of the brain and other whole organs, strictly obey these physical laws. Where organisms differ from inanimate matter is in the organization of their systems and

especially in the possession of coded information (Mayr 1988, p. 2).” It is the task of systems biology to elucidate those differences. This process has barely begun and many researchers are testing computational tools that have been used successfully in other fields for their efficacy in helping to understand many biological systems. Here we are concerned with cellular biochemical networks. Mathematical and statistical network modeling is an important step toward uncovering the organizational principles and dynamic behavior of such networks.

This chapter focuses on methods of constructing discrete dynamic models of biochemical networks from high-throughput experimental data sets, also sometimes referred to as top-down modeling or reverse-engineering. The models discussed here are deterministic, and we will not discuss stochastic methods such as Markov chains and other graphical model approaches. Time-discrete dynamical systems models have long been used in biology, particularly in population dynamics. The models we mainly focus on here are also assumed to have a finite set of possible states for each variable. Boolean networks are an example, using only two possible states for each variable, but one strength of our method is its use of much larger possible state sets that capture more variation in the data. This assumption requires that all experimental measurements, which are real-valued, be first discretized into a finite number of classes.

Because we need to use time series of measurements to make dynamic models and might want to use heterogeneous data sets, great care must be taken during this step so as not to lose too much information. The resulting models will have a lower resolution than, say, ODE models. However, in exchange they are sometimes easier to analyze. We see an important role for discrete models to provide constraints on the structure and dynamics of higher-resolution continuous models. In the language of Ideker and Lauffenburger (2003), discrete models are more high-level than ODE and PDE models.

After a short survey of discrete finite-state modeling frameworks and methods, we present a detailed description of a multi-state modeling technology that has a strong mathematical underpinning, providing mathematical and computational tools for model selection and analysis. We then discuss the issue of linking discrete high-level models with continuous low-level ones. Finally, we exploit the analogy of top-down modeling to the process of system identification in engineering and applied mathematics to outline some steps in a modeling program for cellular pathways.

II. TOP-DOWN MODELING

Traditionally, models of molecular regulatory systems in cells have been created bottom-up, where the model is constructed piece-by-piece by adding new components and characterizing their interactions with other molecules in the model. This process requires that the molecular interactions have been well characterized, usually through quantitative numerical values for kinetic parameters. Note that the

construction of such models is biased toward molecular components that have already been associated with the phenomenon. Still, modeling can be of great help in this bottom-up process, by revealing whether the current knowledge about the system is able to replicate its *in vivo* behavior.

There are many good examples of this process. Teusink et al. (2000) have built a comprehensive model of yeast glycolysis based on detailed kinetics of 15 enzymes of carbohydrate catabolism. Arkin et al. (1998) studied stochastic switching between lysis and lysogeny in a model of lambda phage infection. In a landmark paper, Bray et al. (1993) studied the regulation of chemotactic swimming of *E. coli* cells, correlating the model to the phenotypes of dozens of mutants. For an example of bottom-up modeling of a problem involving spatial distributions of signaling molecules, we refer to a study of calcium waves in neuroblastoma cells by Fink et al. (2000).

Bottom-up modeling is essentially a process of synthesis by which models of isolated cellular components (enzymes, and so on) are merged to become part of a larger model. Note that without applying other steps models built bottom-up are mechanistic (i.e., represent one level of organization with all of the details of the level below). For example, the model of ethanol catabolism mentioned previously contains details of enzyme action of each of its 15 component enzymes.

This modeling approach is well suited to complement experimental approaches in biochemistry and molecular biology, in that models thus created can serve to validate the mechanisms determined *in vitro* by attempting to simulate the behaviors of intact cells. Although this approach has been dominant in cellular modeling, it does not scale very well to genome-wide studies because it requires that proteins be purified and studied in isolation. This is not a practical endeavor due to its large scale, but especially because a large number of proteins act on small molecules that are not available in purified form, as would be required for *in vitro* studies.

With the completion of the human genome sequence and the accumulation of other fully sequenced genomes, research is moving away from the molecular biology paradigm to an approach characterized by large-scale molecular profiling and *in vivo* experiments (or if not truly *in vivo* at least carried out with intact cells). Technologies such as transcript profiling with microarrays, protein profiling with 2-D gels and mass spectrometry, and metabolite profiling with chromatography and mass spectrometry produce measurements that are large-scale characterizations of the state of the biological material probed.

Other new large-scale technologies are also able to uncover groups of molecules that interact (bind), allowing inference of interaction networks. All of these experimental methods are data rich, and some people have recognized (e.g. Loomis and Sternberg 1995; Brenner 1997; Kell 2004) that modeling is necessary to transform these data into knowledge. A new modeling approach is needed to best suit large-scale profiling experiments. Such a top-down approach will start with little knowledge about the system, capturing at first only a coarse-grained image of the system with only a few variables. Then, through iterations of simulation and experiment, the number of variables in the model is increased. At each iteration, novel experi-

ments will be suggested by simulations of the model, which when carried out will provide data to improve the model further, leading to a higher resolution in terms of mechanisms.

Although the processes of bottom-up and top-down modeling are distinct, both have as an objective the identification of molecular mechanisms responsible for cell behavior. The main difference between the two is that the construction of top-down models is biased by the data of the large-scale profiles, whereas bottom-up models are biased by the pre-existing knowledge of particular molecules and mechanisms.

Note that although top-down modeling makes use of genome-wide profiling data it is conceptually very different from other genome-wide data analysis approaches. Top-down modeling needs data produced in experiments that lend themselves to the approach—most likely those designed with that purpose in mind. One should not expect that a random combination of arbitrary molecular snapshots would be of much use for the top-down modeling process. Sometimes they may serve some purpose (e.g., variable selection), but overall, top-down modeling requires perturbation experiments that are carried out with appropriate controls. In the face of modern experimental research methods, the development of an effective top-down modeling strategy is crucial. In addition, we believe that a combination of top-down and bottom-up approaches will eventually have to be used. An example of a first step in this direction is the apoptosis model in Bentele et al. (2004).

III. DISCRETE MODELING METHODS

A. Boolean networks

All top-down discrete modeling methods explored so far have some similarities. They all essentially take the view of a biochemical network as an information-processing system. Each method settles on a particular modeling framework, such as graphical models, Boolean networks, multistate models of a certain type. The resulting model space is then searched to find a model that best fits the given experimental data. Typically, the model space is quite large and different methods employ different “coping strategies” for model selection, such as imposing additional constraints on the models (e.g. sparseness) or random search strategies. As mentioned in the introduction, our focus here is on deterministic dynamical systems models. We survey some of the discrete modeling methods proposed to date, and describe in detail a method whose characteristic is a rigorous mathematical description of the entire model space, together with a mathematical model selection method that takes into account the entire model space. The goal of this section is not to be comprehensive, but to provide a context for the problems that all of these methods face, which are discussed in later sections of this chapter.

The most common approach to the modeling of biochemical regulatory networks is through systems of ordinary differential equations; that is, time-continuous dynamical systems. In 1969, S. Kauffman proposed to model regulatory networks

as logical switching networks, described as Boolean networks (Kauffman 1969). Boolean network models have the advantage of being more intuitive than ODE models, and might be considered as a coarse-grained approximation of the “real” network. They differ from ODE models in that molecules are considered present or absent, rather than ranging over a continuum of values. There is increasing evidence that certain types of regulatory networks have key features that can indeed be represented well through Boolean models (Davidson 2002; Wang et al. 2002; Fischle et al. 2003). Kauffman’s early work has generated a substantial literature on the subject (e.g. Raeymaekers 2002; Sabatti et al. 2002; Albert and Othmer 2003; Kauffman et al. 2004).

Top-down modeling methods using the Boolean framework have been proposed by Liang et al. (1998), Akutsu et al. (1999), and Akutsu et al. (2000). To include stochastic features of gene regulation, probabilistic Boolean networks have been introduced by Shmulevich et al. (2002b). The issue of how the Boolean framework can deal with experimental and biological noise was also addressed by Akutsu et al. (2000).

B. Multi-state discrete models

One of the disadvantages of the Boolean modeling framework is the need to discretize real-valued expression data into an ON/OFF scheme, which loses a large amount of information. Figure 12.1 shows mRNA concentrations of a gene regulatory network simulated with the biochemical network simulator Gepasi (Mendes 1997) on the left. The right side of Figure 12.1 shows two different discretizations: one Boolean and the other allowing 11 possible states.

This example makes it clear that in many cases a finer data discretization is needed in order for a model to capture the essential dynamic features contained in a multivariate data set. Partly in response to this deficiency, multi-state discrete modeling frameworks and hybrid models have been developed. One of the most complex ones (Thomas 1991; Thieffry and Thomas 1998) uses multiple states for the genes in the network corresponding to certain thresholds of gene expression that make possible multiple gene actions. The model includes a mixture of multi-valued logical and real-valued variables, as well as the possibility of asynchronous updating of the variables. A top-down modeling method for this type of model was proposed by Thomas et al. (2004). A software package for analyzing this type of multi-state model is also available (de Jong et al. 2003).

Multiple discrete expression levels were also used in the reverse-engineering method of Repsilber et al. (2002), which uses a genetic algorithm to explore the parameter space of multistage discrete genetic network models. Although this modeling framework is more effective than Boolean networks in capturing the many characteristics of gene regulatory networks, it also introduces substantially more computational complications from a top-down modeling point of view. A hybrid modeling framework was introduced by Brazma and Schlitt (2003) that tries to capture discrete as well as continuous aspects of gene regulation. The authors’ finite-state linear model has a Boolean-network-type of control component, as well

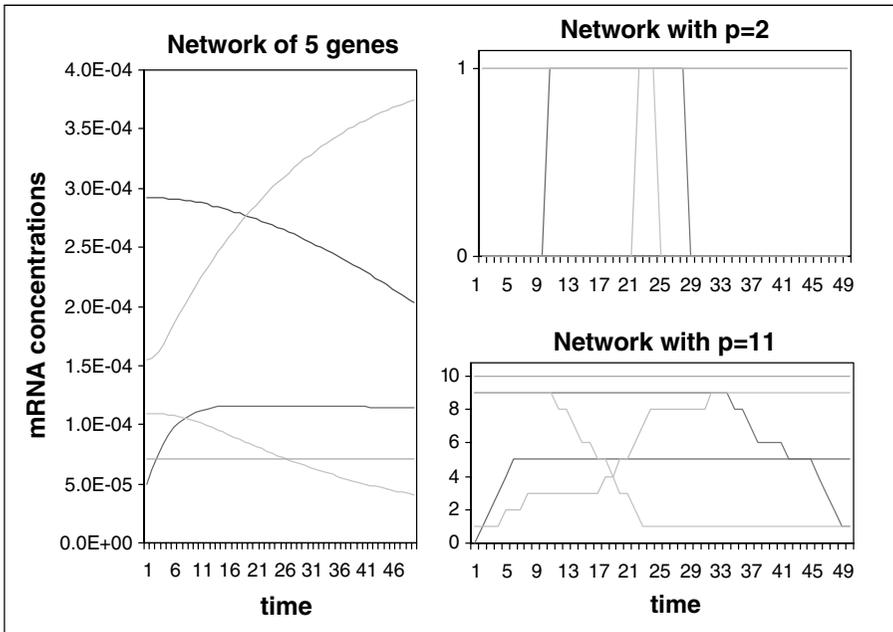


Figure 12.1. Different levels of data discretization.

as linear functions that represent substances that change their concentrations continuously. For a more comprehensive review of modeling methods, see de Jong (2002).

C. Finite-state polynomial models

We now describe a multi-state discrete model approach that leverages existing algorithmic methods from symbolic computation and computational algebraic geometry (Laubenbacher and Stigler 2004). It models a regulatory network as a time-discrete multi-state dynamical system, synchronously updated. The method shares many features with a recently developed continuous top-down method (Yeung et al. 2002), which we first describe in some detail. According to the authors, the method is intended to generate a “first draft of the topology of the entire network, on which further, more local, analysis can be based.” The authors make two assumptions. First, the system is assumed to be operating near a steady state, so that the dynamics can be approximated by a linear system of ordinary differential equations:

$$\frac{dx_i}{dt} = -\lambda_i x_i(t) + \sum_{j=1}^N w_{ij}(t) x_j(t) + b_i(t) + \xi_i(t),$$

for $i = 1, \dots, N$. Here, x_1, \dots, x_N are mRNA concentrations, the λ_i are the self-degradation rates, the b_i are the external stimuli, and the ξ_i represent noise. The (unknown) w_{ij} , which are assumed to be constant over time, describe the type and strength of the influence of the j th gene on the i th gene. They assemble to a square matrix W of real numbers. The output of the reverse-engineering algorithm is this matrix W . The input is a series of data points obtained by applying the stimulus $(b_1, \dots, b_N)^T$ and measuring the concentrations x_1, \dots, x_N M times. Assembling these measurements into a matrix $X = X(t)$, neglecting noise, and absorbing self-degradation into the coupling constants w_{ij} , we obtain a matrix equation

$$\frac{d}{dt}(X) = WX + B.$$

Here, X is an $(N \times M)$ -matrix, W an $(N \times N)$ -matrix, and B an $(N \times M)$ -matrix. Using singular value decomposition (SVD), one obtains

$$X^T = UWW^T,$$

where U and V are orthogonal to each other. The first step is to obtain a particular solution W_0 to the reverse-engineering problem. One then obtains all possible solutions to the problem as

$$W = W_0 + CV^T,$$

where C ranges over the space of all square $(N \times N)$ -matrices whose entries are equal to 0 for a certain range of j and arbitrary otherwise. Equivalently, CV^T ranges over all matrices that vanish on the given time points. The second assumption made in the paper is that gene regulatory networks are sparse. This provides a selection criterion on which to base a particular choice for C , and hence for W . The method selects the sparsest connection matrix W . This is accomplished through a particular choice of norm and robust regression. The algorithm was validated by way of data from three simulated networks.

The modeling framework for the discrete analog of this method is that of time-discrete dynamical systems over a finite state set X . Here, X is to be thought of as the set of discretized experimental values. For instance, in the Boolean case we have $X = \{0, 1\}$. To be precise, a dynamical system of dimension n over X is a function

$$f: X^n \rightarrow X^n$$

with dynamics generated by iteration of f . We will call f a *finite dynamical system*. Here, X^n denotes the set of all n -tuples with entries in X . Abbreviate an n -tuple (x_1, \dots, x_n) by \mathbf{x} . The function f is determined by its coordinate functions $f_i: X^n \rightarrow X$; that is,

$$f(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x})).$$

Suppose that we are given one or more time series of state transitions, measuring concentrations of mRNA, proteins, or metabolites. Our goal is to choose a finite

dynamical system $f: X^n \rightarrow X^n$, which fits the data and “best describes” the network that generated the data. To be precise, we assume that we are given sequences of states

$$\begin{aligned} s_1 &= (s_{11}, s_{21}, \dots, s_{n1}), \dots, s_m = (s_{1m}, \dots, s_{nm}) \\ t_1 &= (t_{11}, t_{21}, \dots, t_{n1}), \dots, t_r = (t_{1r}, \dots, t_{nr}) \\ &\dots \end{aligned}$$

These satisfy the property that if the unknown transition function of the network is f then

$$\begin{aligned} f(s_i) &= s_{i+1}, \text{ for } i = 1, \dots, m - 1 \\ f(t_j) &= t_{j+1}, \text{ for } j = 1, \dots, r - 1 \\ &\dots \end{aligned}$$

Typically, there will be more than one possible choice. In fact, unless *all* state transitions of the system are specified there will always be more than one model that fits the given data set. Because this much information is hardly ever available in practice, any top-down modeling method has to choose from a large set of possible models. As with most methods, ours will also choose the simplest model, in a certain sense. Before describing the selection principle used, we first need to describe the computational framework.

If we do not impose any further mathematical structure, we are left with a problem about set functions. No systematic computational tools for finding dynamical systems that fit the data (and for choosing a particular one) are available in this general setting. The standard mathematical solution is to endow the model space with a suitable additional mathematical structure. One way to do this is by a process analogous to the imposition of a coordinate system onto an affine space, resulting in an algebraic structure on the set of points in the space. Precisely, we assume that our set X is equipped with the structure of a finite number system; that is, a *finite field*.

It is well-known that this can be done whenever the number of elements in X is a power of a prime number p . This assumption is a straight-forward generalization of the Boolean case, where we can take advantage of Boolean arithmetic (e.g., $1 + 1 = 0$). Because the cardinality of X depends on the resolution of the discretization we choose, this is an easy assumption to satisfy in practice by refining the resolution, if needed. One possible approach is to choose a prime number p of possible variable states, in which case the number system can be taken to be \mathbf{Z}/p , the integers modulo p .

An important consequence of this assumption is the well-known fact (Lidl and Niederreiter 1997, p. 369) that each of the coordinate functions of f can be expressed as a polynomial function in n variables, with coefficients in X , and so that the degree of each variable is less than the number of elements in X . For instance, each Boolean function can be expressed as a polynomial, via the correspondence

$x \wedge y = xy$, $x \vee y = x + y + xy$, and $\neg x = x + 1$. In other words, polynomial dynamical systems can serve as a computational model for all finite dynamical systems over a finite field. We are now in a position to use the rich algorithmic theory of polynomial algebra that has been developed over the last 20 years (Cox et al. 1997), including sophisticated symbolic computation software. Thus, we can overcome one disadvantage that discrete models have compared to ODE models, for which there is a mature mathematical theory available.

Thus, assume now that our state set X is a finite field. The model $f: X^n \rightarrow X^n$ we are searching for is determined by its coordinate functions $f_i: X^n \rightarrow X$. We can reverse engineer each coordinate function independently and thus reconstruct the system one variable at a time. The strategy of the method is to first compute the space of all systems that are consistent with the given time series data. The core of this computation is an interpolation algorithm. The method then chooses a particular system $f = (f_1, \dots, f_n)$ that satisfies the following property.

Minimality: For each i , f_i is minimal in the sense that there is no non-zero polynomial g such that $f = h + g$ and g is identically equal to zero on the given time points. That is, we exclude terms in the polynomials f_i that vanish identically on the data. In other words, we do not include interactions in the model that are not manifest in the given data set.

Suppose that f_i and f'_i are two models that fit the given data set. Then, $f_i(\mathbf{x}) = f'_i(\mathbf{x})$ for all data points \mathbf{x} . That is, $(f_i - f'_i)(\mathbf{x}) = 0$ for all \mathbf{x} . Therefore, the set of all such models can be described as $f_i + I$, where f_i is a particular model and I is the set of all models that vanish identically on the given data set. In other words, the situation is very similar to the case of solving a nonhomogeneous system of linear equations, where f_i represents a particular solution to the system and I represents the solution space of the corresponding homogeneous system. The correspondence with the ODE modeling method described by Yeung et al. (2002) is that f_i corresponds to W_0 and I corresponds to the space C . Thus, we need to compute f_i and I .

The particular solution f_i can be computed using a standard formula for Lagrange interpolation (see Laubenbacher and Stigler (2004) for details). To compute I we use mathematical algorithms from computer algebra based on the theory of Groebner bases (Cox et al. 1997). What allows us to do this is the fact that the set of polynomials that vanish on a given data set has the algebraic structure of an *ideal* in the algebraic system $X[x_1, \dots, x_n]$ of all polynomials in n variables with coefficients in X . These algorithms are implemented using the computer algebra system Macaulay2 (Grayson and Stillman, 2003). An important aspect of this computation is that the set of all possible models is described not by enumeration but in terms of a small set of generators, similar to describing a vector space by giving a basis for it. The algorithm to select the simplest model from the set $f_i + I$ uses another fundamental procedure in computer algebra: dividing a polynomial by all polynomials in the ideal I .

One can prove that there is in fact a unique simplest model to choose. However, the algorithm of Laubenbacher and Stigler (2004) depends on an up-front choice

of a total ordering of the variables x_1, \dots, x_n . This choice has the effect that the algorithm uses the "cheapest" (smallest, in this ordering) variables preferentially. On the one hand, this feature allows the incorporation of biological knowledge in the case where certain interactions are already known. On the other hand, it arbitrarily biases the model output in the case where such information is absent.

In Laubenbacher and Stigler (2004), several variable orders were used and common terms in the polynomial models for each order were extracted to circumvent this problem. We briefly describe the validation of this approach. In the absence of a published large multi-state discrete model we used a Boolean model instead. The goal of this validation is not to make statements about the Boolean model and its validity, but rather to test how well the polynomial method is able to recover the Boolean model. Albert and Othmer (2003) presented a Boolean model for a well-characterized network of segment polarity genes in *Drosophila melanogaster*. The network, consisting of five genes and their products, is responsible for pattern formation in the *Drosophila* embryo. The network is a ring of 12 interconnected cells, in which the genes are expressed in patterns resembling stripes. The genes represented in the Albert-Othmer model are *wingless*, *engrailed*, *hedgehog*, *patched*, and *cubitus interruptus*.

The proposed model is a collection of 21 Boolean functions, representing the genes and proteins in the network. Each function governs the state transitions of a single compound. The following are four of the functions defined in the model.

$$f_6 = hh_i^{t+1} = EN_i^t \wedge \neg CIR_i^t$$

$$f_7 = HH_i^{t+1} = hh_i^t$$

$$f_8 = ptc_i^{t+1} = CIA_i^{t+1} \wedge \neg EN_i^{t+1} \wedge \neg CIR_i^{t+1}$$

$$f_9 = PTC_i^{t+1} = ptc_i^t \vee (PTC_i^t \wedge \neg HH_{i-1}^t \wedge \neg HH_{i+1}^t)$$

Representing each biochemical with a variable, the Boolean functions may be translated into polynomial functions, shown below.

$$f_6 = x_5(x_{15} + 1)$$

$$f_7 = x_6$$

$$f_8 = x_{13}((x_{11} + x_{20} + x_{11}x_{20}) + x_{21} + (x_{11} + x_{20} + x_{11}x_{20})x_{21}) \\ (x_4 + 1)(x_{13}(x_{11} + 1)(x_{20} + 1)(x_{21} + 1) + 1)$$

$$f_9 = x_8 + x_9(x_{18} + 1)(x_{19} + 1) + x_8x_9(x_{18} + 1)(x_{19} + 1)$$

Treating this Boolean model as "reality," wild-type and simulated knock-out experiments were generated, creating knock-outs by setting a function representing a gene equal to 0. As the algorithm relies on the choice of an ordering of the variables, causing some variables to have greater weight than the rest, four variable orders were used to counteract this preferential ranking.

Not surprisingly, algorithm performance improved greatly with knock-out data rather than just wild-type data. The algorithm is able to reconstruct approximately

84% of the interactions in the Boolean model, versus only 32% when only wild-type data were used. Furthermore, it correctly identified 92% of the additive interactions and 10% of the nonadditive interactions, whereas none of the nonadditive interactions were identified in the model constructed with only wild-type data.

A more elegant solution was proposed by Allen et al. (2005). Using a large number of randomly generated variable orders to generate models, the authors then rank the variables according to their frequency of appearance in the models for each of these variable orders. This ranking then determines a variable ordering to be used for the final model construction.

Another shortcoming of the algorithm of Laubenbacher and Stigler (2004) is that it relies on exact fitting of data. This makes the method very sensitive to noise that is known to be present in DNA microarray and other “-omics” data. To avoid models that are overly complex due to fitting of noise, the Laubenbacher group is presently developing a genetic algorithm that optimizes between data fit and model complexity. An important feature of the algorithm is that its performance is substantially improved by supplying as initialization the output of the exact data-fitting algorithm described previously versus a random initialization. The key theoretical ingredient in the algorithm is a mathematical characterization of the evolution rules to guarantee that each mutation still satisfies the minimality criterion imposed.

An important tool for working with polynomial models over finite fields is the software package DVD (available at <http://dvd.vbi.vt.edu> as a web interface or for download). The program takes a polynomial system as input. For binary systems, one can also input Boolean functions, which are then translated into polynomial functions. DVD then computes the phase space of the system and outputs statistics such as the number of components, length of limit cycles, and so on. It also outputs the wiring diagram of the system. For small systems, it visualizes the phase space. Figure 12.2 shows the DVD interface.

IV. DATA DISCRETIZATION

The very important issue of data discretization has been studied from the points of view of Bayesian network applications and machine learning (Dougherty et al. 1995; Friedman and Goldszmidt 1996). The first important choice to make is the number of discrete states to use. The second choice is the method by which to map real-valued data to discrete states. There are various ways of labeling real-valued data using finite-state sets. Thresholds with biological relevance are one type of labeling that can be used. This is typically referred to as binning. For example, up-regulation, no regulation, and down-regulation of a gene may be used as thresholds for partitioning the raw data into three groups, labeled 1, 0, and -1, respectively. For binary states, the choice of threshold is particularly crucial, in that even a relatively small change can result in very different discrete time series profiles (Sabatti et al. 2002). Another method of discretization is to normalize the expres-

Discrete Visualizer of Dynamics (DVD) v1.0

If this is your first time, please read the [tutorial](#). It is important that you follow the format specified in the tutorial.
Make your selections and provide inputs (if any) in the form below and click Generate to run the software.
Note: The computation may take some time depending on your internet connection.

Network Description

Enter number of nodes: what is this?

Enter number of states per node: what is this?

Select format of input functions: what is this?

Polynomial

Boolean

Select the updating scheme for the functions: what is this?

Synchronous

Sequential

- Enter update schedule separated by spaces:

Input Functions

Select function file: what is this?

OR (Edit functions below)

```
f1 = (x1+x2)
f2 = (x1*(x2+x3)) * (x4)
f3 = (x4)
f4 = (x1*x4)
```

State Space Specification

Generate state space of what is this?

All trajectories from all possible initial states

One trajectory starting at an initial state

- Enter initialization separated by spaces:

Additional Output Specification (optional)

View what is this?

Select graph(s) to view and image format.

State space graph

Dependency graph

Results will be displayed below.

ANALYSIS OF THE STATE SPACE [m = 2, n = 4]
There are 3 components and 2 fixed point(s)

Components	Size	Cycle Length
1	3	1
2	7	1
3	6	2

TOTAL: 16 = 2^mn nodes

Printing fixed point(s)...

[0 0 0 0] lies in a component of size 3

[0 0 1 1] lies in a component of size 7.

[Click to view the state space graph.](#)

[Click to view the dependency graph.](#)

Figure 12.2. Snapshot of DVD interface.

sion of each gene or protein and use the deviation from the mean to discretize the data.

Any discretization method suitable for our purposes must preserve information about the dynamic relationship between the different variables, and must accommodate several heterogeneous time series simultaneously (e.g., transcription data as well as protein and metabolite concentrations). We have developed a method based on a graph theoretic approach that has the important advantage that the algorithm chooses an optimal number of states, based on the given data (Dimitrova et al. 2005). Most discretization methods require such a choice as part of the input. The algorithm has been implemented in C++ and is freely available. We illustrate it with an example.

Consider the simulated gene regulatory network shown in Figure 12.3 (five genes, whose wiring diagram is given in Figure 3a). The network was generated with the artificial gene network system AGN (Mendes et al. 2003). After simulating the network with the biochemical network simulator Gepasi (Mendes 1997), one finds that it has the positive stable steady state (1.99006, 1.99006, 0.000024814, 0.997525, 1.99994). From the model, we generate six time series, each of length 20, including one wild-type time series and five deletion mutant time series. The discretization algorithm chooses the number system $X = \{0, 1, 2, 3, 4\}$, consisting of five different states for the combined data set.

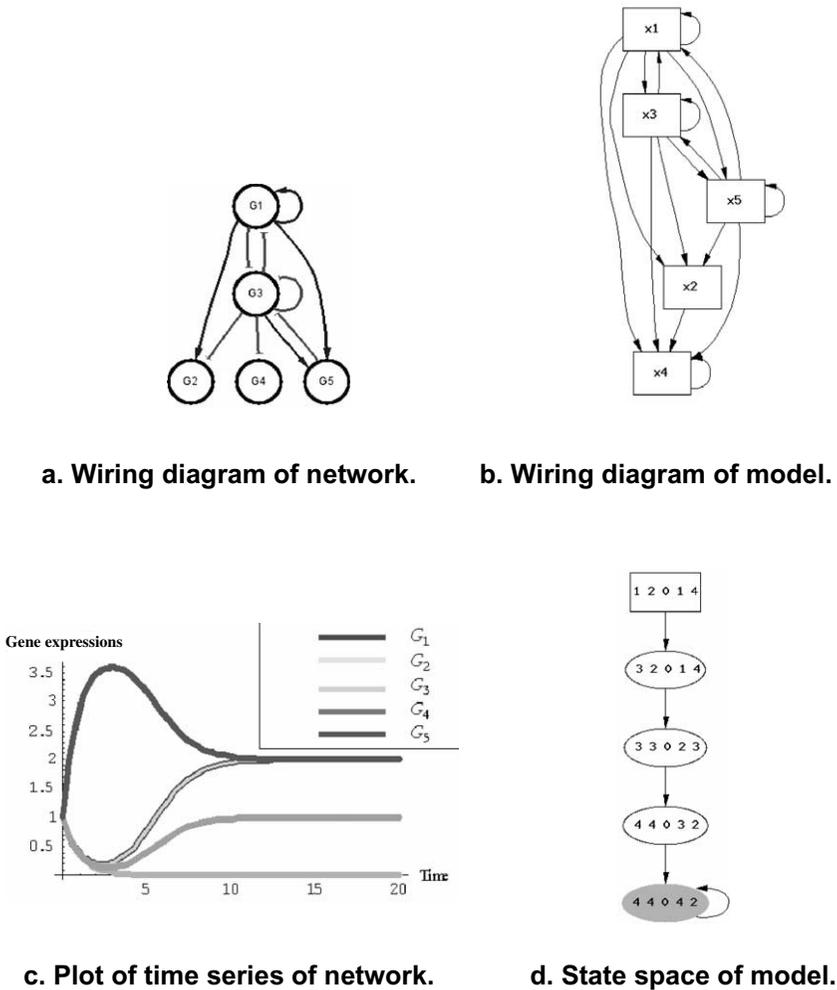


Figure 12.3. Graphs of a network and its associated models. (a) Wiring diagram of network, (b) wiring diagram of model, (c) plot of time series of network, and (d) state space of model (see color plate 8).

After using the multivariate interpolation algorithm, we obtain a “best” polynomial model $f: X^5 \rightarrow X^5$ in five variables. Its phase space consists of a directed graph whose nodes are the 5^5 possible states for the five variables, and there is a directed edge from state **a** to state **b** if $f(\mathbf{a}) = \mathbf{b}$. The model also has a fixed point, like the continuous “real-world” system. Figure 3c shows a particular initialization of the network, simulated in Gepasi, reaching the previously cited steady state. Figure 3d shows a sample of the time series obtained by initializing the discrete model f with the discretization of this initialization. It converges to the discretization (4, 4, 0, 4, 2) of the steady state cited previously.

This example illustrates the fact that the discrete model f exhibits the same qualitative dynamics as the continuous model we started with. Figure 12.3b shows the wiring diagram of the discrete model obtained with our algorithm. The main point of this example is to demonstrate that our discretization method preserves the essential dynamic features of the continuous system representing “reality” in this case, and our interpolation algorithm chooses a model that reflects these dynamic features as well as most of the causal dependencies among the variables.

V. RELATIONSHIP BETWEEN DISCRETE AND CONTINUOUS MODELS

The relationship between discrete and continuous models has been studied extensively in population dynamics (Durrett and Levin 1994; Henson et al. 2001; Domokos and Scheuring 2004; Geritz and Kisdi 2004). For models of biochemical and other biological networks, this relationship was first explored by Glass and Kauffman (1973), with subsequent work by Edwards (2000), Edwards et al. (2001), and Glass et al. (2003). Within the modeling frameworks explored there, (bottom-up) discrete models can be a helpful tool to provide constraints and information about (bottom-up) continuous models of the same network. A good example of how a continuous and a discrete model of the same system can be used together is given by Muraille et al. (1996), where an ODE model of immune response to a replicating pathogen is studied via a discrete logical model using the technique of Thomas (1991). The dynamics of the discrete model, which are easy to analyze, are used to obtain a qualitative picture of the dynamics of the ODE model.

A corresponding mathematical theory for top-down modeling has yet to be developed. How can high-level information from discrete multi-state dynamic models of a network be incorporated into the model selection process for low-level ODE models? For the polynomial system framework described here, we are developing such a theory in parallel with an ODE framework based on a linearization of the dynamics (i.e., the Jacobian, a first-order truncation of the Taylor approximation to the dynamics).

Estimates of the elements of the Jacobian matrix are currently pursued through non-linear least squares. Our aim is to develop ways in which these top-down approaches become synergistic. In particular, we expect the results of the discrete model to be used as initial states for the parameter estimation needed to define a continuous model. We are currently carrying out experiments that will be used to validate both methods, using integrated transcriptomics, proteomics, and metabolomics time courses measuring oxidative stress response in *Saccharomyces cerevisiae*.

VI. A MATHEMATICAL THEORY FOR DISCRETE MODELS

Discrete models are not well understood at a theoretical level. In particular, the relationship between the structure of a model and its dynamics has remained elusive.

There are no general results about the number of components of the state space of Boolean or multi-state discrete models or about the existence of steady states. Especially the question of steady states is an important one for biological models. Having fairly general results about the relationship between structure and dynamics for sufficiently large classes of models is an important problem.

Not surprisingly, these questions can be answered algorithmically for linear systems. Let X be a finite field and $f: X^n \rightarrow X^n$ a linear system. That is, the coordinate functions of f are linear polynomials without constant term. Then f can be represented by a matrix after making a choice of basis. It turns out that the structure of the phase space of f can be completely determined from the factorization of the characteristic polynomial of f , in particular the number of components and the length of all limit cycles (Hernandez Toledo 2003).

Very few results are available for nonlinear systems. A modest first step toward general results for sufficiently large classes of polynomial systems has been made by Colon-Reyes et al. (2004). Suppose that f is a Boolean polynomial system all of whose coordinate functions consist of monomials; that is, f is constructed using the AND operator. Let G be the directed graph whose vertices are the variables of f . There is a directed edge from x_i to x_j if x_j appears in f_i . Reversing the arrows of G , one obtains the wiring diagram of the network. One can define a positive integer, the *loop number* of G , which can be computed in polynomial time (relative to the number of vertices in G). The main result of Colon-Reyes et al. (2004) is that f has only steady states if and only if the loop number of all strongly connected components of G is equal to 1.

VII. TOWARD A MATHEMATICAL THEORY OF BIOLOGICAL SYSTEM IDENTIFICATION

The basic inverse problem we face in modeling biochemical networks is common in engineering and applied mathematics, known as system identification. Our goal is to make a phenomenological (and, ultimately, mechanistic) mathematical model of a multivariate system we can observe as well as perturb, and about which we may have partial knowledge. The major challenges, compared to typical engineered systems, are that the system is very often high-dimensional, the number of observations is small in comparison, and the information we have about the systems is very limited.

The basic procedure is to choose an appropriate modeling framework, use one or more time series of observations to identify some or all possible models within this framework, and choose the “best” one from the possible model space. For engineered systems there is a well-developed mathematical theory that helps in this process. (An important application is the development of controllers for systems.) In particular, there is a theory of system identifiability, which provides criteria for how good a given data set is for the system identification process (Ljung 1999) for a comprehensive treatment of system identification.

No corresponding mathematical theory exists yet for the identification of biological systems. In particular, there is no good understanding about the appropriate experimental design for a particular modeling framework that provides good data sets for top-down modeling. The most commonly studied type of systematic perturbation focuses on single genes in regulatory networks (Karp et al. 1999; Ideker et al. 2000; Rung et al. 2002; Shmulevich et al. 2002a; Tegner et al. 2003). Genetical genomics provides another possible approach (Jansen 2003). Studies of the quantity of data needed have been done by Krupa (2002) and Selinger et al. (2003). The study of appropriate experimental designs for various modeling methods must be part of a long-term systems biology modeling program.

VIII. CONCLUSIONS

We have discussed some top-down modeling methods resulting in time-discrete dynamical system models over finite-state sets. They serve to provide high-level information about systems that can be used as constraints for the construction of low-level models, either top-down or bottom-up. Our method using polynomial dynamical systems over finite fields has the advantageous feature that its mathematical underpinning provides access to a variety of mathematical algorithms and symbolic computation software. Other modeling approaches discussed here have other advantages, such as the capability of asynchronous update or the incorporation of discrete as well as continuous variables. The choice of what method to use in a particular case will depend on the type of data and information available. For instance, the incorporation of asynchronous update is only feasible computationally with prior biological information on the pathways to be modeled. In particular, it provides a mathematical basis for the investigation of questions such as "goodness" measures on data sets. Ultimately, the performance of top-down modeling methods cannot be properly evaluated unless we understand what types of input data are required for optimal performance. That is, the "data must fit the models".

Experimental data sets suitable for the various modeling methods are still difficult to obtain, and the biochemical networks producing the data are typically too poorly understood to truly test modeling performance. An important resource in the field would be a collection of benchmark synthetic biochemical networks and the ability to generate from them data sets covering various types of networks, providing wild-type and perturbation time series. One possible tool for generating such networks and data is described by Mendes et al. (2003).

We believe that the field of system identification can serve as a blueprint for a mathematical top-down modeling program in systems biology. Based on a well-defined collection of model classes, from high-level statistical models down to ODE and PDE models, such a program must include the development of appropriate system identification methods for each model class and quality measures on data sets that can be used to develop confidence measures for the resulting models.

ACKNOWLEDGMENTS

This work was partially supported by NIH grant RO1 GM068947-01. The authors thank E. Dimitrova, A. Jarrah, D. Potter, B. Stigler, J. Tyson, and P. Vera-Licona for help in preparing this manuscript.

REFERENCES

- Akutsu, T., Miyano, S., et al. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 17–28.
- Akutsu, T., Miyano, S., et al. (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.* 7(3/4):331–343.
- Akutsu, T., Miyano, S., et al. (2000). Algorithms for inferring qualitative models of biological networks. *Pac. Symp. Biocomput.* 293–304.
- Akutsu, T., Miyano, S., et al. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* 16(8):727–734.
- Albert, R., and Othmer, H. G. (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.
- Allen, E. E., Fetrow, J. S., et al. Algebraic dependency models of protein signal transduction networks from time-series data. *J Theor Biol* (in press).
- Arkin, A., Ross, J., et al. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 149(4):1633–1648.
- Bray, D., Bourret, R. B., et al. (1993). Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. *Mol. Biol. Cell* 4(5):469–482.
- Brazma, A., and Schlitt (2003). Reverse engineering of gene regulatory networks: a finite state linear model. *Genome Biology* 4(6).
- Brenner, S. (1997). *Loose Ends*. London, Current Biology 73.
- Colon-Reyes, O., Laubenbacher, R., et al. (2004). Boolean monomial dynamical systems. *Annals of Combinatorics* 8:426–439.
- Cox, D., Little, J., et al. (1997). *Ideals, Varieties, and Algorithms*. New York: Springer-Verlag.
- Davidson, E. H., et al. (2002). A genomic regulatory network for development. *Science* 295:1669–1678.
- de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* 9(1):67–103.
- de Jong, H., Geiselman, J., et al. (2003). Genetic Network Analyzer: Qualitative simulation of genetic regulatory networks. *Bioinformatics* 19(3):336–344.
- Dimitrova, E., McGee, J., et al. (2005). A graph-theoretic method for the discretization of gene expression measurements. (Under Review)
- Domokos, G., and Scheuring, I. (2004). Discrete and continuous state population models in a noisy world. *J. Theor. Biol.* 227:535–545.
- Dougherty, J., Kohavi, R., et al. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann.

- Durrett, R., and Levin, S. (1994). The importance of being discrete (and spatial). *Theo. Population Biol.* **46**:363–394.
- Edwards, R. (2000). Analysis of continuous-time switching networks. *Physica* **146**:165–199.
- Edwards, R., Siegelmann, H. T., et al. (2001). Symbolic dynamics and computation in model gene networks. *Chaos* **11**:160–169.
- Fink, C. C., Slepchenko, B., et al. (2000). An image-based model of calcium waves in differentiated neuroblastoma cells. *Biophys. J.* **79**(1):163–183.
- Fischle, W., Wang, Y., et al. (2003). Binary switches and modification cassettes in histone biology and beyond. *Nature* **425**:475–479.
- Friedman, N., and Goldszmidt, M. (1996). Discretization of continuous attributes while learning Bayesian networks. In *Proceedings of the 13th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann.
- Geritz, S. A. H., and Kisdi, E. (2004). On the mechanistic underpinning of discrete-time population models with complex dynamics. *J. Theo. Biol.* **228**:261–269.
- Glass, K., Xia, Y., et al. (2003). Interpreting time-series analyses for continuous-time biological models: Measles as a case study. *J. Theo. Biol.* **223**(1):19–25.
- Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, nonlinear biochemical control networks. *J. Theo. Biol.* **39**:103–129.
- Grayson, D. R., and Stillman, M. E. (2003). <http://www.math.uiuc.edu/Macaulay2>.
- Henson, S. M., Costantino, R. F., et al. (2001). Lattice effects observed in chaotic dynamics of experimental populations. *Science* **294**:602–605.
- Hernandez, Toledo, R. A. (2003). Linear finite dynamical systems. *Comm. Algebra* (in press).
- Ideker, T. E., and Lauffenburger, D. (2003). Building with a scaffold: Emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology* **21**(6):256–262.
- Ideker, T. E., Thorsson, V., et al. (2000). Discovery of regulatory interaction through perturbation: Inference and experimental design. *Pac. Symp. Biocomput.* **5**:302–313.
- Jansen, R. C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4**:145–151.
- Karp, R. M., Stoughton, R., et al. (1999). Algorithms for choosing differential gene expression experiments. RECOMB99.
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theo. Biol.* **22**(3):437–467.
- Kauffman, S. A., Peterson, C., et al. (2004). Genetic networks with canalizing Boolean rules are always stable. *PNAS* **101**(49):17102–17107.
- Kell, D. B. (2004). Metabolomics and systems biology: Making sense of the soup. *Curr. Opin. Microbiol.* **7**(3):296–307.
- Krupa, B. (2002). On the number of experiments required to find the causal structure of complex systems. *J. Theo. Biol.* **219**(2):257–267.
- Laubenbacher, R., and Stigler, B. (2004). A computational algebra approach to the reverse engineering of gene regulatory networks. *J. Theo. Biol.* **229**:523–537.
- Liang, S., Fuhrman, S., et al. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 18–29.
- Lidl, R., and Niederreiter, H. (1997). *Finite Fields*. New York: Cambridge University Press.
- Ljung, L. (1999). *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall.
- Loomis, W. F., and Sternberg, P. W. (1995). Genetic networks. *Science* **269**(5224):649.

- Mayr, E. (1988). *Toward a New Philosophy of Biology*. Cambridge, MA: Harvard University Press.
- Mendes, P. (1997). Biochemistry by numbers: Simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* **22**(9):361–363.
- Mendes, P., Sha, W., et al. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**(2):II122–II129.
- Muraille, E., Thieffry, D., et al. (1996). Toxicity and neuroendocrine regulation of the immune response: A model analysis. *J. Theo. Biol.* **183**:285–305.
- Raeymaekers (2002). Dynamics of Boolean networks controlled by biologically meaningful functions. *J. Theo. Biol.* **218**:331–341.
- Repsilber, D., Liljenstrom, H., et al. (2002). Reverse engineering of regulatory networks: Simulation studies on a genetic algorithm approach for ranking hypotheses. *Biosystems* **66**(1/2):31–41.
- Rung, J., Schlitt, T., et al. (2002). Building and analysing genome-wide gene disruption networks. *Bioinformatics* **18**(2):S202–S210.
- Sabatti, C., Karsten, S. L., et al. (2002). Thresholding rules for recovering a sparse signal from microarray experiments. *Mathematical Biosciences* **176**:17–34.
- Selinger, D. W., Wright, M. A., et al. (2003). On the complete determination of biological systems. *Trends Biotech.* **21**(6):251–254.
- Shmulevich, I., Dougherty, E. R., et al. (2002a). Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics* **18**(10):1319–1331.
- Shmulevich, I., Dougherty, E. R., et al. (2002b). Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**(2):261–274.
- Tegner, J., Yeung, M. K., et al. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA* **100**(10):5944–5949.
- Teusink, B., Passarge, J., et al. (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.
- Thieffry, D., and Thomas, R. (1998). Qualitative analysis of gene networks. *Pac. Symp. Biocomput.* 77–88.
- Thomas, R. (1991). Regulatory networks seen as asynchronous automata: A logical description. *J. Theo. Biol.* **153**:1–23.
- Thomas, R., Mehrotra, S., et al. (2004). A model-based optimization framework for the inference on gene regulatory networks from DNA array data. *Bioinformatics* **20**(17):3221–3235.
- Wang, W., Cherry, J. M., et al. (2002). A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **99**(26):16893–16898.
- Yeung, M. K., Tegner, J., et al. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**(9):6163–6168.

Computational Models for Circadian Rhythms: Deterministic Versus Stochastic Approaches

Jean-Christophe Leloup, Didier Gonze,
and Albert Goldbeter

*Unité de Chronobiologie théorique, Faculté des
Sciences, Université Libre de Bruxelles, Brussels, Belgium*

Chapter 13

ABSTRACT

Circadian rhythms originate from intertwined feedback processes in genetic regulatory networks. Computational models of increasing complexity have been proposed for the molecular mechanism of these rhythms, which occur spontaneously with a period on the order of 24 h. We show that deterministic models for circadian rhythms in *Drosophila* account for a variety of dynamical properties, such as phase shifting or long-term suppression by light pulses and entrainment by light/dark cycles. Stochastic versions of these models allow us to examine how molecular noise affects the emergence and robustness of circadian oscillations. Finally, we present a deterministic model for the mammalian circadian clock and use it to address the dynamical bases of physiological disorders of the sleep/wake cycle in humans.

I. INTRODUCTION: THE COMPUTATIONAL BIOLOGY OF CIRCADIAN RHYTHMS

Most living organisms have developed the capability of generating autonomously sustained oscillations with a period close to 24 h. The function of these so-called *circadian rhythms* is to allow the organisms to adapt their physiology to the natural alternation of day and night. Circadian rhythms are endogenous because they can occur in constant environmental conditions (e.g., constant darkness). During the last two decades, experimental studies have shed much light on the molecular mechanism of circadian rhythms, which represents a long-standing problem in biology. In all eukaryotic organisms investigated so far, the molecular mechanism

of circadian oscillations relies on the negative feedback exerted by a clock protein on the expression of its gene (Hardin et al. 1990; Glossop et al. 1999; Lee et al. 2000; Alabadi et al. 2001; Reppert and Weaver 2002).

Even before details were known about their molecular origin, abstract mathematical models were used to probe the dynamic properties of circadian rhythms. A popular model of this type was provided by the van der Pol equations, which were originally proposed for sustained oscillations in electrical circuits. Thus, the van der Pol oscillator has been used for more than three decades for modeling circadian rhythms (e.g., to account for phase shifts of these rhythms by light pulses (Jewett and Kronauer 1998)). Another application involving this model pertains to modeling the enhanced fitness due to the resonance of circadian rhythms with the external light/dark cycle in cyanobacteria (Gonze et al. 2002c).

However, now that the molecular mechanism of circadian rhythms has largely been uncovered, mathematical models based on experimental observations have been proposed. Taking the form of a system of coupled ordinary differential equations, these deterministic models predict that in a certain range of parameter values the genetic regulatory network at the core of the clock mechanism can produce sustained oscillations of the limit cycle type. Deterministic models for circadian rhythms were first proposed for *Drosophila* and *Neurospora* (Goldbeter 1995, 1996; Leloup and Goldbeter 1998; Leloup et al. 1999; Smolen et al. 2001; Ueda et al. 2001), and later for mammals (Forger and Peskin 2003; Leloup and Goldbeter 2003, 2004; Becker-Weimann et al. 2004). The first model showing that oscillations can originate from negative feedback on gene expression was due to Goodwin (1965), who showed (already four decades ago) that periodic behavior may originate from such mode of genetic regulation. Modified versions of the Goodwin model are still being used to probe properties of circadian rhythms in organisms such as *Neurospora* (Ruoff et al. 2001). In this chapter we will focus on more recent models, which rely on more detailed molecular mechanisms.

One limitation of deterministic models is that they do not take into consideration the fact that the number of molecules involved in the regulatory mechanism within the rhythm-producing cells may be small as observed, for example, in *Neurospora* (Morrow et al. 1997). At low concentrations of protein or messenger RNA molecules, molecular fluctuations are likely to have a marked impact on circadian oscillations (Barkai and Leibler 2000). To assess the effect of molecular noise, it is necessary to resort to a stochastic approach. Comparing the predictions of deterministic and stochastic models for circadian rhythms shows that robust circadian oscillations can be observed even when the maximum number of mRNA and protein molecules is of the order of some tens and hundreds, respectively (Gonze et al. 2002a, 2002b, 2004a).

The goal of this chapter is to present an overview of deterministic and stochastic models for circadian rhythms. We will begin by presenting (in Section II) deterministic models for circadian oscillations of the PER protein and its mRNA in *Drosophila*. A core model will be presented, which also provides a useful model for circadian rhythms in *Neurospora*. This model for *Drosophila* circadian rhythms will

be extended to take into account the role of the TIM protein and the control of circadian behavior by light.

In Section III, we consider stochastic versions of these models. We examine how molecular noise affects the emergence of circadian oscillations and determine the influence of a variety of factors, such as number of protein and mRNA molecules, degree of cooperativity of repression, distance from bifurcation point, and rate constants characterizing the binding of the repressor protein to the gene. Two types of stochastic models are presented: one involves a fully detailed description of individual reaction steps, whereas a second relies on a non-developed description of nonlinear kinetic steps. Both types of models yield largely similar results. The study of stochastic models for circadian oscillations will allow us to characterize the domain of validity of deterministic models for circadian rhythms.

In Section IV we return to deterministic approaches and present a model for the mammalian circadian clock. We use this model to address the molecular bases of disorders of the sleep/wake cycle in humans, which are associated with dysfunctions of the clock. Computational models can thus be applied to investigating not only the molecular mechanism of circadian rhythms but the origin of associated physiological disorders. As discussed in Section V, the example of circadian rhythms illustrates how more and more complex models have been presented over the years to account for new experimental observations. We consider the need for such an increase in complexity of computational models for circadian rhythms, and the added insights these complex models provide for a better understanding of circadian behavior.

II. MODELING THE DROSOPHILA CIRCADIAN CLOCK

A. Overview of experimental observations

Some of the most remarkable advances in elucidating the molecular basis of circadian rhythms have been made in mutants of the fly *Drosophila* (Konopka 1979; Hall and Rosbash 1988; Baylies et al. 1993; Dunlap 1993), in which circadian rhythms affect the rest/activity cycle and the daily eclosion peaks of pupae. Both rhythms persist in constant darkness or temperature (Pittendrigh 1960). The classic work of Konopka and Benzer (1971) yielded *Drosophila* flies altered in their circadian system, owing to mutations in a single gene called *per* (for “period”). Four phenotypes were characterized: the wild type (per^+) has a free-running period of activity and eclosion close to 24 h; short-period mutants (per^s) have a period close to 19 h; in long-period mutants (per^l), the periodicity increases up to 29 h; and arrhythmic mutants (per^0) have lost the circadian pattern of eclosion or activity (Konopka and Benzer 1971; Konopka 1979). Interestingly, whereas in the wild type the period remains independent of temperature—a property known as temperature compensation, which is common to all circadian rhythms (Pittendrigh 1960)—the mutants per^l and per^s have lost this property (Konopka et al. 1989). In contrast to the wild

type, the period of their activity rhythm respectively increases and decreases with temperature. Accounting for temperature compensation of circadian rhythms remains an important challenge for computational biology.

A breakthrough for the mechanism of circadian rhythms in *Drosophila* was the finding (Hardin et al. 1990, 1992) that *per* mRNA is produced in a circadian manner. This periodic variation is accompanied by a circadian rhythm in the degree of abundance of PER. The peak in *per* mRNA precedes the peak in PER by 4 to 8 h (Zerr et al. 1990; Zeng et al. 1994). On the basis of this observation, Hardin et al. (1990, 1992) suggested that the *Drosophila* circadian rhythm results from a negative feedback exerted by the PER protein on the synthesis of the *per* mRNA. Post-translational modification of PER is also involved in the mechanism of circadian oscillations. Experimental evidence indeed indicates that PER is multiply phosphorylated (Edery et al. 1994). It appears that PER phosphorylation plays a role in the circadian oscillatory mechanism, by controlling the nuclear localization of PER and/or its degradation (Grima et al. 2002; Ko et al. 2002).

Overexpression of PER in *Drosophila* eyes represses *per* transcription and suppresses circadian rhythmicity in these cells, without affecting circadian oscillations in other *per*-expressing cells in the brain or the circadian rhythm in locomotor activity. This work shows that the action of PER on transcription is intracellular, and suggests that "each *per*-expressing cell contains an autonomous oscillator of which the *per* feedback loop is a component" (Zeng et al. 1994). Such a mechanism, based on negative autoregulation of transcription, has also been found in *Neurospora* (Aronson et al. 1994). The current view is that negative autoregulation of gene expression by a clock protein represents a unified mechanism for the generation of circadian rhythmicity in a wide variety of experimental systems (Dunlap 1999; Young and Kay 2001).

B. A core deterministic model for circadian oscillations of the PER protein and its mRNA

A first model for circadian oscillations in the *Drosophila* PER protein and its mRNA is based on multiple phosphorylation of PER and on the inhibition of *per* transcription by a phosphorylated form of the protein (Goldbeter 1995). This model, schematized in Figure 13.1a, can be viewed as a minimal core model because it takes into account a limited number of phosphorylated residues of PER. The model also applies to oscillations of FRQ and *frq* mRNA in *Neurospora*.

In the model, the *per* gene is first expressed in the nucleus and transcribed into *per* messenger RNA (mRNA). The latter is transported into the cytosol, where it is translated into the PER protein, P_0 , and degraded. The PER protein undergoes multiple phosphorylation, from P_0 into P_1 and from P_1 into P_2 . These modifications, catalyzed by a protein kinase, are reverted by a phosphatase. The fully phosphorylated form of the protein is marked up for degradation and transported into the nucleus in a reversible manner. The nuclear form of the protein (P_N) represses the transcription of the gene.

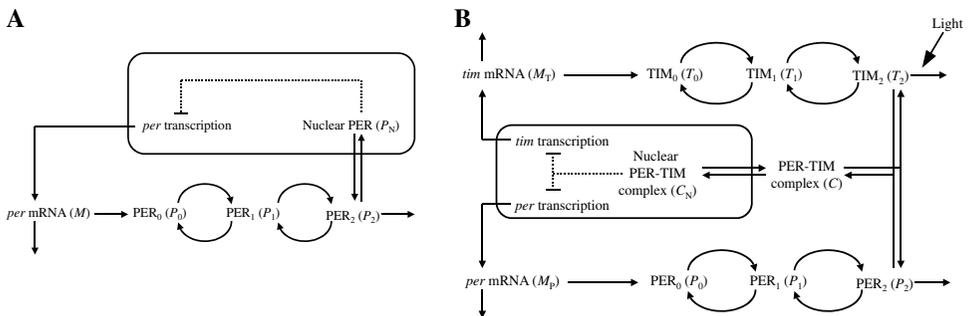


Figure 13.1. Schemes of the models for circadian oscillations in *Drosophila*. (a) The PER model is based on the sole negative regulation exerted by the PER protein on the expression of its gene (Goldbeter 1995). (b) The PER-TIM model incorporates the *tim* gene and its product, which forms a complex with the PER protein. This model is based on the negative regulation exerted by the PER-TIM complex on the expression of the *per* and *tim* genes. The effect of light is to increase the rate of TIM degradation (Leloup and Goldbeter 1998).

In the model, we consider two successive phosphorylations of PER, which is the minimal implementation of multiple phosphorylation. A single phosphorylation step would yield similar results. In fact, sustained oscillations can occur in the absence of phosphorylation, as shown by the study of a three-variable model representing an even simpler model for circadian oscillations (Leloup et al. 1999; Gonze and Goldbeter 2000; Gonze et al. 2000). We nevertheless focus on a model that includes multiple phosphorylation, because this process contributes to the mechanism of circadian oscillations by introducing a delay in the negative feedback loop.

In the model, the temporal variation of the concentrations of mRNA (M) and of the various forms of the regulatory protein—cytosolic (P_0 , P_1 , P_2) or nuclear (P_N)—is governed by the following system of kinetic equations (see Goldbeter (1995, 1996) for further details):

$$\begin{aligned}
 \frac{dM}{dt} &= v_s \frac{K_I^n}{K_I^n + P_N^n} - v_m \frac{M}{K_m + M} \\
 \frac{dP_0}{dt} &= k_s M - v_1 \frac{P_0}{K_1 + P_0} + v_2 \frac{P_1}{K_2 + P_1} \\
 \frac{dP_1}{dt} &= v_1 \frac{P_0}{K_1 + P_0} - v_2 \frac{P_1}{K_2 + P_1} - v_3 \frac{P_1}{K_3 + P_1} + v_4 \frac{P_2}{K_4 + P_2} \\
 \frac{dP_2}{dt} &= v_3 \frac{P_1}{K_3 + P_1} - v_4 \frac{P_2}{K_4 + P_2} - v_d \frac{P_2}{K_d + P_2} - k_1 P_2 + k_2 P_N \\
 \frac{dP_N}{dt} &= k_1 P_2 - k_2 P_N
 \end{aligned} \tag{13.1}$$

In these equations, the phosphorylation and dephosphorylation terms (with maximum rates v_1 , v_3 , and v_2 , v_4 , respectively)—as well as the degradation terms for

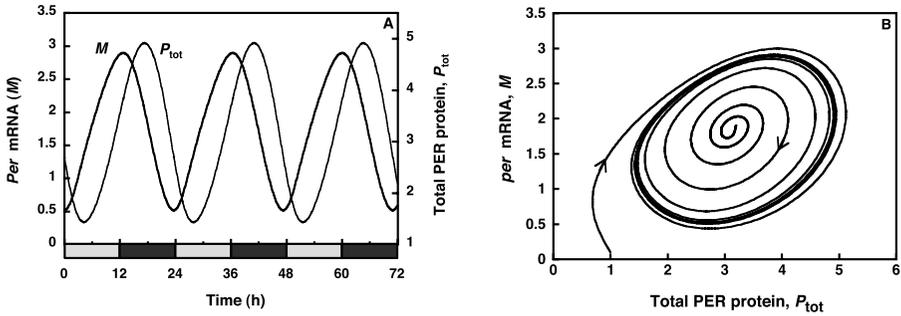


Figure 13.2. Sustained oscillations and limit cycle generated by the PER model. (a) Temporal variation in *per* mRNA (M) and in the total amount of PER protein (P_{tot}). (b) Sustained oscillations in total PER protein and *per* mRNA (expressed in nM) correspond to the evolution toward a limit cycle when the system's trajectory is projected onto the (M, P_{tot}) plane. Starting from two different initial conditions, the system reaches a unique closed curve characterized by a period and amplitude that are fixed for the given set of parameter values. The curves have been obtained by numerical integration of Equations 13.1. Parameter values are $v_s = 0.76$ nM/h, $v_m = 0.65$ nM/h, $k_s = 0.38$ h⁻¹, $v_d = 0.95$ nM/h, $k_1 = 1.9$ h⁻¹, $k_2 = 1.3$ h⁻¹, $K_1 = 1$ nM, $K_d = 0.2$ nM, $K_1 = K_2 = K_3 = K_4 = 2$ nM, $n = 4$, $V_1 = 3.2$ nM/h, $V_2 = 1.58$ nM/h, $V_3 = 5$ nM/h, and $V_4 = 2.5$ nM/h. Initial conditions are $M = 0.1$, $P_0 = P_1 = P_2 = P_N = 0.25$ ($P_{\text{tot}} = 1$), $M = 1.9$, and $P_0 = P_1 = P_2 = P_N = 0.8$ ($P_{\text{tot}} = 3.2$) (see Goldbeter (1995, 1996)).

mRNA and fully phosphorylated PER protein (with maximum rates v_m and v_d , respectively)—are all of Michaelian form corresponding to non-cooperative enzyme kinetics. The repression term takes the form of a Hill equation characterized by the Hill coefficient n . Repression by P_N becomes steeper and steeper as the degree of cooperativity n increases above unity. Although higher cooperativity favors the occurrence of sustained oscillations, periodic behavior can also be obtained for $n = 1$ (i.e., in the absence of cooperativity in repression).

For an appropriate set of parameter values, the model accounts for the occurrence of sustained oscillations in continuous darkness (Figure 13.2a). When plotting the time evolution of one variable (e.g., *per* mRNA (M)) as a function of another variable (e.g., the total amount of PER protein (P_{tot})), these oscillations correspond in such a phase plane to the evolution toward a closed curve, known as a limit cycle (Figure 13.2b). This name stems from the fact that the same closed trajectory is reached regardless of initial conditions, as illustrated in Figure 13.2b. In addition to accounting for the circadian rhythms in mRNA and for protein level, the model shows how variations in parameters such as the rate of degradation of PER or the rate of its translocation into the nucleus may change the period of the oscillations, or even suppress rhythmic behavior (Goldbeter 1995, 1996).

When the model based on PER alone was proposed, the way light affects circadian rhythms in *Drosophila* was still unknown. In 1996, a series of papers showed, concomitantly, that a second protein—TIM (for TIMELESS)—forms a complex with PER, and that light acts by inducing degradation of TIM (Hunter-Ensor et al. 1996; Lee et al. 1996; Myers et al. 1996; Zeng et al. 1996). These observations paved the

way for the construction of a more detailed computational model incorporating the formation of a PER-TIM complex as well as the enhancement of TIM degradation during the light phase.

C. A ten-variable deterministic model for circadian oscillations in *Drosophila*

The ten-variable model for circadian oscillations of the PER and TIM proteins and of *per* and *tim* mRNAs in *Drosophila* (Leloup and Goldbeter 1998; Leloup et al. 1999) is schematized in Figure 13.1b. The mechanism is based on the negative feedback exerted by the complex between the nuclear PER and TIM proteins on the expression of their genes. For each of these proteins, transcription, translation, and multiple phosphorylation are treated as in the PER model of Figure 13.1a. The fully phosphorylated proteins PER and TIM are marked up for degradation, and form a complex that is transported into the nucleus in a reversible manner. The nuclear form of the PER-TIM complex represses the transcription of the *per* and *tim* genes.

Recent experiments indicate that repression is in fact of indirect nature: a complex between two activators, the CLOCK and CYC proteins, promotes the expression of the *per* and *tim* genes. The PER-TIM complex prevents this activation by forming a complex with CLOCK and CYC (Darlington et al. 1998; Rutila et al. 1998; Lee et al. 1999). We return to the effect of such an indirect negative feedback in Section IV, restricting the present discussion to the PER-TIM model. In this model, the variables are the concentrations of the mRNAs (M_P and M_T), the various forms of the PER and TIM proteins ($P_0, P_1, P_2, T_0, T_1, T_2$), and the cytosolic (C) and nuclear (C_N) forms of the PER-TIM complex. The temporal evolution of the concentration variables is governed by the following system of 10 kinetic equations (see Leloup and Goldbeter (1998) and Leloup et al. (1999) for further details):

$$\begin{aligned}
 \frac{dM_P}{dt} &= v_{sP} \frac{K_{IP}^n}{K_{IP}^n + C_N^n} - v_{mP} \frac{M_P}{K_{mP} + M_P} - k_d M_P \\
 \frac{dP_0}{dt} &= k_{sP} M_P - V_{1P} \frac{P_0}{K_P + P_0} + V_{2P} \frac{P_1}{K_{2P} + P_1} - k_d P_0 \\
 \frac{dP_1}{dt} &= V_{1P} \frac{P_1}{K_P + P_0} - V_{2P} \frac{P_1}{K_{2P} + P_1} - V_{3P} \frac{P_1}{K_{3P} + P_1} + V_{4P} \frac{P_2}{K_{4P} + P_2} - k_d P_1 \\
 \frac{dP_2}{dt} &= V_{3P} \frac{P_1}{K_{3P} + P_1} - V_{4P} \frac{P_2}{K_{4P} + P_2} - k_3 P_2 T_2 + k_4 C - v_{dP} \frac{P_2}{K_{dP} + P_2} - k_d P_2 \\
 \frac{dM_T}{dt} &= v_{sT} \frac{K_{IT}^n}{K_{IT}^n + C_N^n} - v_{mT} \frac{M_T}{K_{mT} + M_T} - k_d M_T \\
 \frac{dT_0}{dt} &= k_{sT} M_T - V_{1T} \frac{T_0}{K_{1T} + T_0} + V_{2T} \frac{T_1}{K_{2T} + T_1} - k_d T_0 \\
 \frac{dT_1}{dt} &= V_{1T} \frac{T_0}{K_{1T} + T_0} - V_{2T} \frac{T_1}{K_{2T} + T_1} - V_{3T} \frac{T_1}{K_{3T} + T_1} + V_{4T} \frac{T_2}{K_{4T} + T_2} - k_d T_1
 \end{aligned} \tag{13.2}$$

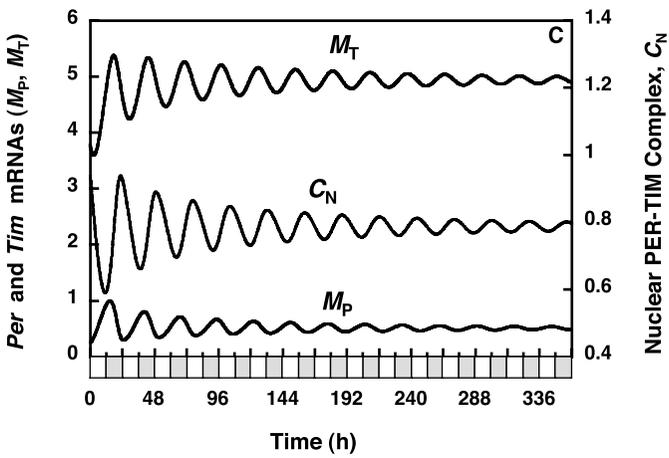
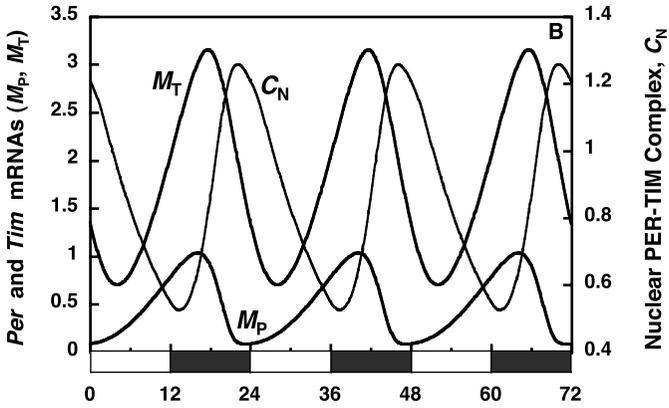
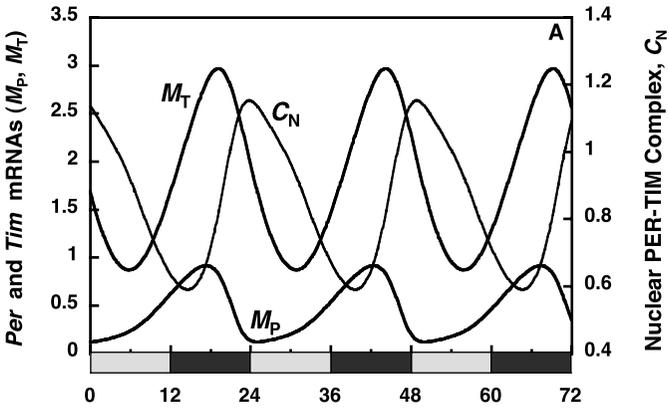
$$\begin{aligned}\frac{dT_2}{dt} &= V_{3T} \frac{T_1}{K_{3T} + T_1} - V_{4T} \frac{T_2}{K_{4T} + T_2} - k_3 P_2 T_2 + k_4 C - v_{dT} \frac{T_2}{K_{dT} + T_2} - k_d T_2 \\ \frac{dC}{dt} &= k_3 P_2 T_2 - k_4 C - k_1 C + k_2 C_N - k_{dC} C \\ \frac{dC_N}{dt} &= k_1 C - k_2 C_N - k_{dN} C_N\end{aligned}$$

These equations correspond to one particular version in a family of possible models, which differ by details of the molecular implementation of the feedback mechanism. Thus, rather than considering the formation of a complex between the fully phosphorylated forms of PER and TIM the complex could be made also (or instead) between the non-phosphorylated or mono-phosphorylated forms of the proteins. These other versions of the basal model yield largely similar results.

The various terms appearing in Equations 13.2 are similar to those of Equations 13.1. We have added nonspecific degradation terms, characterized by the rate constants k_d , k_{dC} , and k_{dN} . These linear terms are generally of negligible magnitude, and are not essential for oscillations. Their inclusion ensures the existence of a steady state when the specific protein degradation processes are inhibited. In Equations 13.2, parameter v_{dT} represents the maximum value of the TIM degradation rate. This is the light-sensitive parameter, which will be set to a constant low value during continuous darkness, and to a constant high value during continuous light. In a light/dark cycle, v_{dT} will vary in a square-wave manner between these two extreme values. The square-wave corresponds well to laboratory conditions under which light varies in an all-or-none manner. The natural variation of light is of course smoother, and other waveforms should be considered to address the effect of variations of luminosity under natural light/dark cycles.

Much as the PER model, the model based on the formation of the PER-TIM complex can account for sustained autonomous oscillations originating from negative auto-regulatory feedback. Now, however, we may address the dynamic behavior of the model in various lighting conditions, by incorporating suitable changes in parameter v_{dT} . Thus, as illustrated in Figure 13.3, sustained oscillations can occur

Figure 13.3. Circadian oscillations in the PER-TIM model. From top to bottom, the curves correspond to (a) sustained oscillations in continuous darkness, (b) entrainment by a light/dark cycle of 24 h period (12 : 12 LD), and (c) damped oscillations in continuous light. The LD cycle is symbolized by the alternation of white and black bars. Continuous darkness is symbolized by the alternation of gray and black bars. Shown is the temporal variation in *per* and *tim* mRNAs (M_P , M_T) and in the concentration of nuclear PER-TIM complex (C_N). The curves have been obtained by numerical integration of Equations 13.2 (Leloup and Goldbeter 1998). Parameter values are $v_{SP} = 0.8 \text{ nM h}^{-1}$, $v_{ST} = 1 \text{ nM h}^{-1}$, $v_{MP} = 0.8 \text{ nM h}^{-1}$, $v_{MT} = 0.7 \text{ nM h}^{-1}$, $K_{mP} = K_{mT} = 0.2 \text{ nM}$, $k_{SP} = k_{ST} = 0.9 \text{ h}^{-1}$, $v_{dP} = v_{dT} = 2 \text{ nM h}^{-1}$, $k_1 = 1.2 \text{ h}^{-1}$, $k_2 = 0.2 \text{ h}^{-1}$, $k_3 = 1.2 \text{ nM}^{-1} \text{ h}^{-1}$, $k_4 = 0.6 \text{ h}^{-1}$, $K_{IP} = K_{IT} = 1 \text{ nM}$, $K_{dP} = K_{dT} = 0.2 \text{ nM}$, $n = 4$, $K_{1P} = K_{1T} = K_{2P} = K_{2T} = K_{3P} = K_{3T} = K_{4P} = K_{4T} = 2 \text{ nM}$, $k_d = k_{dC} = k_{dN} = 0.01 \text{ h}^{-1}$, $V_{1P} = V_{1T} = 8 \text{ nM h}^{-1}$, $V_{2P} = V_{2T} = 1 \text{ nM h}^{-1}$, $V_{3P} = V_{3T} = 8 \text{ nM h}^{-1}$, and $V_{4P} = V_{4T} = 1 \text{ nM h}^{-1}$. Parameter v_{dT} is increased from 2 nM/h in the dark phase to 5 nM/h in the light phase (Leloup and Goldbeter 1998).



in continuous darkness (DD), but damped oscillations occur in conditions corresponding to continuous light (LL), as observed in *Drosophila* (Qiu and Hardin 1996). In LL, the light-sensitive parameter was chosen so that it takes a high value corresponding to a stable steady state. The disappearance of oscillations can be explained intuitively: because of enhanced degradation, the TIM protein cannot reach a level allowing effective repression by the PER-TIM complex. Oscillations observed in DD with a period close to 24 h can be entrained by a 12 : 12 LD cycle (12 h of light followed by 12 h of darkness). Experimentally, there exists a window of entrainment, ranging typically from 21 to 28 h (Moore-Ede et al. 1982).

The PER-TIM model allows us to compare theoretical predictions with experimental observations in a variety of cases. A first comparison pertains to entrainment by LD cycles of varying photoperiod. As shown by the experiments of Qiu and Hardin (1996), the peak in *per* mRNA always follows the transition from the L to the D phase by about 4 h. A similar result is obtained in the PER-TIM model (Figure 13.4). The lag after the L to D transition appears to be the same regardless of the duration of the light phase, because the level of TIM has decreased to a minimum value at the end of the L phase, and the time required for the PER-TIM complex to accumulate during the dark phase above the threshold for repression remains unchanged.

Another key comparison pertains to the phase shifts induced by light pulses in continuous darkness. Depending on the phase at which these perturbations are made, circadian oscillations can be either advanced or delayed. Alternatively, no phase shift may occur. These data yield a phase response curve (PRC) when the phase shift is plotted as a function of the phase of perturbation. The PRC is an

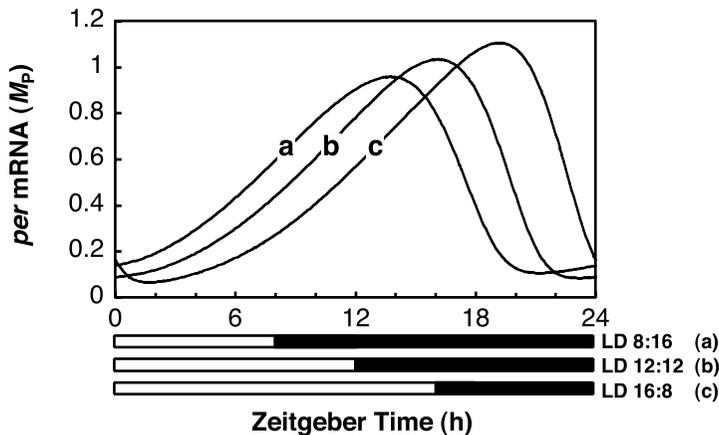


Figure 13.4. Phase locking of the *per* mRNA oscillations in the PER-TIM model. The three curves correspond to entrainment by a light/dark cycle of 24 h period but with different photoperiod: (a) 8 : 16 LD cycle, (b) 12 : 12 LD cycle, and (c) 16 : 8 LD cycle. The LD cycles are symbolized by the alternation of white and black bars. The curves have been obtained by numerical integration of Equations 13.2. Parameter values are as in Figure 13.3.

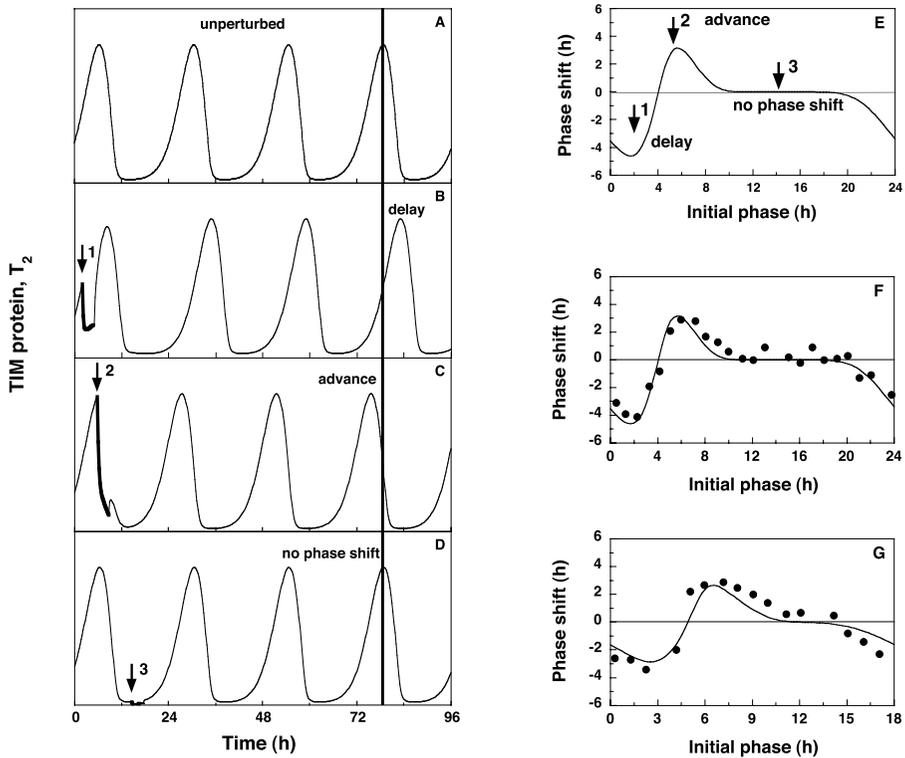


Figure 13.5. Phase shifting by a light pulse: comparison with experiments. (a) Unperturbed oscillations of phosphorylated TIM (T_2). The vertical line through the fourth peak serves as reference for determining phases shifts. (b–d) Transient perturbations at three different phases of the oscillations, producing, respectively, a phase delay, a phase advance, or an absence of phase shift. The arrows mark the beginning of the light pulse and the thick lines indicate both the duration and the effect of this perturbation (see following). (e) Phase response curve (PRC) obtained by plotting the phase shift as a function of the phase at which the perturbation is applied. The perturbation takes the form of a 3-h twofold increase in TIM maximum degradation rate (v_{dT}), triggered by the light pulse. (f and g) PRCs obtained theoretically (solid lines) for the wild type (panel F) and for the *per^S* mutant (panel G) in *Drosophila*. The theoretical predictions compare well with the experimental observations (dots) based on data obtained by Konopka and Orr using a 1-min light pulse (redrawn from Figure 2 of Hall and Rosbash (1987)). The oscillations of the TIM protein (panels A through D) and the PRCs (panels E through G) have been obtained by numerical integration of Equations 13.2 (Leloup and Goldbeter 1998). Parameter values are listed in Figure 2 of Leloup and Goldbeter (1998). For the PRCs, the zero phase is chosen, as in the experiments (Hall and Rosbash 1987), so that the minimum in *per* mRNA occurs after 12 h.

important tool in the study of circadian rhythms. We may simulate the effect of light pulses in the PER-TIM model by transiently increasing the maximum rate of TIM degradation, v_{dT} . Unperturbed oscillations of fully phosphorylated TIM (T_2) are shown in Figure 13.5a, where the vertical line through the fourth peak will serve as reference for determining phase shifts triggered by transient perturbations.

As shown in Figure 13.5b, when the perturbation is applied during the rising phase of TIM a phase delay is observed. In contrast, a phase advance occurs when the perturbation is made at the maximum of TIM (Figure 13.5c), whereas no phase shift is observed when the pulse is given at the minimum of TIM (Figure 13.5d). The latter result stems from the fact that when TIM is already at its minimum a transient increase in TIM degradation remains without effect. Plotting the phase shifts as a function of the phase of perturbation yields the PRC shown in Figure 13.5e, where the arrows 1 through 3 refer to the situations depicted in panels B through D, respectively. The predictions of the model compare well with the experimental PRC both for wild-type flies (Figure 13.5f, where the solid curve is the same PRC as in panel E) and for the *per^s* mutant (Figure 13.5g). The model indicates that the dead zone in which no phase shift occurs is nearly absent in the *per^s* mutant because TIM remains near its minimum for a relatively much shorter time, as a result of the faster degradation of PER in this mutant (see Figure 6 in Leloup and Goldbeter (1998)).

Obtaining good agreement with experimental observations is not straightforward, as this requires an appropriate characterization of the biochemical effects of a light pulse on the circadian clock. In constructing the theoretical PRC of Figure 13.5, we assumed that the effect of the light pulse is to double during 3 h the maximum rate of TIM degradation. Other combinations of multiplication factor and duration of increase may also yield satisfactory agreement. The interest of this result is to predict that the light pulse should have long-lasting biochemical consequences that may outlast the light pulse itself. This prediction is in fact corroborated by recent experimental observations (Busza et al. 2004).

Other results obtained with the PER-TIM model are of a more counter-intuitive nature. First, the model shows that in a certain range of parameter values sustained oscillations of the limit cycle type may coexist with a stable steady state. Such a situation, known as hard excitation, provides a possible explanation for the suppression of circadian rhythms by a single light pulse and for the subsequent restoration of periodic behavior by a second such pulse. This puzzling phenomenon, which has been observed in a variety of organisms, remains largely unexplained. The model indicates that over a range of phases corresponding to TIM increase in *Drosophila* transient increases in parameter v_{dT} may bring the system from the limit cycle into the basin of attraction of the stable steady state. A second pulse in v_{dT} may then bring back the oscillations (Figure 13.6a). Suppression is only possible over a finite portion of the limit cycle, as shown in Figure 13.6b. The characteristics (duration and amplitude) of the suppressing pulse change with the phase of perturbation in this domain (Leloup and Goldbeter 2001). In contrast, a single critical perturbation suppressing the rhythm exists in the situation described by Winfree (1980), wherein the stable limit cycle surrounds an unstable steady state. However, suppression is only transient in that case. The coexistence between a stable steady state and a stable limit cycle (illustrated in Figure 13.6a) is by no means uncommon, but a computational model is clearly needed to predict the occurrence of such a phenomenon.

We were at first surprised to observe that the deterministic PER-TIM model was also capable of producing chaotic behavior in constant environmental conditions

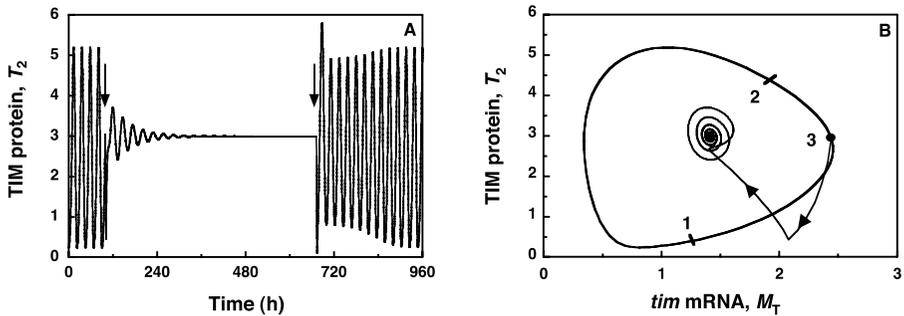


Figure 13.6. Long-term suppression of circadian rhythms by a single pulse of light. (a) Permanent rhythm suppression by a single pulse of light in the PER-TIM model, and restoration of the rhythm by a similar pulse. At the time indicated by the first arrow, to mimic the effect of a light pulse parameter v_{dT} , which measures the maximum rate of TIM degradation, is increased during 2 h from the basal value of 1.3 nM h^{-1} up to 4.0 nM h^{-1} . Initial conditions correspond to point 3 in panel B. At the time indicated by a second arrow, a similar change in v_{dT} , mimicking a second light pulse, is initiated, and the system returns to the oscillatory regime. The curve is obtained by numerical integration of Equations 13.2 for the parameter values of Figure 4 in Leloup and Goldbeter (2001). (b) Light pulses, translated into transient increases in v_{dT} , can permanently suppress the rhythm when applied over a portion of the limit cycle bounded by the two black bars marked 1 and 2. The trajectory starting from point 3 on the limit cycle corresponds to the rhythm suppression by the first pulse in a.

(e.g., continuous darkness (Leloup and Goldbeter 1999)). Such autonomous chaos has previously been shown to originate from the interplay between two instability-generating mechanisms (e.g., two feedback loops, each of which may produce sustained oscillations (Goldbeter 1996)). Here, the model contains but a single negative feedback loop, exerted by the PER-TIM complex. However, the formation of this complex involves two branches leading to the synthesis of PER and TIM. Chaos occurs in a relatively small parameter domain when a dynamical imbalance arises between the synthesis and degradation of the PER and TIM proteins or their mRNAs. Nonautonomous chaos can also be found in models for circadian rhythms, as a result of the periodic forcing of the circadian clock by light/dark cycles. The theoretical study indicates (Gonze and Goldbeter 2000) that the occurrence of such nonautonomous chaos is favored by the square wave nature of LD cycles: the domain of entrainment indeed enlarges at the expense of the domain of chaos when the waveform of the LD cycle progressively changes from square wave to sinusoidal.

Another use of the PER and PER-TIM models for circadian oscillations in *Drosophila* is to address the dynamical bases of temperature compensation (i.e., the relative independence of the period of circadian oscillations with respect to temperature (see Section II.A)). The analysis of the models supports the view (Ruoff and Rensing 1996) that temperature compensation originates from a balance between two opposing tendencies: the acceleration of some reactions with temperature tends to increase the period, whereas the acceleration of other reactions tends to lower it (Leloup and Goldbeter 1997). When the balance is lost (as a

result of a mutation), temperature compensation fails to occur, as observed in long- and short-period *Drosophila* mutants.

This discussion shows how useful deterministic models of moderate complexity may prove for the study of circadian rhythms. However, the question arises as to the validity of these computational models when the numbers of molecules involved in the oscillatory mechanism are small, as may occur for proteins and mRNAs in cellular conditions. Then, deterministic models may reach their limits, and it becomes necessary to resort to stochastic approaches. We shall now examine how stochastic models may account for the emergence of circadian rhythms, and will turn thereafter to more complex deterministic models proposed for the mammalian circadian clock.

III. STOCHASTIC MODELS FOR CIRCADIAN RHYTHMS

A. Core molecular model for circadian oscillations

To illustrate the stochastic approach to modeling circadian rhythms, it will be useful to resort to a relatively simple model for circadian oscillations. The model examined in Section II.A and schematized in Figure 13.1a provides a core model for circadian rhythms based on the negative feedback exerted by a protein (which is referred to in the following as clock protein) on the expression of its gene. As previously indicated, this model applies to circadian oscillations of the PER protein and *per* mRNA in *Drosophila*, and to the case of *Neurospora* (Leloup et al. 1999, Gonze et al. 2000) for which circadian rhythms originate from the negative feedback exerted by the FRQ protein on the expression of its gene (Aronson et al. 1994; Lee et al. 2000). The core model contains five variables and is described by Equations 13.1. When the effect of light is incorporated—as was done for the PER-TIM model discussed in Section II.B—this model accounts for the occurrence of sustained oscillations in continuous darkness, phase-shifting by light pulses, and entrainment by light/dark cycles. The model shown in Figure 13.1a will thus serve as a convenient core model for testing the effect of molecular noise on circadian oscillations. An even simpler model (governed by a set of three kinetic equations) is obtained when disregarding multiple phosphorylation of the clock protein (Leloup et al. 1999; Gonze et al. 2000). The following discussion pertains to the five-variable model, which includes PER reversible phosphorylation.

B. Molecular noise in the fully developed stochastic version of the core model

The decrease in the total number (N) of molecules in a system of chemical reactions is accompanied by a rise in the amplitude of fluctuations around the state predicted by the deterministic evolution of this chemical system. These fluctuations, which reflect intrinsic molecular noise, can be taken into account by describing the chemical reaction system as a birth-and-death stochastic process governed by a

master equation (Nicolis and Prigogine 1977). In a given reaction step, molecules of participating species are either produced (birth) or consumed (death). At each step is associated a transition probability proportional to the numbers of molecules of involved chemical species and to the chemical rate constant of the corresponding deterministic model.

To implement such a master equation approach to stochastic chemical dynamics, Gillespie (1976, 1977) introduced a rigorous numerical algorithm. In addition to other approaches (Morton-Firth and Bray 1998), this method of the Monte Carlo type is widely used to determine the effect of molecular noise on the dynamics of chemical (Baras et al. 1990; Baras 1997), biochemical (McAdams and Arkin 1997), or genetic (Arkin et al. 1998) systems. The Gillespie method associates a probability with each reaction. At each time step the algorithm stochastically determines the reaction that takes place according to its probability, as well as the time interval to the next reaction. The numbers of molecules of the different reacting species as well as the probabilities are updated at each time step. In this approach (Gillespie 1976, 1977), a parameter denoted Ω permits the modulation of the number of molecules present in the system.

To assess the effect of molecular noise on circadian oscillations, we have used the Gillespie method to perform stochastic simulations of the core deterministic model governed by Equations 13.1. When the degree of cooperativity of repression—given by the Hill coefficient n in Equations 13.1—is equal to 4, the core mechanism can be decomposed in 30 elementary steps, as indicated in Table 13.1. A probability of occurrence, proportional to the deterministic rate constant, is associated with each of these individual steps. This approach rests on the analysis of a fully developed stochastic version of the core model for circadian oscillations. In the following we will show that an alternative (more compact) approach—in which the nonlinear functions in Equations 13.1 are not decomposed into elementary steps—yields largely similar results.

C. Robustness of circadian oscillations with respect to molecular noise

The first result obtained with the fully developed stochastic version of the core model for circadian rhythms is that it is also capable of producing sustained oscillations in conditions of continuous darkness. These oscillations correspond to the evolution toward a limit cycle, which is shown in the right-hand panels of Figure 13.7b as a projection onto the (M, P_N) plane. For comparison, the deterministic oscillations and the corresponding limit cycle are shown in Figure 13.7a. The effect of molecular noise is merely to induce variability in the maxima of the oscillations. This is reflected by the noisy appearance of the limit cycle and a thickening of its upper portion linking the maximum in mRNA with the maximum in nuclear (or total) clock protein. The noisy stochastic limit cycle surrounds the deterministic limit cycle (shown as the closed white curve in the lower right-hand panel in Figure 13.7b) obtained by numerical integration of Equations 13.1 in corresponding conditions (Gonze et al. 2002a, 2002b).

Table 13.1. Decomposition of the deterministic model into elementary reaction steps.

Reaction Number	Reaction Step	Probability of Reaction
1	$G + P_N \xrightarrow{a_1} GP_N$	$w_1 = a_1 \times G \times P_N / \Omega$
2	$GP_N \xrightarrow{d_1} G + P_N$	$w_2 = d_1 \times GP_N$
3	$GP_N + P_N \xrightarrow{a_2} GP_{N2}$	$w_3 = a_2 \times GP_N \times P_N / \Omega$
4	$GP_{N2} \xrightarrow{d_2} GP_N + P_N$	$w_4 = d_2 \times GP_{N2}$
5	$GP_{N2} + P_N \xrightarrow{a_3} GP_{N3}$	$w_5 = a_3 \times GP_{N2} \times P_N / \Omega$
6	$GP_{N3} \xrightarrow{d_3} GP_{N2} + P_N$	$w_6 = d_3 \times GP_{N3}$
7	$GP_{N3} + P_N \xrightarrow{a_4} GP_{N4}$	$w_7 = a_4 \times GP_{N3} \times P_N / \Omega$
8	$GP_{N4} \xrightarrow{d_4} GP_{N3} + P_N$	$w_8 = d_4 \times GP_{N4}$
9	$[G, GP_N, GP_{N2}, GP_{N3}] \xrightarrow{V_s} M$	$w_9 = v_s \times (G + GP_N + GP_{N2} + GP_{N3})$
10	$M + E_m \xrightarrow{k_{m1}} C_m$	$w_{10} = k_{m1} \times M \times E_m / \Omega$
11	$C_m \xrightarrow{k_{m2}} M + E_m$	$w_{11} = k_{m2} \times C_m$
12	$C_m \xrightarrow{k_{m3}} E_m$	$w_{12} = k_{m3} \times C_m$
13	$M \xrightarrow{k_5} M + P_0$	$w_{13} = k_5 \times M$
14	$P_0 + E_1 \xrightarrow{k_{11}} C_1$	$w_{14} = k_{11} \times P_0 \times E_1 / \Omega$
15	$C_1 \xrightarrow{k_{12}} P_0 + E_1$	$w_{15} = k_{12} \times C_1$
16	$C_1 \xrightarrow{k_{13}} P_1 + E_1$	$w_{16} = k_{13} \times C_1$
17	$P_1 + E_2 \xrightarrow{k_{21}} C_2$	$w_{17} = k_{21} \times P_1 \times E_2 / \Omega$
18	$C_2 \xrightarrow{k_{22}} P_1 + E_2$	$w_{18} = k_{22} \times C_2$
19	$C_2 \xrightarrow{k_{23}} P_0 + E_2$	$w_{19} = k_{23} \times C_2$
20	$P_1 + E_3 \xrightarrow{k_{31}} C_3$	$w_{20} = k_{31} \times P_1 \times E_3 / \Omega$
21	$C_3 \xrightarrow{k_{32}} P_1 + E_3$	$w_{21} = k_{32} \times C_3$
22	$C_3 \xrightarrow{k_{33}} P_2 + E_3$	$w_{22} = k_{33} \times C_3$
23	$P_2 + E_4 \xrightarrow{k_{41}} C_4$	$w_{23} = k_{41} \times P_2 \times E_4 / \Omega$
24	$C_4 \xrightarrow{k_{42}} P_2 + E_4$	$w_{24} = k_{42} \times C_4$
25	$C_4 \xrightarrow{k_{43}} P_1 + E_4$	$w_{25} = k_{43} \times C_4$
26	$P_2 + E_d \xrightarrow{k_{d1}} C_d$	$w_{26} = k_{d1} \times P_2 \times E_d / \Omega$
27	$C_d \xrightarrow{k_{d2}} P_2 + E_d$	$w_{27} = k_{d2} \times C_d$
28	$C_d \xrightarrow{k_{d3}} E_d$	$w_{28} = k_{d3} \times C_d$
29	$P_2 \xrightarrow{k_1} P_N$	$w_{29} = k_1 \times P_2$
30	$P_N \xrightarrow{k_2} P_2$	$w_{30} = k_2 \times P_N$

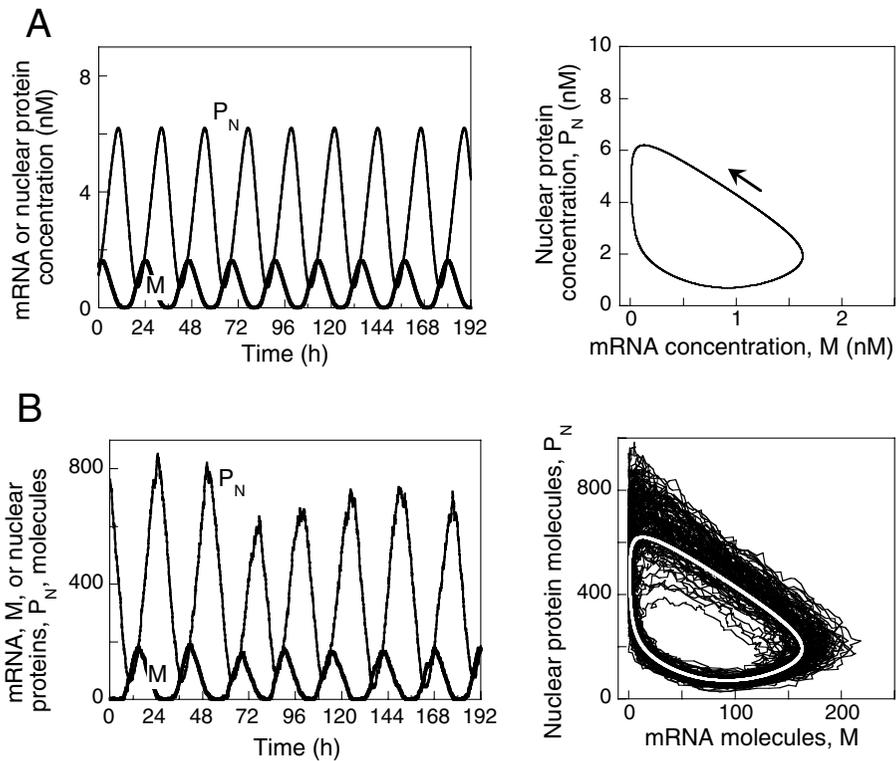


Figure 13.7. Deterministic versus stochastic simulations of the core model for circadian oscillations (schematized in Figure 13.1a). (a) Oscillations obtained in the absence of noise for the deterministic model governed by Equations 13.1. Sustained oscillations of mRNA (M) and nuclear clock protein (P_N) in the left-hand panel correspond to the evolution toward a limit cycle shown as a projection onto the (M, P_N) plane in the right-hand panel. (b) Oscillations generated by the stochastic version of the core model in the presence of noise, for $\Omega = 100$ and $n = 4$. The data, expressed in numbers of molecules of mRNA and of nuclear clock protein, are obtained by stochastic simulations of the detailed reaction system (Table 13.1) corresponding to the deterministic version of the core model. In the lower right-hand panel, the white curve corresponds to the deterministic limit cycle. The latter is surrounded by the stochastic trajectory which takes the form of a noisy limit cycle.

To assess the robustness of circadian oscillations at low numbers of molecules, we performed stochastic simulations for decreasing values of Ω . For $\Omega = 500$, the number of mRNA molecules varies in the range 0 to 1,000, whereas the numbers of nuclear and total clock protein molecules oscillate in the ranges 200 to 4,000 and 800 to 8,000, respectively (see left-hand panel in Figure 13.8a). The results in Figure 13.8 show that as Ω decreases progressively from the value of 500 down to a value of 100 or 50 robust circadian oscillations continue to occur in continuous darkness. The number of mRNA molecules oscillates from 0 to 200 or 0 to 120, whereas the number of nuclear clock protein molecules oscillates in the range 20 to 800 or 10

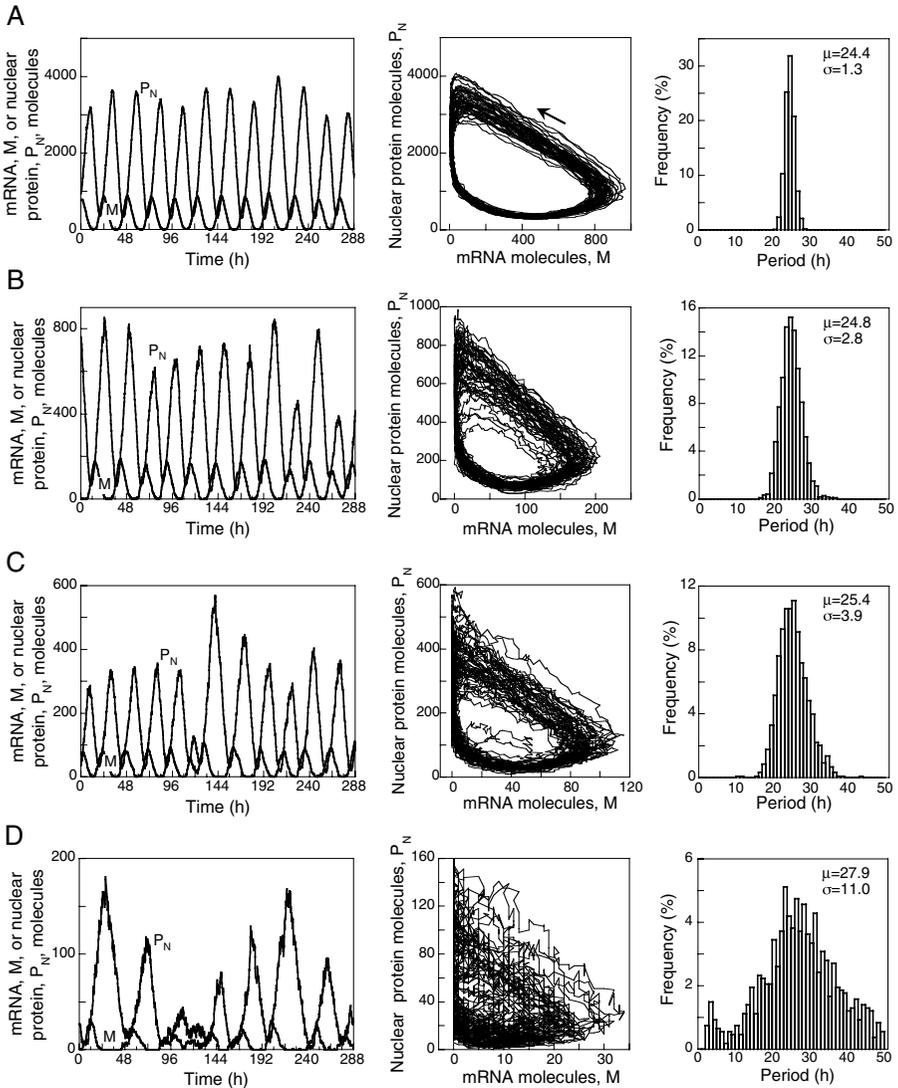


Figure 13.8. Effect of number of molecules on the robustness of circadian oscillations. Shown in rows A through D are the oscillations in the numbers of molecules of mRNA and nuclear clock protein, the projection of the corresponding limit cycle, and the histogram of periods of 1,200 successive cycles, for Ω varying from 500 (A), to 100 (B), 50 (C), and 10 (D). The curves are obtained by stochastic simulations of the core model (Table 13.1), for $n = 4$ (other parameters are listed in Table 13.2 where “mol” stands for “molecule”). For period histograms, the period was determined as the time interval separating two successive upward crossings of the mean level of mRNA or clock protein. In B and C, the decrease in the numbers of mRNA and protein molecules still permits robust circadian oscillations (see histograms where the mean value (μ) and standard deviation (σ) of the period are indicated in h), whereas at still lower numbers of molecules (D) noise begins to obliterate rhythmic behavior (Gonze et al. 2002b).

Table 13.2. Parameter values for stochastic simulations.

Reaction Steps	Parameter Values
Steps 1–8	For $n = 4$: $a_1 = \Omega \text{ mol}^{-1} \text{ h}^{-1}$, $d_1 = (160 \times \Omega) \text{ h}^{-1}$, $a_2 = (10 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_2 = (100 \times \Omega) \text{ h}^{-1}$, $a_3 = (100 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_3 = (10 \times \Omega) \text{ h}^{-1}$, $a_4 = (100 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_4 = (10 \times \Omega) \text{ h}^{-1}$ For $n = 3$: $a_1 = \Omega \text{ mol}^{-1} \text{ h}^{-1}$, $d_1 = (80 \times \Omega) \text{ h}^{-1}$, $a_2 = (100 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_2 = (100 \times \Omega) \text{ h}^{-1}$, $a_3 = (100 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_3 = \Omega \text{ h}^{-1}$ For $n = 2$: $a_1 = \Omega \text{ mol}^{-1} \text{ h}^{-1}$, $d_1 = (40 \times \Omega) \text{ h}^{-1}$, $a_2 = (100 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_2 = (10 \times \Omega) \text{ h}^{-1}$ For $n = 1$: $a_1 = (10 \times \Omega) \text{ mol}^{-1} \text{ h}^{-1}$, $d_1 = (20 \times \Omega) \text{ h}^{-1}$
Step 9	$v_3 = (0.5 \times \Omega) \text{ mol h}^{-1}$
Steps 10–12	$k_{m1} = 165 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{m2} = 30 \text{ h}^{-1}$, $k_{m3} = 3 \text{ h}^{-1}$, $E_{m \text{ tot}} = E_m + C_m = (0.1 \times \Omega) \text{ mol}$
Steps 13	$k_5 = 2.0 \text{ h}^{-1}$
Steps 14–16	$k_{11} = 146.6 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{12} = 200 \text{ h}^{-1}$, $k_{13} = 20 \text{ h}^{-1}$ $E_{1 \text{ tot}} = E_1 + C_1 = (0.3 \times \Omega) \text{ mol}$
Steps 17–19	$k_{21} = 82.5 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{22} = 150 \text{ h}^{-1}$, $k_{23} = 15 \text{ h}^{-1}$, $E_{2 \text{ tot}} = E_2 + C_2 = (0.2 \times \Omega) \text{ mol}$
Steps 20–22	$k_{31} = 146.6 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{32} = 200 \text{ h}^{-1}$, $k_{33} = 20 \text{ h}^{-1}$, $E_{3 \text{ tot}} = E_3 + C_3 = (0.3 \times \Omega) \text{ mol}$
Steps 23–25	$k_{41} = 82.5 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{42} = 150 \text{ h}^{-1}$, $k_{43} = 15 \text{ h}^{-1}$, $E_{4 \text{ tot}} = E_4 + C_4 = (0.2 \times \Omega) \text{ mol}$
Steps 26–28	$k_{d1} = 1650 \text{ mol}^{-1} \text{ h}^{-1}$, $k_{d2} = 150 \text{ h}^{-1}$, $k_{d3} = 15 \text{ h}^{-1}$, $E_{d \text{ tot}} = E_d + C_d = (0.1 \times \Omega) \text{ mol}$
Steps 29–30	$k_1 = 2.0 \text{ h}^{-1}$, $k_2 = 1.0 \text{ h}^{-1}$

to 600. For these smaller values of Ω , the limit cycles are more noisy but the period histograms calculated for some 1,200 successive cycles indicate that the distribution remains narrow with a mean free running period μ close to a circadian value. The standard deviation σ remains small with respect to the mean period but slightly increases as the number of molecules diminishes.

A further decrease in the number of molecules (e.g., down to $\Omega = 10$) will eventually obliterate circadian rhythmicity, and the latter is overcome by noise (Figure 13.8d). At such a low value of Ω , highly irregular oscillations occur, during which the number of mRNA molecules varies from 0 to 30 and the number of nuclear protein

molecules oscillates in the range 5 to 160. Even for such reduced numbers of mRNA and protein molecules, however, oscillations are not fully destroyed by noise. The histogram of periods indicates that the mean is still close to a circadian value, but the standard deviation is greatly increased. The stochastic approach illustrated in Figures 13.7 and 13.8 provides us with the unique opportunity of witnessing the emergence of a biological rhythm out of molecular noise (Gonze et al. 2004a, 2004b).

The results in Figure 13.8 were obtained in conditions in which the mean levels of mRNA and of clock protein differ by one to two orders of magnitude. Similar results are obtained by means of stochastic simulations when the level of mRNA is considerably lower than that of the clock protein, as long as the former remains above a few tens of molecules.

The degree of cooperativity is another parameter that affects the robustness of circadian oscillations in the presence of molecular noise. Stochastic simulations were performed with $\Omega = 100$ for values of n ranging from 1 to 4, where n denotes the total number of protein molecules that bind to the promoter to repress transcription. The results indicate that robustness significantly increases when n passes from 1 (absence of cooperativity) to values of 2 and above. Changes in standard deviation of the period show that cooperative repression enhances the robustness of circadian oscillations with respect to molecular noise (Gonze et al. 2002b).

Stochastic simulations further indicate that circadian oscillations can be entrained by LD cycles. The effect of light is incorporated into the model by assuming that the probability of occurrence of the reaction step corresponding to degradation of phosphorylated clock protein increases during the light phase, as observed in *Drosophila*. Of particular interest is that the phase of the entrained rhythm is then stabilized through periodic forcing by the LD cycle (Figure 13.9). The phase of the maximum in mRNA of clock protein is of course not constant in these conditions, because of fluctuations, but its mean value occurs a few hours after the L-to-D transition, as observed in the case of *Drosophila* (see also Figure 13.4 for the results obtained in the corresponding deterministic case).

Additional factors influence the robustness of circadian oscillations with respect to molecular noise. Among these are the distance from a bifurcation point, and the magnitude of the rate constants characterizing binding of the repressor to the gene. To illustrate the first aspect, it is useful to consider the bifurcation diagram showing the onset of sustained oscillations as a function of a control parameter such as the maximum rate of clock protein degradation, v_d (Figure 13.10). This diagram, obtained for the core deterministic model of Figure 13.1a governed by Equations 13.1, shows that as v_d is progressively increased from a low initial value the system at first settles in a stable non-oscillatory state before sustained oscillations of the limit cycle type arise when v_d exceeds a critical value. The amplitude of the oscillations progressively increases as the value of v_d moves away from this bifurcation point. We now select four increasing values of v_d located well below (a) or just below (b) the bifurcation value, and just above (c) or well beyond (d) it. Stochastic simulations performed for a given value of Ω with the fully developed

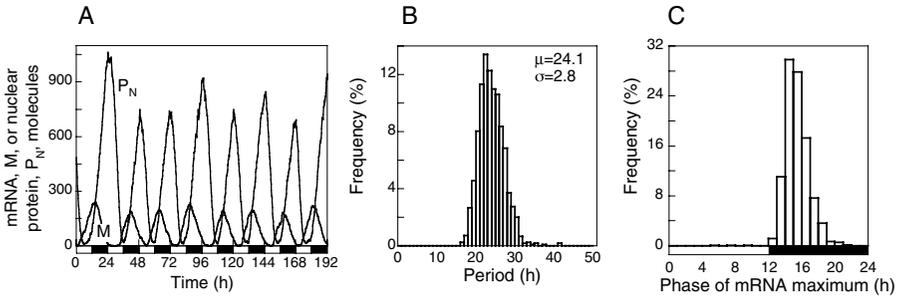


Figure 13.9. Effect of molecular noise on circadian oscillations under conditions of periodic forcing by a light/dark cycle. The data are obtained for $\Omega = 100$ and $n = 4$. (a) Circadian oscillations in the numbers of mRNA and nuclear clock protein molecules. (b) Histogram of periods with mean value (μ) and standard deviation (σ) indicated in h. (c) Histogram of the time corresponding to the maximum number of mRNA molecules over a period. Periodic forcing is achieved by doubling during each light phase the value ascribed during the dark phase to the parameter (k_{d3}) measuring the probability of the protein degradation step (Table 13.1). Histograms are determined for some 1,200 successive cycles (Gonze et al. 2002b).

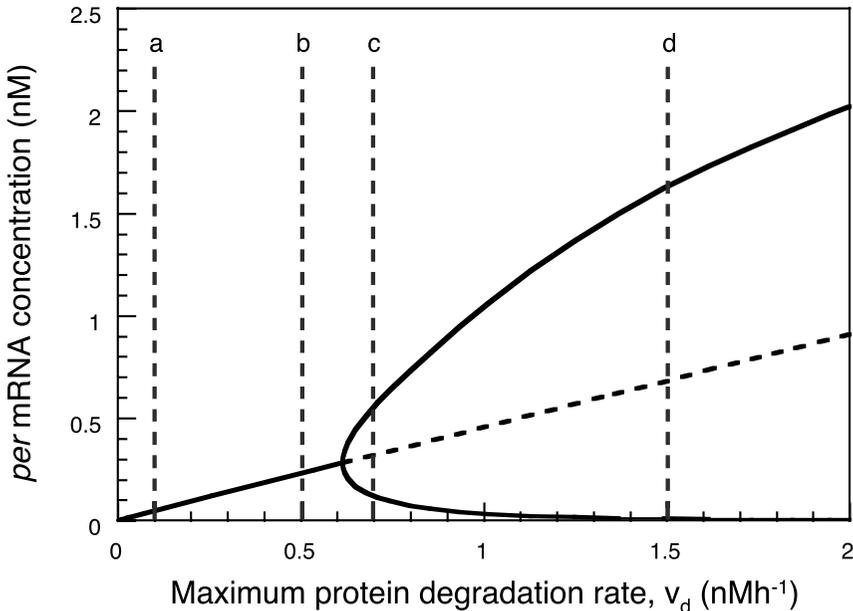


Figure 13.10. Bifurcation diagram showing the onset of sustained oscillations in the deterministic core model for circadian rhythms, as a function of parameter v_d (which measures the maximum rate of protein degradation). The curve shows the steady-state level of *per* mRNA, stable (solid line), or unstable (dashed line), as well as the maximum and minimum concentration of *per* mRNA in the course of sustained circadian oscillations. The diagram is established by means of the program AUTO (Doedel 1981) applied to Equations 13.1. Parameter values are given in Table 13.2 (Gonze et al. 2002a).

version of the core model indicate (Figure 13.11) that circadian oscillations become less sensitive to molecular noise as the system moves away from the bifurcation point, well into the domain of periodic behavior.

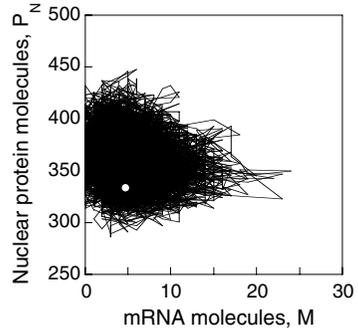
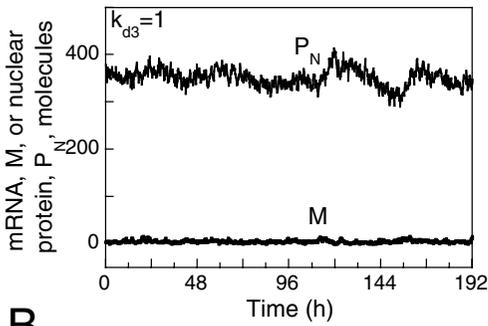
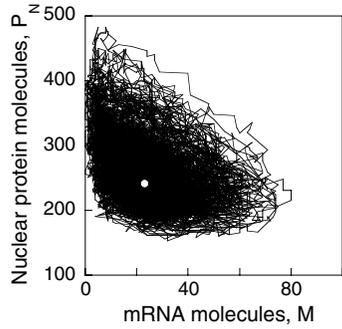
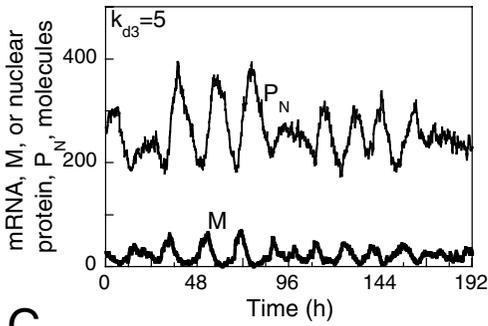
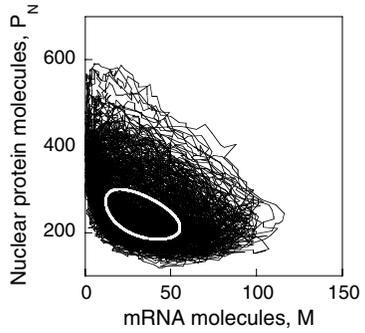
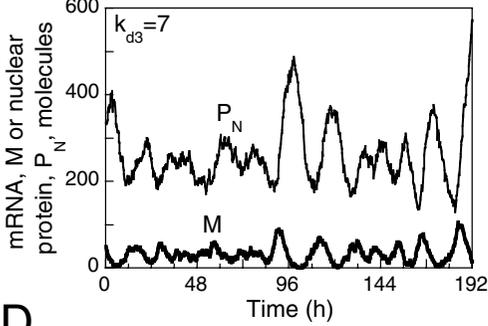
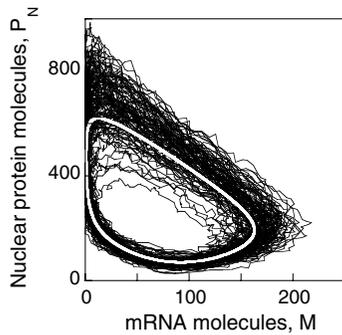
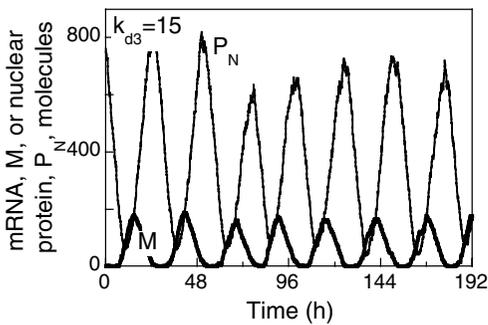
Finally, among the kinetic parameters that govern the probability of occurrence of the various individual steps listed in Table 13.1 few have as much influence on the robustness of circadian oscillations as the rate constants characterizing the successive binding of repressor molecules to the gene promoter of the clock protein. In the case of cooperative binding of four repressor molecules, we have to consider four successive steps of association and dissociation characterized by the rate constants a_i and d_i ($i = 1, \dots, 4$) (see steps 1 through 8 in Table 13.1). It will be useful to divide these rate constants by a scaling parameter γ to assess their influence on the robustness of circadian rhythms with respect to molecular noise. An increase in γ will thus correspond to a decrease in the rate constants a_i and d_i .

In Figure 13.12 are shown the results of stochastic simulations of the core model for $\gamma = 1$ (a), $\gamma = 100$ (b), and $\gamma = 1000$ (c). As γ increases up to 100 and 1,000, oscillations with larger and larger amplitude and increasing variability of the period are observed. The oscillations obtained for $\gamma = 1$ are much more regular. To clarify the nature of this phenomenon, we examined the deterministic version of the detailed stochastic model considered in Table 13.1. To the 30 reaction steps listed in Table 13.1 corresponds a deterministic system of 22 ordinary differential equations (Gonze et al. 2004a). In this fully developed version of the deterministic model, parameters a_i and d_i appear explicitly, whereas they only appear in the form of a single equilibrium inhibition constant (K_i) in the reduced five-variable deterministic model governed by Equations 13.1.

The results obtained with the fully developed deterministic model demonstrate the existence of a bifurcation as a function of the scaling parameter γ , as shown by the bifurcation diagram in Figure 13.13. When γ increases above a critical value close to 100, the system ceases to oscillate and evolves toward a stable steady



Figure 13.11. Effect of the proximity from a bifurcation point on the effect of molecular noise in the stochastic model for circadian rhythms. The different panels are established for the four increasing values of parameter k_{43} corresponding to the v_4 values shown in Figure 13.10: 0.1 (A), 0.5 (B), 0.7 (C) and 1.5 (d). The values of k_{43} listed in the left panels, are expressed here in molecules per h. The right-hand panels show the evolution in the phase plane, whereas the left-hand panels represent the corresponding temporal evolution of the number of *per* mRNA and nuclear PER molecules. (A) Fluctuations around a stable steady state. (B) Fluctuations around a stable steady state close to the bifurcation point. Damped oscillations occur in these conditions when the system is displaced from the stable steady state. In A and B, the white dot in the right-hand panel represents the stable steady state predicted by the deterministic version of the model in corresponding conditions. (C) Oscillations observed close to the bifurcation point. (D) Oscillations observed further from the bifurcation point, well into the domain of sustained oscillations. In C and D, the thick white curve in the right-hand panel represents the limit cycle predicted by the deterministic version of the model governed by Equations 13.1, in corresponding conditions. The smaller amplitude of the limit cycle in C as compared to the limit cycle in D is associated with an increased influence of molecular noise. The curves are obtained by means of the Gillespie algorithm applied to the model of Table 13.1 (Gonze et al. 2002a).

A**B****C****D**

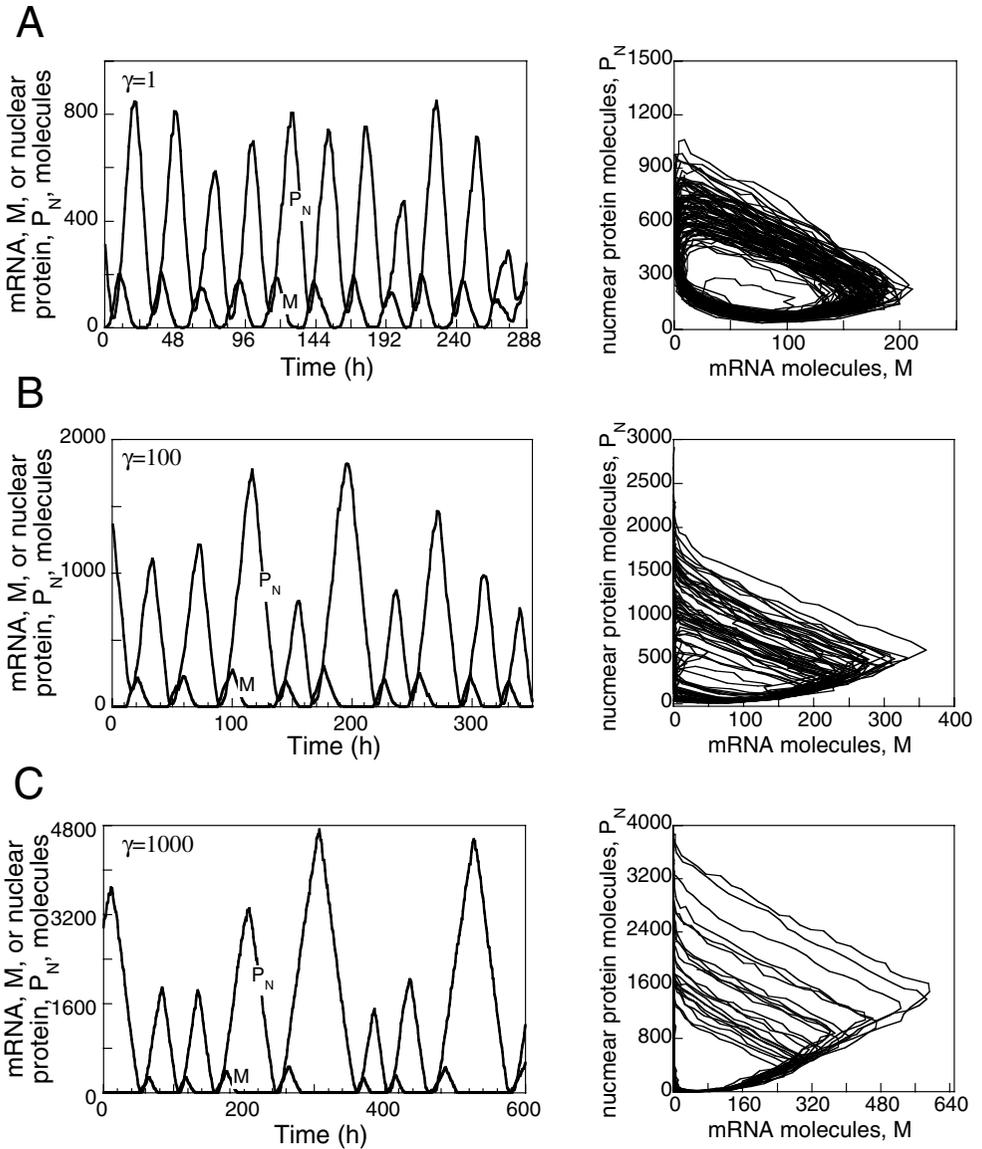


Figure 13.12. Irregular time series and trajectory in the phase space obtained by stochastic simulations of the core model for circadian rhythms for $\gamma = 1$ (a), $\gamma = 100$ (b), and $\gamma = 1,000$ (c). The curves were obtained for the model of Table 13.1, with $\Omega = 100$. Other parameter values are given in Table 13.2. The results should be compared with the bifurcation diagram established in Figure 13.13 as a function of γ for the corresponding fully developed version of the deterministic model. This diagram predicts that the steady state is stable and excitable for $\gamma = 100$ and 1,000, whereas sustained oscillations occur for $\gamma = 1$ when the steady state is unstable (Gonze et al. 2004a).

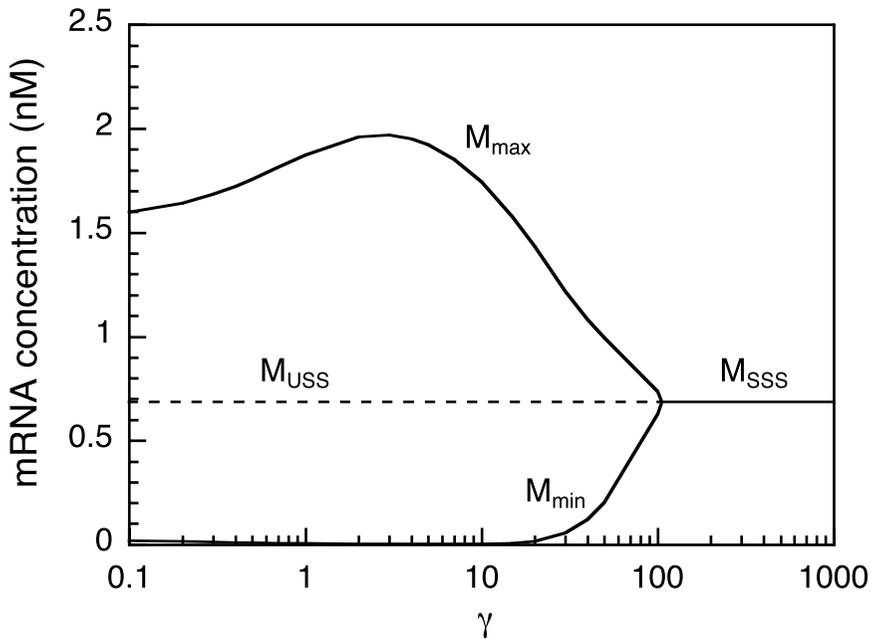


Figure 13.13. Bifurcation diagram showing the onset of circadian oscillations in the fully developed version of the deterministic core model, as a function of the scaling parameter γ . The latter parameter divides the association and dissociation rate constants a , and d , characterizing the binding of the repressor protein to the gene. The curve shows the steady-state level of mRNA, stable (solid line, M_{SSS}) or unstable (dashed line, M_{USS}), as well as the maximum (M_{max}) and minimum (M_{min}) mRNA concentration in the course of sustained oscillations. The diagram was determined by numerical integration of the 22 kinetic equations governing the dynamics of the fully developed deterministic model (Gonze et al. 2004a).

state. Numerical simulations performed with the 22-variable deterministic model for $\gamma = 1,000$, $\gamma = 100$, and $\gamma = 1$ show (Gonze et al. 2004a) that for $\gamma = 100$ the system still undergoes sustained low-amplitude oscillations. For $\gamma = 1,000$, the system evolves toward a stable steady state, as predicted by the bifurcation diagram of Figure 13.13, but this steady state is excitable: a small perturbation bringing the system slightly away from the steady state triggers a large excursion in the phase space, which corresponds to a burst of transcriptional activity, before the system returns to the stable steady state. This property of excitability also holds for the limit cycle observed for $\gamma = 100$. Thus, it is also possible to trigger large-amplitude peaks in gene transcription starting from such small-amplitude oscillations.

These results explain why oscillations predicted by stochastic simulations become highly irregular when the rate constants a_i and d_i decrease below a critical value. As shown by the study of the corresponding detailed deterministic model, such irregular oscillations reflect repetitive noise-induced large excursions away from a stable excitable steady state or from a small-amplitude limit cycle close to

the bifurcation point. The values of the bimolecular rate constants a_i used by Barkai and Leibler (2000) for simulating the circadian models of Figures 13.1a and 13.1b were probably below the critical value corresponding to sustained oscillations, which may explain their failure to obtain robust circadian oscillations in these models. When γ decreases (i.e., when the values of parameters a_i and d_i increase)—as in the case considered in Figure 13.8, which corresponds to $\gamma = 1$ —the oscillations become more regular and more robust, because the system operates well into the domain of sustained large-amplitude oscillations. The high values of parameters a_i and d_i corresponding to $\gamma = 1$ are of the order of those determined experimentally (Gonze et al. 2002b).

D. Non-developed stochastic models for circadian rhythms

The nonlinear terms appearing in the kinetic Equations 13.1 of the deterministic core model do not correspond to single reaction steps. These terms rather represent compact kinetic expressions obtained after application of quasi-steady-state hypotheses on enzyme-substrate or gene-repressor complexes. The resulting expressions are of the Michaelis—Menten type for enzyme reaction rates, or of the Hill type for cooperative binding of the repressor to the gene promoter. In the fully developed stochastic version of the core model, all reactions were decomposed into elementary steps (see Table 13.1).

Alternatively, we may resort to a simpler approach in which we attribute to each linear or nonlinear term of the kinetic equations a probability of occurrence of the corresponding reaction step (Gonze et al. 2002a). Then, in contrast to the treatment presented previously for the fully developed stochastic version we do not decompose the binding of the repressor P_N to the gene promoter into successive elementary steps, and rather retain the Hill function description for cooperative repression. A similar approach is taken for describing degradation of mRNA; translation of mRNA into protein, phosphorylation, or dephosphorylation reactions; and enzymatic degradation of fully phosphorylated clock protein and its reversible transport into and out of the nucleus. Some of these steps are of the Michaelian type, whereas others correspond to linear kinetics.

The comparison of stochastic simulations performed with the fully developed and non-developed versions of the core model showed that the two versions yield largely similar results (Gonze et al. 2002a). On the basis of these findings, a non-developed stochastic version of the 10-variable deterministic model governed by Equations 13.2, incorporating the formation of the PER-TIM complex, was considered. This version corresponds to a set of 30 reaction steps (listed in Table 13.3). Stochastic simulations show how sustained oscillations occur in this model under conditions corresponding to continuous darkness. As for the core model considered previously, the robustness of the oscillations is enhanced when the number of protein and mRNA molecules increases.

A conspicuous property of the 10-variable deterministic PER-TIM model for circadian rhythms in *Drosophila* is that it can produce autonomous chaotic behavior

Table 13.3. Nondeveloped stochastic version of the PER-TIM model for circadian rhythms [Gonze et al. 2003].

Reaction Number	Reaction Step	Probability of Reaction
1	$V_{sP} \rightarrow M_P$	$w_1 = (V_{sP} \times \Omega) \frac{(K_{IP} \times \Omega)^n}{(K_{IP} \times \Omega)^n + C_N^n}$
2	$M_P \xrightarrow{V_{mP}} \rightarrow$	$w_2 = (V_{mP} \times \Omega) \frac{M_P}{(K_{mP} \times \Omega) + M_P}$
3	$M_P \xrightarrow{k_{sP}} M_P + P_0$	$w_3 = k_{sP} \times M_P$
4	$P_0 \xrightarrow{V_{IP}} P_1$	$w_4 = (V_{IP} \times \Omega) \frac{P_0}{(K_{IP} \times \Omega) + P_0}$
5	$P_1 \xrightarrow{V_{2P}} P_0$	$w_5 = (V_{2P} \times \Omega) \frac{P_1}{(K_{2P} \times \Omega) + P_1}$
6	$P_1 \xrightarrow{V_{3P}} P_2$	$w_6 = (V_{3P} \times \Omega) \frac{P_1}{(K_{3P} \times \Omega) + P_1}$
7	$P_2 \xrightarrow{V_{4P}} P_1$	$w_7 = (V_{4P} \times \Omega) \frac{P_2}{(K_{4P} \times \Omega) + P_2}$
8	$P_2 + T_2 \xrightarrow{k_3} C$	$w_8 = k_3 \times P_2 \times T_2 / \Omega$
9	$C \xrightarrow{k_4} P_2 + T_2$	$w_9 = k_4 \times C$
10	$P_2 \xrightarrow{V_{dP}} \rightarrow$	$w_{10} = (V_{dP} \times \Omega) \frac{P_2}{(K_{dP} \times \Omega) + P_2}$
11	$V_{sT} \rightarrow M_T$	$w_{11} = (V_{sT} \times \Omega) \frac{(K_{IT} \times \Omega)^n}{(K_{IT} \times \Omega)^n + C_N^n}$
12	$M_T \xrightarrow{V_{mT}} \rightarrow$	$w_{12} = (V_{mT} \times \Omega) \frac{M_T}{(K_{mT} \times \Omega) + M_T}$
13	$M_T \xrightarrow{k_{sT}} M_T + T_0$	$w_{13} = k_{sT} \times M_T$
14	$T_0 \xrightarrow{V_{IT}} T_1$	$w_{14} = (V_{IT} \times \Omega) \frac{T_0}{(K_{IT} \times \Omega) + T_0}$
15	$T_1 \xrightarrow{V_{2T}} T_0$	$w_{15} = (V_{2T} \times \Omega) \frac{T_1}{(K_{2T} \times \Omega) + T_1}$
16	$T_1 \xrightarrow{V_{3T}} T_2$	$w_{16} = (V_{3T} \times \Omega) \frac{T_1}{(K_{3T} \times \Omega) + T_1}$
17	$T_2 \xrightarrow{V_{4T}} T_1$	$w_{17} = (V_{4T} \times \Omega) \frac{T_2}{(K_{4T} \times \Omega) + T_2}$
18	$T_2 \xrightarrow{V_{dT}} \rightarrow$	$w_{18} = (V_{dT} \times \Omega) \frac{T_2}{(K_{dT} \times \Omega) + T_2}$

Continues

Table 13.3. Nondeveloped stochastic version of the PER-TIM model for circadian rhythms [Gonze et al. 2003].—cont'd

Reaction Number	Reaction Step	Probability of Reaction
19	$C \xrightarrow{k_1} C_N$	$w_{19} = k_1 \times C$
20	$C_N \xrightarrow{k_2} C$	$w_{20} = k_2 \times C_N$
21	$M_P \xrightarrow{k_d} \rightarrow$	$w_{21} = k_d \times M_P$
22	$P_0 \xrightarrow{k_d} \rightarrow$	$w_{22} = k_d \times P_0$
23	$P_1 \xrightarrow{k_d} \rightarrow$	$w_{23} = k_d \times P_1$
24	$P_2 \xrightarrow{k_d} \rightarrow$	$w_{24} = k_d \times P_2$
25	$M_T \xrightarrow{k_d} \rightarrow$	$w_{25} = k_d \times M_T$
26	$T_0 \xrightarrow{k_d} \rightarrow$	$w_{26} = k_d \times T_0$
27	$T_1 \xrightarrow{k_d} \rightarrow$	$w_{27} = k_d \times T_1$
28	$T_2 \xrightarrow{k_d} \rightarrow$	$w_{28} = k_d \times T_2$
29	$C \xrightarrow{k_{dC}} \rightarrow$	$w_{29} = k_{dC} \times C$
30	$C_N \xrightarrow{k_{dN}} \rightarrow$	$w_{30} = k_{dN} \times C_N$

in a restricted domain in parameter space (see Section II.C). It was therefore interesting to check whether stochastic simulations were capable of reproducing this mode of dynamic behavior, which corresponds to the evolution to a strange attractor in the phase space. As shown in Figure 13.14, the strange attractor obtained by numerical integration of the deterministic Equations 13.2 can be recovered in corresponding conditions by simulations of the non-developed version of the stochastic model of Table 13.3. Here again, as illustrated in Figure 13.14, the larger the number of molecules of mRNAs and proteins involved in the oscillatory mechanism the closer the noisy stochastic trajectory is from the deterministic chaotic attractor.

The results obtained with stochastic models help to clarify the limits of validity of deterministic models for circadian oscillations. It appears that the deterministic approach provides a faithful picture as long as the number of molecules involved in the oscillatory mechanism exceeds a few tens or hundreds of molecules. Above this range, the larger the number of molecules the closer the stochastic trajectory from that predicted by the deterministic model.

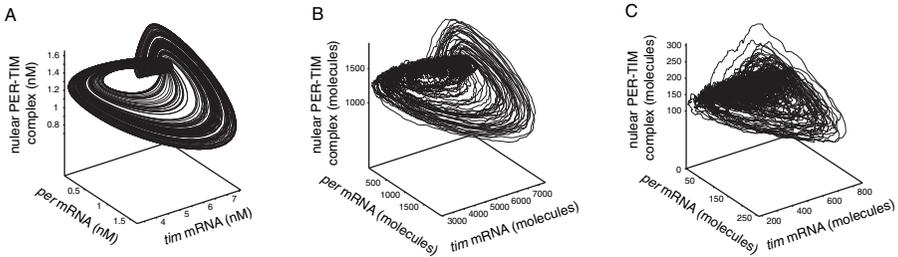


Figure 13.14. Effect of molecular noise on autonomous chaos. (a) Strange attractor corresponding to chaotic oscillations in the deterministic PER-TIM model for circadian rhythms. (b and c) Progressive dissolution of the strange attractor in the presence of molecular noise, for $\Omega = 1,000$ and 100 , respectively. The curve in a is obtained by numerical integration of Equations 13.2. In b and c, the curves are obtained by means of the Gillespie algorithm applied to the non-developed stochastic version of the PER-TIM model listed in Table 13.3 (Gonze et al. 2003).

IV. MODELING THE MAMMALIAN CIRCADIAN CLOCK

The molecular mechanism of circadian rhythms in mammals resembles that brought to light for *Drosophila*. In this organism, the negative feedback exerted by the PER-TIM complex is of an indirect rather than direct nature (Glossop et al. 1999). Thus, the transcription of the *per* and *tim* genes is triggered by a complex formed by the activators CYC and CLOCK. Binding of the PER-TIM complex to CYC and CLOCK prevents the activation of *per* and *tim* expression (Lee et al. 1999). In mammals the situation resembles that observed in *Drosophila*, but it is the CRY protein that forms a regulatory complex with a PER protein (Shearman et al. 2000; Reppert and Weaver 2002). Several forms of these proteins exist (PER1, PER2, PER3, CRY1, and CRY2). The complex CLOCK—BMAL1, formed by the products of the *Clock* and *Bmal1* genes, activates *Per* and *Cry* transcription. As in *Drosophila*, the PER-CRY complex inhibits the expression of the *Per* and *Cry* genes in an indirect manner, by binding to the complex CLOCK—BMAL1 (Lee et al. 2001; Reppert and Weaver 2002).

The mechanism of circadian rhythms in *Drosophila* and mammals thus relies on interlocked negative and positive feedback loops. In addition to the negative regulation of gene expression described previously, indirect positive regulation is involved. In *Drosophila*, the PER-TIM complex de-represses the transcription of *Clock* by binding to CLOCK, which exerts a negative autoregulation on the expression of its gene (Bae et al. 1998) via the product of the *vri* gene (Blau and Young 1999). In mammals, likewise, *Bmal1* expression is subjected to negative autoregulation by BMAL1, via the product of the *Rev-Erb α* gene (Preitner et al. 2002). The PER-CRY complex enhances *Bmal1* expression in an indirect manner (Reppert and Weaver 2002) by binding to CLOCK—BMAL1 and thereby decreasing the transcription of the *Rev-Erb α* gene (Preitner et al. 2002).

Models based on intertwined positive and negative regulatory loops have been proposed for *Drosophila* (Smolen et al. 2001; Ueda et al. 2001) and mammals (Forger and Peskin 2003, 2005; Leloup and Goldbeter 2003, 2004; Becker-Weimann et al. 2004). We shall focus here on the model proposed for the mammalian circadian clock, as it allows us to address the molecular dynamical bases of disorders of the human sleep/wake cycle associated with dysfunctions of the circadian clock.

A. Toward a detailed computational model for the mammalian circadian clock

The model for the mammalian circadian clock is schematized in Figure 13.15, both in a compact (a) and in a detailed manner (b). It describes the regulatory interactions between the products of the *Per*, *Cry*, *Bmal1*, and *Clock* genes. For simplicity, we do not distinguish between the *Per1*, *Per2*, and *Per3* genes and represent them in the model by a single *Per* gene. Similarly, *Cry1* and *Cry2* are represented by a single *Cry* gene. Moreover, as the *Clock* mRNA and its product (the CLOCK protein) are constitutively high in comparison to *Bmal1* mRNA and BMAL1 protein, they are considered in the model as parameters rather than variables.

We shall treat the regulatory effect of BMAL1 on *Bmal1* expression as a direct negative autoregulation. We have shown (Leloup and Goldbeter 2003) that similar conclusions are reached in an extended model in which the action of the REV-ERB α protein in the indirect negative feedback exerted by BMAL1 on the expression of its gene is considered explicitly. The version of the model without REV-ERB α is governed by a set of 16 kinetic equations (Leloup and Goldbeter 2003, 2004), whereas three more equations are needed in the extended model that incorporates the Rev-Erb α mRNA and the Rev-Erb α protein (Leloup and Goldbeter 2003).

In a certain range of parameter values, the 16- or 19-variable model for the mammalian clock produces sustained oscillations with a circadian period. These oscillations are endogenous, in that they occur for parameter values that remain constant in time, in agreement with the observation that circadian rhythms persist in continuous darkness or light. As observed experimentally (Lee et al. 2001; Reppert and Weaver 2002), *Bmal1* mRNA oscillates in antiphase with *Per* and *Cry* mRNAs (Figure 13.16a). The proteins also undergo antiphase oscillations and follow their mRNAs by a few hours (Figure 13.16b). Because most parameter values remain to be determined experimentally—as for the case of *Drosophila* (see Figures 13.2 and 13.3)—these oscillations were obtained for a semi-arbitrary choice of parameter values in a physiological range so as to yield a period of oscillations in continuous darkness (DD) close to 24 h.

To probe for entrainment of the circadian clock by LD cycles, we must incorporate the effect of light on *Per* expression. In continuous darkness, the maximum rate of *Per* expression, v_{sP} , remains at a low constant value. In LD, this rate varies periodically (e.g., as a square wave, going from a constant low value during the dark phase up to a higher constant value v_{sPmax} during the light phase). In such conditions, entrainment by a 12 : 12 LD cycle (12 h of light followed by 12 h of darkness) can be obtained over an appropriate range of v_{sPmax} values (Leloup and Goldbeter 2003).

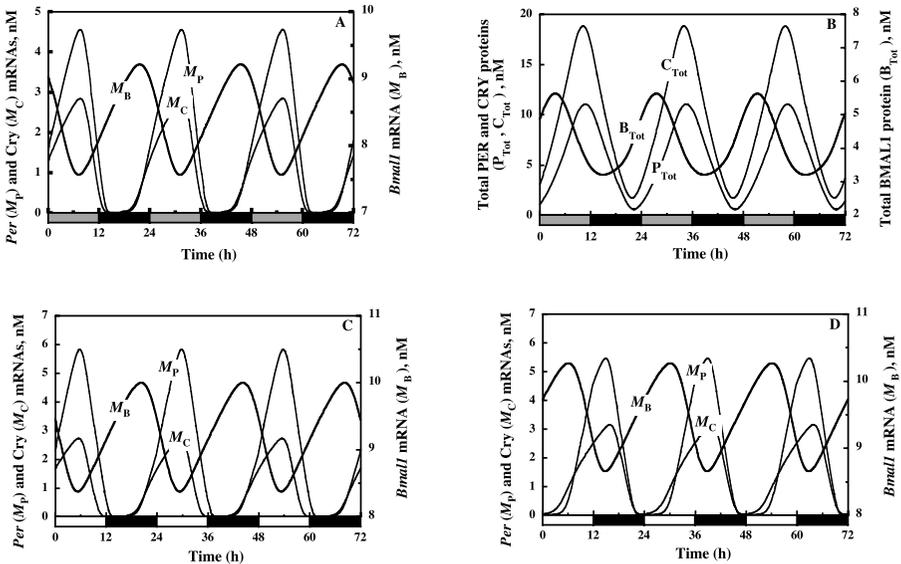


Figure 13.16. Circadian oscillations predicted by the mammalian clock model. (a) In constant darkness, the mRNA of *Bmal1* oscillates in antiphase with respect to the mRNAs of *Per* and *Cry*. (b) Corresponding protein oscillations in constant darkness. (c) Oscillations of the mRNAs after entrainment by 24-h light/dark (LD) cycles. The peak in *Per* mRNA occurs in the middle of the light phase. (d) Oscillations are delayed by 9 h and the peak in *Per* mRNA occurs in the dark phase when the value of parameter K_{AC} is decreased from 0.6 to 0.4 nM. Other parameter values correspond to the basal set of values listed in Table 1 in Leloup and Goldbeter (2003). In c and d, the maximum value of the rate of *Per* expression, v_{sp} , varies in a square-wave manner so that it remains at a constant low value of 1.5 nM/h during the 12-h-dark phase (black rectangle), and is raised up to the high value of 1.8 nM/h during the 12-h-light phase (white rectangle). The curves have been obtained by numerical integration of Equations 1 through 16 of the model without REV-ERB α (listed, together with parameter values, by Leloup and Goldbeter (2003)).

Figure 13.16c is a 10% change in parameter v_{mC} . The autonomous period in DD is 23.85 h and 23.70 h in Figures 13.16c and 13.16d, respectively, whereas the phase of *Per* mRNA is delayed by about 9 h in the latter case—so that *Per* mRNA reaches its maximum during the D phase instead of peaking in the L phase. This result is counterintuitive, in that we expect the maximum in *Per* mRNA to occur in phase L, because *Per* expression is enhanced by light. The virtue of the computational model is to alert us to the possibility that the phase of oscillations in LD may be highly labile, with the peak in *Per* mRNA shifting well into the D phase as a result of a small change in a light-insensitive parameter.

B. Multiple sources for oscillations in the circadian regulatory network

The genetic regulatory network underlying circadian rhythms contains intertwined positive and negative feedback loops. In view of the complexity of these regula-

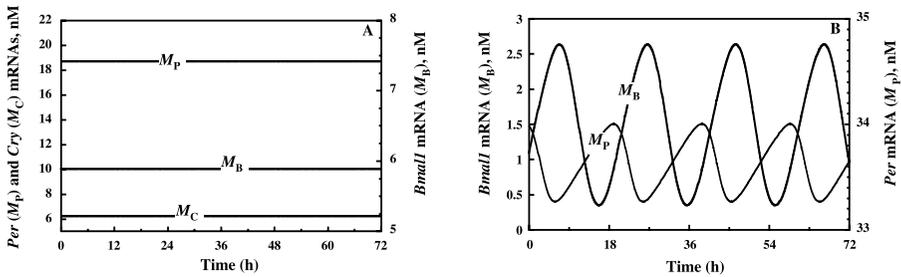


Figure 13.17. Multiple sources of oscillatory behavior in the genetic regulatory network controlling circadian rhythms. (a) Oscillations shown in Figures 13.16a and 13.16b disappear in the absence of PER protein synthesis ($k_{sp} = 0$). The curves show the asymptotic stable steady state reached after transients have subsided. (b) Sustained oscillations can nevertheless be restored when choosing a slightly different set of parameter values, even though $k_{sp} = 0$ (Leloup and Goldbeter 2003). The fact that oscillations can occur in the absence of PER protein indicates the existence of another oscillatory mechanism, which relies only on CLOCK-BMAL1 negative auto-regulation (see scheme in Figure 13.15a).

tory interactions, it should not be a surprise that more than one mechanism in the network may give rise to sustained oscillations. Evidence pointing to the existence of a second oscillatory mechanism (Leloup and Goldbeter 2003, 2004) stems from the fact that sustained oscillations generally disappear in the absence of PER protein (Figure 13.17a). However, even in such conditions sustained oscillations may occur with a period that is not necessarily circadian (Figure 13.17b). This second oscillator is based on the negative autoregulation exerted by BMAL1 on the expression of its gene, via the *Rev-Erb α* gene (see Figure 13.15).

Experimental observations so far suggest that if a second oscillator exists in the circadian regulatory network it does not manifest itself in producing rhythmic behavior. Thus, *mPer1/mPer2* (Zheng et al. 2001) or *mCry1/mCry2* (Van der Horst et al. 1999) double-knockout mice are arrhythmic. In some conditions, however, an extended light pulse can restore rhythmic behavior in a low proportion of *mPer1/mPer2* double-knockout mice (K. Bae and D. Weaver, personal communication).

In the absence of the negative feedback exerted by BMAL1 on the expression of its gene, oscillations can still originate from the PER—CRY negative feedback loop involving BMAL1. This result holds with the observation that circadian oscillations occur in the absence of REV-ERB α in mice (Preitner et al. 2002). Preventing altogether the synthesis of BMAL1 suppresses oscillations, because BMAL1 is involved in the mechanism of the two oscillators described previously.

C. Sensitivity analysis of the computational model for circadian rhythms

To assess the sensitivity of circadian oscillatory behavior to changes in parameter values, we determined for each parameter (one at a time) the range of values producing sustained oscillations (as well as the variation of the period over this range)

while keeping the other parameters set to their basal values (Leloup and Goldbeter, 2004; for an alternative sensitivity analysis, see Stelling et al. 2004). Such a sensitivity analysis was performed by constructing a series of bifurcation diagrams for four different sets of basal parameter values, each yielding circadian oscillations. Parameter set 1 was chosen so that oscillations disappear in the absence of PER protein or in the absence of negative autoregulation by BMAL1. Parameter set 2 corresponds to a situation in which oscillations can occur in the absence of PER, as a result of the negative autoregulation of BMAL1. Parameter set 3 corresponds to a situation in which circadian oscillations can occur in the absence of negative autoregulation by BMAL1. Finally, parameter set 4 was selected because oscillations can occur in the absence of PER or in the absence of negative autoregulation of BMAL1. On the basis of this analysis we may distinguish between two types of sensitivity: the first relates to the size of the oscillatory domain and the other to the influence on the period.

For some parameters the range of values producing sustained oscillations is quite narrow, less than one order of magnitude, whereas for other parameters it is much larger and extends over several orders of magnitude. The largest variation in period, by a factor close to 3, is observed for parameters that measure, respectively, the entry of the PER-CRY complex into the nucleus, and the formation of the inactive complex between PER-CRY and CLOCK-BMAL1 in the nucleus. For some sets of parameter values, the period may vary significantly (by a factor close to 2) over the oscillatory domain, whereas for other sets of parameter values the change in period as a function of this parameter may be reduced. Parameters for which the range of values yielding oscillations is narrowest are mainly those linked to BMAL1 and its mRNA. On the basis of these results, we may conclude that parameters affecting the level of BMAL1 possess the narrowest range of values producing sustained oscillations, whereas the period is most affected by the parameters measuring the entry of the PER-CRY complex into the nucleus and the formation of the inactive complex between PER-CRY and CLOCK-BMAL1.

D. From molecular mechanism to physiological disorders

The computational model for circadian oscillations in mammals provides us with the unique opportunity to address not only the molecular mechanism of a key biological rhythm but the dynamical bases of physiological disorders resulting from perturbations of the human circadian clock. Several disorders of the sleep/wake cycle are indeed associated with dysfunctions of the circadian clock in humans. In the familial advanced sleep/phase syndrome (FASPS), the phase of the sleep/wake cycle in LD is advanced by several hours, as a result of a decreased rate of PER phosphorylation (Toh et al. 2001). In a family in which FASPS is present over five generations, those affected by the syndrome tend to go to sleep around 7:30 p.m. and awake around 4:30 a.m. Moreover, in a patient affected by FASPS the period of the circadian clock in DD was reduced down to 23.5 h from a normal mean value of 24.4 h (Jones et al. 1999).

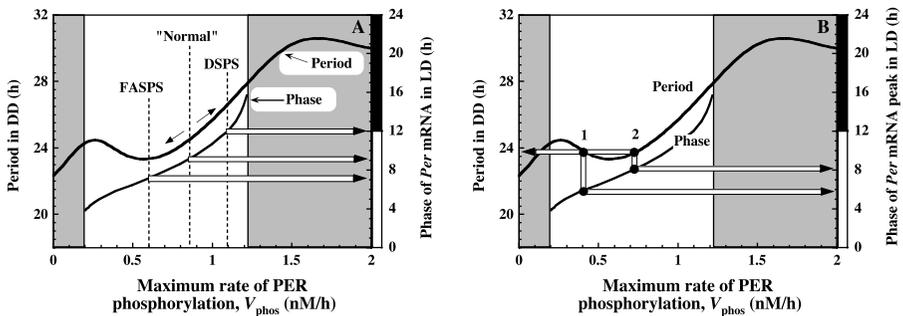


Figure 13.18. Relating the mammalian clock model to syndromes associated with disorders of the sleep/wake cycle in humans (Leloup and Goldbeter 2003). (a) Effect of the maximum rate of PER phosphorylation on the free running period in DD and on the phase of the oscillations in LD. The phase corresponds to the time (in h) at which the maximum in *Per* mRNA occurs after the onset of the L phase. Decreasing (increasing) the rate of phosphorylation of the PER protein, V_{phos} , with respect to the “normal” situation can produce a phase advance (delay) as well as a decrease (increase) in free running period that accounts for the phase shift observed in the familial advanced sleep phase syndrome (FASPS) or the delayed sleep phase syndrome (DSPS). (b) Situations 1 and 2 show that different values of the control parameter can produce different phases after entrainment, even though they correspond to the same free running period in DD. The gray areas on the left and right in the two panels refer to absence of entrainment (see Figure 13.19).

The phase advance characteristic of FASPS can be accounted for by the model as a result of a decrease in parameter V_{phos} , which measures the maximum rate of PER phosphorylation by the protein kinase CK1 ϵ . As in clinical observations (Jones et al. 1999), the advance of the phase in LD then accompanies a decrease in autonomous period as the phosphorylation rate decreases (Leloup and Goldbeter 2003). Such a decrease in period in DD can be observed over parts of the bifurcation diagram established as a function of V_{phos} (see Figure 13.18a). The model could be used similarly to address the delayed sleep phase syndrome, which is the mirror physiological disorder of the sleep/wake cycle and appears to be associated with increased rate of PER phosphorylation (Ebisawa et al. 2001; Archer et al. 2003). The bifurcation diagram of Figure 13.18a indicates that an increase in V_{phos} may correspond to a delayed phase of the sleep/wake cycle in LD, and to an increase in the autonomous period of circadian oscillations in DD. An interesting prediction arising from Figure 13.18b is that two distinct values of V_{phos} may yield the same period in DD and different phases upon entrainment in LD.

For a long time the model for the mammalian circadian clock placed us in a quandary, as the model failed to account for the most conspicuous property of circadian rhythms; namely, their entrainment by LD cycles. There is generally a range of parameter values in which entrainment occurs, but we failed to find any such range when the light-sensitive parameter (the maximum rate of *Per* expression) was made to vary in a square wave manner. Regardless of the magnitude of the periodic variation, entrainment did not occur. We then realized that the level of CRY

protein was critical for entrainment by LD cycles. When the level of CRY remains too low, free PER builds up during successive light phases, as there is not enough CRY with which to form a complex. Consequently, entrainment fails to occur (Leloup and Goldbeter 2003). It was sufficient to raise the level of CRY—by increasing the rate of PER synthesis or the rate of *Per* expression, or by decreasing the rate of degradation of either PER or *Per* mRNA—for entrainment to occur.

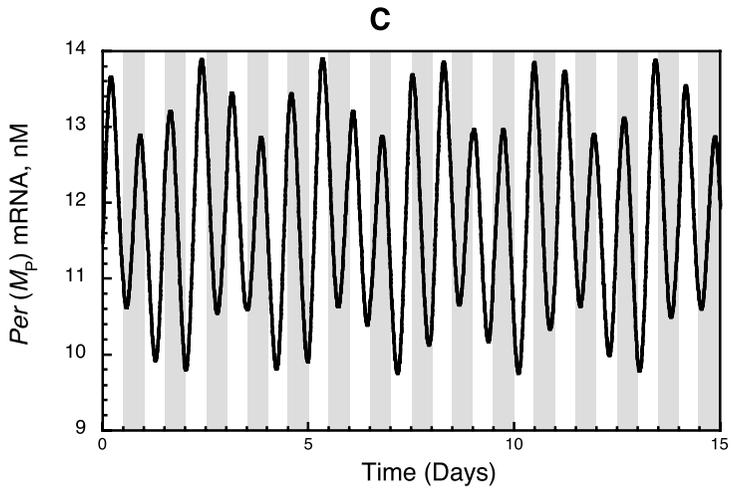
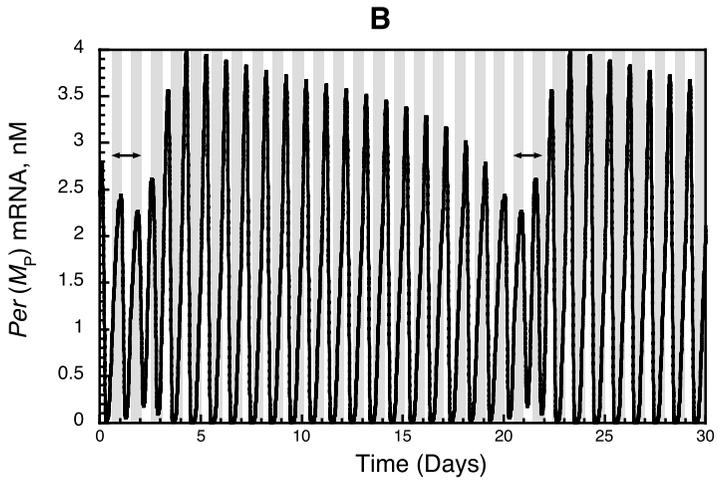
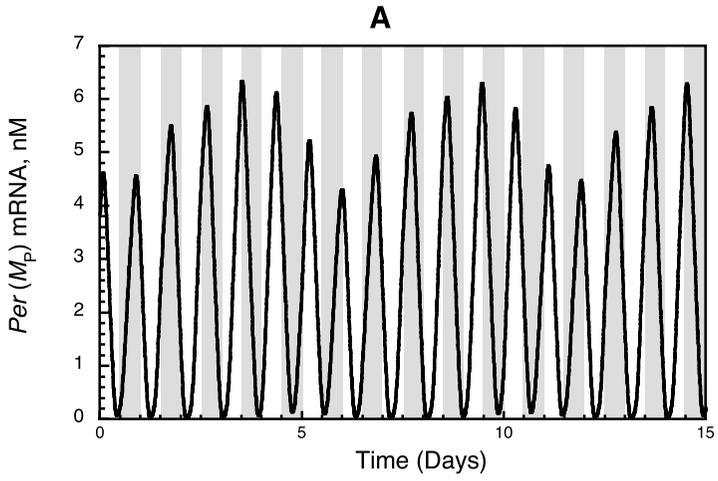
If entrainment failure is so easy to obtain in the model, could it be that a corresponding syndrome exists in human physiology? The answer is yes: there is a condition known as non-24-h sleep/wake syndrome (Richardson and Malin 1996), in which the time at which the subject goes to sleep is drifting every day. This slow drift is sometimes accompanied by “jumps” in the phase ϕ of the sleep/wake cycle in LD conditions. During such jumps, ϕ rapidly traverses one phase of the LD cycle in a few days, and slowly drifts across the other phase of the LD cycle during a much longer time (on the order of several weeks). The absence of entrainment in the model corresponds to quasi-periodic oscillations in LD. These oscillations can be associated or not with phase jumps, as shown in Figure 13.19 in panels A and B, respectively. Chaotic oscillations may also result from the periodic forcing by LD cycles (Figure 13.19c).

We are currently using the model to search for conditions other than decreased levels of CRY, which might also lead to the failure of entrainment in LD. If the non-24-h sleep/wake cycle syndrome is indeed due to altered levels of CRY, the results suggest that restoring adequate levels of the protein might allow entrainment to occur.

V. CONCLUSIONS

Remarkable advances have been made during the last two decades in unraveling the molecular bases of circadian rhythms—first in *Drosophila* and *Neurospora*, and more recently in cyanobacteria, plants, and mammals. Based on experimentally determined mechanisms, computational models of increasing complexity have been proposed for these rhythms. As reviewed in this chapter, computational approaches throw light on the precise conditions in which circadian oscillations occur as a result of genetic regulation. The models also account for a variety of

Figure 13.19. Absence of entrainment and the non-24-h sleep/wake cycle syndrome. The phase of the circadian oscillations does not always lock to a constant value with respect to the 24-h LD cycle, in contrast to what occurs in the case of entrainment. Lack of entrainment can lead to quasi-periodic behavior (a), which is sometimes accompanied by phase jumps (b) corresponding to slow drifts of the phase followed by rapid progression through the L or D phase (horizontal arrows). Chaotic behavior (c) can also be observed as a result of forcing by the LD cycle. Gray and white columns represent the D and L phases of the LD cycles, respectively. Parameter values are as in Table 1 of Leloup and Goldbeter (2003), with $v_{mP} = 0.95 \text{ nMh}^{-1}$ (a), 1.45 nMh^{-1} (b), or 0.70 nMh^{-1} (c).



properties of circadian rhythms, such as phase shifting or long-term suppression by light pulses, entrainment by light/dark cycles, and temperature compensation.

When the numbers of molecules of protein or mRNA involved in the oscillatory mechanism are very low, it becomes necessary to resort to stochastic approaches. We have shown by means of stochastic simulations that coherent sustained oscillations emerge from molecular noise in the genetic regulatory network as soon as the maximum numbers of mRNA and clock protein molecules are in the tens and hundreds, respectively. At higher numbers of molecules, the stochastic models yield results that are largely similar to the predictions of the corresponding deterministic models. The latter therefore provide a useful representation of circadian oscillatory behavior over a wide range of conditions.

Among the factors that contribute to the robustness of circadian rhythms with respect to molecular noise are the degree of cooperativity of repression, the distance from a bifurcation point, and the rate constants measuring the binding of the repressor to the gene. All models considered here pertain to the onset of circadian rhythms at the cellular level. The intercellular coupling of oscillatory cells—for example, in the suprachiasmatic nuclei (SCN), which represent the central circadian pacemaker in mammals (Kunz and Achermann 2003; Gonze et al. 2005)—may further contribute to the robustness of circadian rhythms.

The computational approach supports the view (Reppert and Weaver 2002) that the genetic regulatory mechanism of sustained circadian oscillations is similar in both the central and peripheral (Schibler et al. 2003; Yoo et al. 2004) oscillators, and that the observed differences in phase are of a quantitative rather than qualitative nature.

We have used the case of circadian rhythms to show how more and more complex computational models must be considered to accommodate the accelerating flux of new experimental observations. A question that arises naturally is whether such an ever-increasing complexity of the models is really needed. It appears that as with geographical maps a balance must be made between the necessity of including the most relevant details and the desire to not become lost in a too meticulous description, because the model might quickly become so complex that its detailed numerical study would become highly cumbersome.

An example of molecular detail that has to be incorporated is the phosphorylation of the PER protein: even if sustained oscillations are possible, in principle, in the absence of PER covalent modification the phosphorylation step is needed not only to account for the effect of mutations in the protein kinase that phosphorylates PER but also to account for some disorders of the sleep/wake cycle in humans related to altered PER phosphorylation. Moreover, as described in this chapter, several results can only be obtained in models that possess a minimum degree of complexity. Thus, autonomous chaos was obtained in the 10-variable model for circadian rhythms in *Drosophila* incorporating the formation of a PER-TIM complex, but not in the five-variable model based on PER alone. In the mammalian clock model, incorporation of additional feedback loops brought to light the possibility of multiple sources of oscillatory behavior.

Finally, circadian rhythms provide a case in point for showing how computational models can be used to address a wide range of issues, extending from molecular mechanism to physiological disorders. Identifying the origin of dysfunctions and predicting ways of obviating them in metabolic or genetic regulatory networks on the basis of numerical simulations presents a key challenge for computational biology.

ACKNOWLEDGMENTS

This work was supported by grants 3.4607.99 and 3.4636.04 from the Fonds de la Recherche Scientifique Médicale (F.R.S.M., Belgium), DARPA-AFRL grant F30602-02-0554, and the BIOSIM Network of Excellence within the FP6 Program of the European Union. J.-C. L. and D. G. are, respectively, Chercheur qualifié and Chargé de recherches du Fonds National de la Recherche Scientifique (F.N.R.S., Belgium). This chapter was prepared while A. G. held a Chaire Internationale de Recherche Blaise Pascal de l'Etat et de la Région d'Ile-de-France, gérée par la Fondation de l'Ecole Normale Supérieure at the University of Paris Sud-Orsay (France), in the Institute of Genetics and Microbiology directed by Professor Michel Jacquet, whose hospitality is gratefully acknowledged.

REFERENCES

- Alabadi, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Mas, P., and Kay, S. A. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science* **293**:880–883.
- Archer, S. N., Robilliard, D. L., Skene, D. J., Smits, M., Williams, A., Arendt, J., and von Schantz, M. (2003). A length polymorphism in the circadian clock gene *Per3* is linked to delayed sleep phase syndrome and extreme diurnal preference. *Sleep* **26**:413–415.
- Arkin, A., Ross, J., and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**:1633–1648.
- Aronson, B. D., Johnson, K. A., Loros, J. J., and Dunlap, J. C. (1994). Negative feedback defining a circadian clock: Autoregulation of the clock gene frequency. *Science* **263**:1578–1584.
- Bae, K., Lee, C., Sidote, D., Chuang, K.-Y., and Edery, I. (1998). Circadian regulation of a *Drosophila* homolog of the mammalian *Clock* gene: PER and TIM function as positive regulators. *Mol. Cell. Biol.* **18**:6142–6151.
- Baras, F. (1997). Stochastic analysis of limit cycle behaviour. In *Stochastic Dynamics: Lecture Notes in Physics* (L. Schimansky-Geier and T. Poeschel eds.), pp. 167–178, Berlin: Springer.
- Baras, F., Pearson, J. E., and Mansour, M. M. (1990). Microscopic simulation of chemical oscillations inhomogeneous systems. *J. Chem. Phys.* **93**:5747–5750.
- Barkai, N., and Leibler, S. (2000). Circadian clocks limited by noise. *Nature* **403**:267–268.

- Baylies, M. K., Weiner, L., Vosshall, L. B., Saez, L., and Young, M. W. (1993). Genetic, molecular, and cellular studies of the *per* locus and its products in *Drosophila melanogaster*. In *Molecular Genetics of Biological Rhythms* (M. W. Young ed.), pp. 123–153, New-York: Marcel Dekker.
- Becker-Weimann, S., Wolf, J., Herzel, H., and Kramer, A. (2004). Modeling feedback loops of the mammalian circadian oscillator. *Biophys. J.* **87**:3023–3034.
- Blau, J., and Young, M. W. (1999). Cycling vrille expression is required for a functional *Drosophila* clock. *Cell* **99**:661–671.
- Busza, A., Emery-Le, M., Rosbash, M., and Emery, P. (2004). Roles of the two *Drosophila* CRYPTOCHROME structural domains in circadian photoreception. *Science* **304**:1503–1506.
- Darlington, T. K., Wager-Smith, K., Ceriani, M. F., Staknis, D., Gekakis, N., Steeves, T. D. L., Weitz, C. J., Takahashi, J. S., and Kay, S. A. (1998). Closing the circadian loop: CLOCK-induced transcription of its own inhibitors *per* and *tim*. *Science* **280**:1599–1603.
- Doedel, E. J. (1981). AUTO: A program for the automatic bifurcation analysis of autonomous systems. *Cong. Numer.* **30**:265–384.
- Dunlap, J. C. (1993). Genetic analysis of circadian clocks. *Annu. Rev. Physiol.* **55**:683–728.
- Dunlap, J. C. (1999). Molecular bases for circadian clocks. *Cell* **96**:271–290.
- Ebisawa, T., Uchiyama, M., Kajimura, N., Mishima, K., Kamei, Y., Katoh, M., Watanabe, T., Sekimoto, M., Shibui, K., Kim, K., et al. (2001). Association of structural polymorphisms in the human *period3* gene with delayed sleep phase syndrome. *EMBO Rep.* **2**:342–346.
- Edey, I., Zwiebel, L. J., Dembinska, M. E., and Rosbash, M. (1994). Temporal phosphorylation of the *Drosophila* period protein. *Proc. Natl. Acad. Sci. USA* **91**:2260–2264.
- Forger, D. B., and Peskin, C. S. (2003). A detailed predictive model of the mammalian circadian clock. *Proc. Natl. Acad. Sci. USA* **100**:14806–14811.
- Forger, D. B., and Peskin, C. S. (2005). Stochastic simulation of the mammalian circadian clock. *Proc. Natl. Acad. Sci. USA* **102**:321–324.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**:403–434.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**:2340–2361.
- Glossop, N. R. J., Lyons, L. C., and Hardin, P. E. (1999). Interlocked feedback loops within the *Drosophila* circadian oscillator. *Science* **286**:766–768.
- Goldbeter, A. (1995). A model for circadian oscillations in the *Drosophila* period protein (PER). *Proc. R. Soc. London Ser. B* **261**:319–324.
- Goldbeter, A. (1996). *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behavior*. Cambridge UK: Cambridge University Press.
- Gonze, D., and Goldbeter, A. (2000). Entrainment versus chaos in a model for a circadian oscillator driven by light-dark cycles. *J. Stat. Phys.* **101**:649–663.
- Gonze, D., Leloup, J.-C., and Goldbeter, A. (2000). Theoretical models for circadian rhythms in *Neurospora* and *Drosophila*. *C. R. Acad. Sci. III.* **323**:57–67.
- Gonze, D., Halloy, J., and Goldbeter, A. (2002a). Deterministic versus stochastic models for circadian rhythms. *J. Biol. Phys.* **28**:637–653.
- Gonze, D., Halloy, J., and Goldbeter, A. (2002b). Robustness of circadian rhythms with respect to molecular noise. *Proc. Natl. Acad. Sci. USA* **99**:673–678.
- Gonze, D., Roussel, M. R., and Goldbeter, A. (2002c). A model for the enhancement of fitness in cyanobacteria based on resonance of a circadian oscillator with the external light-dark cycle. *J. Theor. Biol.* **214**:577–597.

- Gonze, D., Halloy, J., Leloup, J.-C., and Goldbeter, A. (2003). Stochastic models for circadian rhythms: Effect of molecular noise on periodic and chaotic behaviour. *C. R. Biologies* **326**:189–203.
- Gonze, D., Halloy, J., and Goldbeter, A. (2004a). Emergence of coherent oscillations in stochastic models for circadian rhythms. *Physica A* **342**:221–233.
- Gonze, D., Halloy, J., and Goldbeter, A. (2004b). Stochastic models for circadian oscillations: Emergence of a biological rhythm. *Int. J. Quantum Chem.* **98**:228–238.
- Gonze, D., Bernard, S., Waltermann, C., Kramer, A., and Herzog, H. (2005). Spontaneous synchronization of coupled circadian oscillators. *Biophys. J.* **89**:120–129.
- Goodwin, B. C. (1965). Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.* **3**:425–438.
- Grima, B., Lamouroux, A., Chelot, E., Papin, C., Limbourg-Bouchon, B., and Rouyer, F. (2002). The F-box protein Slimb controls the levels of clock proteins Period and Timeless. *Nature* **420**:178–182.
- Hall, J. C., and Rosbash, M. (1988). Mutations and molecules influencing biological rhythms. *Annu. Rev. Neurosci.* **11**:373–393.
- Hardin, P. E., Hall, J. C., and Rosbash, M. (1990). Feedback of the *Drosophila* period gene product on circadian cycling of its messenger RNA levels. *Nature* **343**:536–540.
- Hardin, P. E., Hall, J. C., and Rosbash, M. (1992). Circadian oscillations in period gene mRNA levels are transcriptionally regulated. *Proc. Natl. Acad. Sci. USA* **89**:11711–11715.
- Hunter-Ensor, M., Ousley, A., and Sehgal, A. (1996). Regulation of the *Drosophila* protein timeless suggests a mechanism for resetting the circadian clock by light. *Cell* **84**:677–685.
- Jewett, M. E., and Kronauer, R. E. (1998). Refinement of a limit cycle oscillator model of the effects of light on the human circadian pacemaker. *J. Theor. Biol.* **192**:455–465.
- Jones, C. R., Campbell, S. S., Zone, S. E., Cooper, F., DeSano, A., Murphy, P. J., Jones, B., Czajkowski, L., and Ptacek, L. J. (1999). Familial advanced sleep-phase syndrome: A short-period circadian rhythm variant in humans. *Nat. Med.* **5**:1062–1065.
- Ko, H. W., Jiang, J., and Edery, I. (2002). Role for Slimb in the degradation of *Drosophila* Period protein phosphorylated by Doubletime. *Nature* **420**:673–678.
- Konopka, R. J. (1979). Genetic dissection of the *Drosophila* circadian system. *Fed. Proc.* **38**:2602–2605.
- Konopka, R. J., and Benzer, S. (1971). Clock mutants of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **68**:2112–2116.
- Konopka, R. J., Pittendrigh, C., and Orr, D. (1989). Reciprocal behaviour associated with altered homeostasis and photosensitivity of *Drosophila* clock mutants. *J. Neurosci.* **6**:1–10.
- Kunz, H., and Achermann, P. (2003). Simulation of circadian rhythm generation in the suprachiasmatic nucleus with locally coupled self-sustained oscillators. *J. Theor. Biol.* **224**:63–78.
- Lee, C., Parikh, V., Itsukaichi, T., Bae, K., and Edery, I. (1996). Resetting the *Drosophila* clock by photic regulation of PER and a PER-TIM complex. *Science* **271**:1740–1744.
- Lee, C., Bae, K., and Edery, I. (1999). PER and TIM inhibit the DNA binding activity of a *Drosophila* CLOCK-CYC/DBMAL1 heterodimer without disrupting formation of the heterodimer: A basis for circadian transcription. *Mol. Cell. Biol.* **19**:5316–5325.
- Lee, K., Loros, J. J., and Dunlap, J. C. (2000). Interconnected feedback loops in the *Neurospora* circadian system. *Science* **289**:107–110.
- Lee, C., Etchegaray, J. P., Cagampang, F. R., Loudon, A. S., and Reppert, S. M. (2001). Post-translational mechanisms regulate the mammalian circadian clock. *Cell* **107**:855–867.

- Leloup, J.-C., and Goldbeter, A. (1997). Temperature compensation of circadian rhythms: Control of the period in a model for circadian oscillations of the PER protein in *Drosophila*. *Chronobiol. Int.* **14**:511–520.
- Leloup, J.-C., and Goldbeter, A. (1998). A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J. Biol. Rhythms* **13**:70–87.
- Leloup, J.-C., and Goldbeter, A. (1999). Chaos and birhythmicity in a model for circadian oscillations of the PER and TIM proteins in *Drosophila*. *J. Theor. Biol.* **198**:445–459.
- Leloup, J.-C., and Goldbeter, A. (2001). A molecular explanation for the long-term suppression of circadian rhythms by a single light pulse. *Am. J. Physiol. Regul. Integrat. Comp. Physiol.* **280**:R1206–R1212.
- Leloup, J.-C., and Goldbeter, A. (2003). Toward a detailed computational model for the mammalian circadian clock. *Proc. Natl. Acad. Sci. USA* **100**:7051–7056.
- Leloup, J.-C., and Goldbeter, A. (2004). Modeling the mammalian circadian clock: Sensitivity analysis and multiplicity of oscillatory mechanisms. *J. Theor. Biol.* **230**:541–562.
- Leloup, J.-C., Gonze, D., and Goldbeter, A. (1999). Limit cycle models for circadian rhythms based on transcriptional regulation in *Neurospora* and *Drosophila*. *J. Biol. Rhythms* **14**:433–448.
- McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94**:814–819.
- Morrow, M. W., Garceau, N. Y., and Dunlap, J. C. (1997) Dissection of a circadian oscillation into discrete domains. *Proc. Natl. Acad. Sci. USA* **94**:3877–3882.
- Moore-Ede, M. C., Sulzman, F. M., and Fuller, C. A. (1982). *The Clocks That Time Us: Physiology of the Circadian Timing System*. Cambridge, MA: Harvard University Press.
- Morton-Firth, C. J., and Bray, D. (1998). Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.* **192**:117–128.
- Myers, M. P., Wager-Smith, K., Rothenfluh-Hilfiker, A., and Young, M. W. (1996). Light-induced degradation of TIMELESS and entrainment of the *Drosophila* circadian clock. *Science* **271**:1736–1740.
- Nicolis, G., and Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations*. New York: Wiley.
- Pittendrigh, C. S. (1960). Circadian rhythms and the circadian organization of living systems. *Cold Spring Harbor Symp. Quant. Biol.* **25**:159–184.
- Preitner, N., Damiola, F., Lopez-Molina, L., Zakany, J., Duboule, D., Albrecht, U., and Schibler, U. (2002). The orphan nuclear receptor REV-ERB α controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell* **110**:251–260.
- Qiu, J., and Hardin, P. E. (1996). *per* mRNA cycling is locked to lights-off under photoperiodic conditions that support circadian feedback loop function. *Mol. Cell. Biol.* **16**:4182–4188.
- Reppert, S., and Weaver, D. (2002). Coordination of circadian timing in mammals. *Nature* **418**:935–941.
- Richardson, G. S., and Malin, H. V. (1996). Circadian rhythm sleep disorders: Pathophysiology and treatment. *J. Clin. Neurophysiol.* **13**:17–31.
- Ruoff, P., and Rensing, L. (1996). The temperature-compensated Goodwin model simulates many circadian clock properties. *J. Theor. Biol.* **179**:275–285.
- Ruoff, P., Vinsjevik, M., Monnerjahn, C., and Rensing, L. (2001). The Goodwin model: Simulating the effect of light pulses on the circadian sporulation rhythm of *Neurospora crassa*. *J. Theor. Biol.* **209**:29–42.

- Rutila, J. E., Suri, V., Le, M., So, W. V., Rosbash, M., and Hall, J. C. (1998). CYCLE is a second bHLH-PAS clock protein essential for circadian rhythmicity and transcription of *Drosophila period* and *timeless*. *Cell* **93**:805–814.
- Sassone-Corsi, P. (1994). Rhythmic transcription and autoregulatory loops: Winding up the biological clock. *Cell* **78**:361–364.
- Schibler, U., Ripperger, J., and Brown, S. A. (2003). Peripheral circadian oscillators in mammals: Time and food. *J. Biol. Rhythms* **18**:250–260.
- Shearman, L. P., Sriram, S., Weaver, D. R., Maywood, E. S., Chaves, I., Zheng, B., Kume, K., Lee, C. C., van der Horst, G. T., Hastings, M. H., and Reppert, S. M. (2000). Interacting molecular loops in the mammalian circadian clock. *Science* **288**:1013–1019.
- Smolen, P., Baxter, D. A., and Byrne, J. H. (2001). Modeling circadian oscillations with interlocking positive and negative feedback loops. *J. Neurosci.* **21**:6644–6656.
- Stelling, J., Gilles, E. D., and Doyle, F. J. 3rd (2004). Robustness properties of circadian clock architectures. *Proc. Natl. Acad. Sci. USA* **101**:13210–13215.
- Takahashi, J. S. (1992). Circadian clock genes are ticking. *Science* **258**:238–240.
- Toh, K. L., Jones, C. R., He, Y., Eide, E. J., Hinz, W. A., Virshup, D. M., Ptacek, L. J., and Fu, Y.-H. (2001). An *hPer2* phosphorylation site mutation in familial advanced sleep-phase syndrome. *Science* **291**:1040–1043.
- Ueda, H. R., Hagiwara, M., and Kitano, H. (2001). Robust oscillations within the interlocked feedback model of *Drosophila* circadian rhythm. *J. Theor. Biol.* **210**:401–406.
- van der Horst, G. T., Muijtjens, M., Kobayashi, K., Takano, R., Kanno, S., Takao, M., de Wit, J., Verkerk, A., Eker, A. P., van Leenen, D., et al. (1999). Mammalian *Cry1* and *Cry2* are essential for maintenance of circadian rhythms. *Nature* **398**:627–630.
- Yoo, S. H., Yamazaki, S., Lowrey, P. L., Shimomura, K., Ko, C. H., Buhr, E. D., Siepkka, S. M., Hong, H. K., Oh, W. J., Yoo, O. J., Menaker, M., and Takahashi, J. S. (2004). PERIOD2: LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proc. Natl. Acad. Sci. USA* **101**:5339–5346.
- Young, M. W., and Kay, S. A. (2001). Time zones: A comparative genetics of circadian clocks. *Nat. Rev. Genet.* **2**:702–715.
- Zeng, H., Hardin, P. E., and Rosbash, M. (1994). Constitutive overexpression of the *Drosophila period* protein inhibits *period* mRNA cycling. *EMBO J.* **13**:3590–3598.
- Zeng, H., Qian, Z., Myers, M. P., and Rosbash, M. (1996). A light-entrainment mechanism for the *Drosophila* circadian clock. *Nature* **380**:129–135.
- Zerr, D. M., Hall, J. C., Rosbash, M., and Siwicki, K. K. (1990). Circadian fluctuations of period protein immunoreactivity in the CNS and the visual system of *Drosophila*. *J. Neurosci.* **10**:2749–2762.
- Zheng, B., Albrecht, U., Kaasik, K., Sage, M., Lu, W., Vaishnav, S., Li, Q., Sun, Z. S., Eichele, G., Bradley, A., and Lee, C. C. (2001). Nonredundant roles of the *mPer1* and *mPer2* genes in the mammalian circadian clock. *Cell* **105**:683–694.

Multistability and Multicellularity: Cell Fates as High-Dimensional Attractors of Gene Regulatory Networks

Sui Huang

*Vascular Biology Program, Children's Hospital,
Harvard Medical School, Boston, Massachusetts, USA*

Chapter 14

ABSTRACT

Cells in multicellular organisms exhibit discrete mutually exclusive phenotypic states, such as proliferation, apoptosis, or differentiation into various cell types. Each of these “cell fates” is associated with a particular stable genome-wide gene expression profile defined by 25,000 genes. To explain the collapse of the hyperastronomical number of combinatorially possible expression configurations into those characteristic of observable cell fates, the latter have been proposed to be high-dimensional attractors in gene activity state space.

Here we review the biology of cell fate regulation from a “systems” perspective and discuss two gene network models (small systems of differential equations and high-dimensional Boolean networks) to illustrate how molecular interactions produce multistability and attractors. Implications for cell fate regulation, stem cell multipotency, stochastic fate decisions, and cancer are discussed. This chapter also illustrates the necessity for embracing both pathway details as well as simplifying abstraction in computational systems biology.

I. INTRODUCTION

A hallmark of multicellular organisms is the differentiation of cells into functionally distinct cell types, such as a resting nerve cell or a proliferating skin cell. A cell type represents a special case of a cell fate—the more general and abstract term that encompasses any distinct functional phenotypic state a cell can occupy—such as *proliferation* (the state in which the cell undergoes the cell division cycle), *quiescence* (the state in which the cell is not dividing, but viable), *apoptosis* (state of

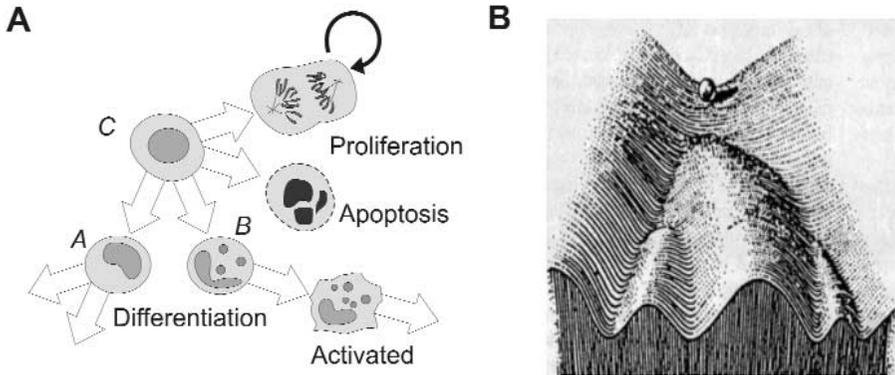


Figure 14.1. (a) Schematic representation of cell fates and cell fate transitions. Note that for every tissue the “fate transition map” is different in that not all fates are equally accessible in all tissues. In this example, cell C may be regarded as a multipotent progenitor cell capable of differentiating in the two mature cells A and B, which themselves can undergo various fate transitions. (b) Waddington’s “epigenetic landscape” captures the discrete nature of cell fate decisions. (Reprinted from C. F. Waddington, *The Strategy of the Gene*, 1957, Allen and Unwin.)

commitment to programmed cell death), or *differentiation into a mature cell type* (the typically nonproliferative state in which the cell exerts tissue-specific functions), and the various *activation states* of the differentiated cell (Figure 14.1a).

Development and tissue homeostasis require that cells undergo transitions between these cell fates in a tightly controlled manner. Cell fates appear as discrete, stable, and mutually exclusive states or processes, and hence represent behavioral entities (or “programs”) identifiable as such. Consequently, cells “switch” in an almost discontinuous manner between cell fates. Depending on the particular tissue, the terminology can overlap; that is, the mutual exclusiveness is relative. For instance, in the adult a liver cell can switch between a differentiated state and a proliferative state. In the latter, many liver-specific genes are shut down and the cell undergoes repeated cell divisions. However, this cell is still a liver cell. In contrast, a nerve cell (like many mature cells) usually cannot access the proliferative state anymore. Such cell fates are often referred to as post-mitotic or terminally differentiated.

The past decades of molecular biology have focused on the biochemical pathways that control the switch between cell fates. One general paradigm that has emerged from the gene-centered view of this period is that of signal transduction pathways: a cascade of biochemical reactions, typically protein-to-protein interactions, connects the extracellular input signal, carried by a soluble growth factor, to the effector genes that in turn control gene regulatory pathways to produce proteins necessary for the new cell fate. For instance, a nerve cell differentiation factor would induce the expression of neuronal genes and repress genes specific to other tissues.

This molecular analysis has brought great insight into the very material basis of how a biological signal is propagated along a pathway in a cell. However, in this pathway-focused view of biology we sometimes miss the larger picture. First, genes and proteins do not form independent pathway modules to which a distinct “biological function” can be assigned. Instead, they collectively establish an almost genome-wide network of regulatory interactions. This has been heralded since the early days of signal transduction research by the ever-increasing discovery of “cross talks” between historically defined molecular pathways (Bouvier 1990). The post-genomic systematic characterization of interaction between genes and proteins has led to the picture of a genome-scale network rather than the collection of parallel pathways devoted to individual biological functions (Huang 2004; Marcotte 2001). In fact, many molecular biologists now even equate systems biology with “network biology.”

Second, the broader picture is important because cells exhibit a distinct behavior in the tissue context with respect to the dynamics of their phenotypic states. The actual macroscopic dynamics of cell fate regulation itself must also be considered when studying the details of regulatory pathways because we cannot understand the *regulating* mechanisms without knowing the *regulated* behavior. The fundamental feature of discreteness of cell fates, which current biologists interested in molecular processes within the cell need not give heed to, has most elegantly been articulated by the great embryologist of the last century, C. F. Waddington. He noted “well-recognizable types” and that “intermediates” are rare and unstable (Waddington 1956), and proposed in the 1940s the “epigenetic landscape” (shown in Figure 14.1b) as an intuitive metaphor to explain the fact that cells are forced to take all-or-none decisions between distinct cell fates. Because each cell fate is associated with a distinct gene expression profile, the latter must also exhibit discrete states that change into one another through quasi-discontinuous transitions.

In this chapter, we will emphasize the biology of cell fate regulation by focusing on the picture of discrete cell fates and their transitions and develop the concepts that may explain how the natural dynamics of large (genome-scale) gene regulatory networks produce the characteristic cell fate behavior. A major aim of this chapter is to review the ideas of Boolean gene networks as an abstract model able to capture generic cell fate behavior in the light of systems biology. We will also discuss specific implications for understanding properties of whole-cell behavior that remain difficult to explain in the signaling pathway paradigm.

II. GENE EXPRESSION PROFILES IN GENE EXPRESSION STATE SPACE

Traditionally, an individual protein that is expressed only in a certain cell type is referred to as a cell-type specific *marker*. For instance, the protein albumin is expressed “specifically” in liver cells, and its promoter element can be used to direct liver-specific expression of transgenes. With the advent of massively parallel

measurement of gene expression by DNA microarrays, it has become natural to think of each macroscopically discernible cell fate as uniquely associated with a distinct gene expression profile; that is, a characteristic transcriptome (or proteome) (Hsiao et al. 2001; Perou et al. 2000). The transcriptomes, although distinct as a whole, of course overlap with respect to individual genes. Those constitutively expressed in all cell states are called *housekeeping genes*, and estimates are that about 10% of the genes in the human genome are housekeeping genes (Eisenberg and Levanon 2003).

How does a genome produce the distinct stable gene expression profiles that define a given cell fate, such as a cell type? The most primitive form of differentiation into various cell types encountered in evolution is the differentiation into *somatic* cells and *germ line* cells in simple metazoa. Using the roundworm *ascaris* as a model for differentiation, T. Boveri found in 1910 that the somatic cells lose a portion of their genome, as evidenced by what he called “chromatin diminution” (Muller et al. 1996). Of course, this is not the case for the germ-line cells, which have to maintain the entire genomic information and pass it on to the next generation. Thus, one could envision a mechanism of cell differentiation in which specialized somatic cells keep only the genes needed to exert their function, and lose DNA containing the genes not needed (e.g., hemoglobin genes in nerve cells), whereas the germ-line cells would contain the entire genome. It turned out that the *ascaris* mechanism represents rather an exception and is not seen in most organisms, not even in many other worms, such as the model animal *C. elegans*.

We have taken for granted that (at least in mammals) all healthy somatic cells in the body contain the same entire genomic information (with notable exceptions in the immune system). All genes are present, so to speak, as dormant instructions in all cells. Thus, the cell-fate-specific transcriptomes are established and stabilized purely epigenetically (i.e., by the regulation of expression of genes from the intact genomic DNA). Sometimes, this is referred to loosely as a “genetic program.”

The human genome contains roughly $N = 25,000$ genes (according to the latest estimate). For the sake of simplicity, assume that each gene can be either *expressed* (i.e., the protein it encodes is present and active in the cell) or *not expressed* (repressed). Thus, let’s symbolize each gene as a *bit*, g_i , where $i = 1, 2, \dots, N$. The variable g_i can take the values $g_i = 1$ (gene i is turned ON = expressed) and $g_i = 0$ (gene i is OFF = repressed). Each state $S(t) = (g_1(t), g_2(t) \dots g_N(t))$ at time t , which can be written as a string of length N —for example, {1010110 . . . }—would then represent a genome-wide gene expression profile. With $N = 25,000$ binary genes we would have $2^{25000} \approx 10^{7526}$ possible configurations of strings, or genome-wide profiles.

The entirety of these possible combinatorial gene activation configurations across N genes constitutes the N -dimensional gene expression *state space*—each gene spanning one dimension. Remember that we have made a simplifying assumption of ON/OFF genes and are thus on the conservative side. Each one of these gene expression profiles S is a unique configuration, one point in the N -dimensional state space, and could theoretically represent a phenotypic cell state. Despite our simplification—which omits multilevel activation, splicing, post-

transcriptional and post-translational modifications, metabolites, and so on—the number of 10^{7526} possible gene expression profiles is hyperastronomic. (Compare: there are approximately 10^{80} protons in the universe, and there are approximately 10^{17} cells in our body).

In other words, we would have almost an endless quasi-continuum of cell phenotypes if all of these gene activation configurations would be realizable. Many cell types would be very similar to each other. For instance, the one specified by the string {00000...001} would be almost undistinguishable from the one specified by {00000...011}. In addition, the activation of a gene (including transcription, translation, and protein activation) is subjected to molecular noise that is manifest as random fluctuations in its activity, hence further contributing to “smearing out” the genetic programs to an unfathomable continuum.

III. CELL FATES AND CELL TYPES AS DYNAMIC ENTITIES IN MULTICELLULAR ORGANISMS

The reality is that we do not observe a quasi-continuum of cell phenotypes but distinct and almost discrete cell fates that are in general mutually exclusive to each other. Hence, there also cannot be a continuum of gene expression profiles in the huge gene expression state space. In other words, despite sharing the identical genome cell types have their own type identities, which are separated by some sort of barrier. The gene expression profiles characteristic of the various cell fates do not easily morph between each other, much as cell types do not *ad libitum* or randomly differentiate into each other. In textbooks, the number of cell types in the human body that are “plainly distinguishable” and identified under one single name is given as “more than 200” (Alberts et al. 2004).

More detailed analysis based on gene expression profiles using DNA microarray technology now reveals that this is an underestimate. If a cell type is characterized by its distinct gene expression profile (including post-translational modifications), many nominal cell types, as defined by traditional histology, actually encompass a set of multiple molecularly distinguishable subtypes. This is well documented for cell types that appear in different regions of the body, such as fibroblasts and endothelial cells (Chang et al. 2002; Chi et al. 2003). For instance, lung endothelial cells have a gene expression profile that is different from that of endothelial cells in the brain or the intestine. Moreover, each individual cell can occupy distinct functional states; for example, inflammatory (activated states) versus quiescent states. Thus, as long as the ontogenetic hierarchy and the similarity relationships between these functionally distinct cellular states are not fully clarified and there is no formal definition of “cell type,” the term *cell fate* is preferred for general purposes, designating a dynamic entity that can be defined molecularly based on the unique gene activity profile.

The characteristic dynamics of cell fate regulation and the lack of continuous morphing between gene expression profiles leads to the restricted rule-governed behavior consisting of conditional cell fate switching. In other words, from this

higher-level perspective development and homeostasis occur according to a defined *cell fate transition map*. Each cell fate can only switch to a defined, relatively small, set of other accessible fates, and some fate transitions are reversible whereas others are irreversible. Branching and sequential cell fate map patterns can be observed. For instance, a multipotent stem cell or progenitor cell in the adult tissue makes an all-or-none decision as to whether to enter the cell division cycle (self-renewal) or to commit to one of a few accessible cell lineages leading to a particular differentiated cell type via discrete stages.

IV. LIMITATIONS OF MOLECULAR PATHWAYS AS EXPLANATION OF CELL FATE BEHAVIOR

With the previously presented picture of the actual dynamics of cell fate behavior in mind, we can now formulate our question in more concrete ways. Where does the discreteness of cell fates, their mutual exclusivity, the discontinuity of fate transitions, and the restricted choice of alternative fates come from? And why do cell types, despite sharing the same genome, in general represent stable entities and do not gradually “drift away” and “morph” into one another but instead get “stuck” in precisely those expression profiles that represent the observable cell fates? What is the molecular basis of the rules that govern such cell fate dynamics?

These “emergent” or “system” properties cannot be easily explained by individual signaling pathways. Attempts to explain cell fates and cell-fate-specific gene expression in the framework of traditional gene-centered regulation rest largely on *ad hoc* chains of causation that are embodied by molecular pathways. Because of the lack of appreciation of the distinct dynamics of cell fates discussed above, there was no need for an encompassing self-consistent formal explanation. Instead, the molecular biology explanation for why a cell expresses liver-specific genes is that there must be liver-specific transcription factors that activate the transcription of those genes that are only expressed in liver, and that hence contribute to the establishment of the liver-specific expression “program” (Odom et al. 2004). Although “master genes” (such as MyoD, PPAR, PU.1, GATA, and so on) have been found that when ectopically expressed can induce an entire or partial cell type program, this of course only shifts the explanation one step further up the chain of causation and is not a “closed” explanation in a strict sense. What regulates the (tissue-specific) regulators? Moreover, the often-used metaphor “gene expression program” lacks a formal definition.

Similarly, the stability of gene expression associated with an enduring cell fate (i.e., the “memory” of differentiated cells of their lineage identity) is typically explained by the regulation at the level of chromatin modification, which is mediated by covalent changes of histones (the protein component of chromatin)—including methylation and acetylation—and by direct methylation of DNA (Georgopoulos 2002; Khorasanizadeh 2004; Arney and Fisher 2004). These changes

affect accessibility of DNA for the transcriptional machinery. These processes are commonly referred to by molecular biologists as epigenetic mechanism *sensu strictiore*. However, it is important to note here that epigenetics is a much broader concept, within which the covalent histone and DNA modifications only constitute a subset (Jablonka and Lamb 2002). All regulation of stable gene expression programs that do not involve changes in the DNA sequence, as discussed earlier in the case of *ascaris*, can be referred to as epigenetic.

Importantly, maintenance of stable cell state and associated gene expression need not and cannot solely depend on the covalent changes that enjoy the intuitive attribute of stability. First, if a gene is sustainably turned ON or OFF by methylation and/or acetylation, again the question is simply shifted; namely, to the one about what controls the activity of the responsible enzymes (acetylases and methylases) and directs them to the appropriate gene loci. Second, the picture is emerging that histone modification is a dynamic process. In fact, it has long been known that histone acetyltransferases enzymes are counterbalanced by histone deacetylases, and there is increasing evidence that this is also the case for histone methylation (Kubicek and Jenuwein 2004). Moreover, gene expression dynamics appears now to be less dependent on large-scale chromatin packaging than previously thought (Gilbert et al. 2004). Thus, epigenetic chromatin modification only superficially, but not in principle, explains the persistence of stable gene expression profiles. What is needed is a self-consistent comprehensive “closed-loop” explanation for the existence of stable and discrete expression profiles in a dynamic system.

V. CELL FATE DYNAMICS: CONSTRAINED BY THE GENE REGULATORY NETWORK

An obvious reason for the lack of consistent “intermediate types,” as Waddington called them, that would correspond to intermediate gene expression profiles in a continuous expression state space is that the expression of individual genes is not independent of each other. Therefore, not all of the 10^{7526} gene expression configurations in our binary gene model are realizable, but only a tiny subset of it. For instance, if *Gene 1* unconditionally inhibits *Gene 2* then all the gene activation configurations S in which both *Gene 1* and *Gene 2* are active—that is, $S = (1, 1, \dots, g_N)$ —would be logically unstable, and would (driven by the regulatory interactions) be forced to move into a “neighboring state” that complies with this regulatory constraint (e.g., $S = (1, 0, \dots, g_N)$). In other words, not all gene expression profiles S are equally stable, because the network of mutual influences of gene expression imposes constraints on the collective dynamics of gene expression.

The question now is how the particular architecture of the regulatory network of a large number N of genes can give rise to precisely the type of constraints of the dynamics of the gene expression profile so that the architecture governs cell fate dynamics and produces the macroscopic system behavior of cells that we observe. More generally, whether a large system of interacting elements can exhibit coher-

ent globally ordered (stable) behavior or becomes disordered and chaotic is an old question in the physics of dynamic systems. From the study of relatively small networks as systems of differential equations, it has been shown that in general sparsely connected networks (below a certain connectivity) are stable, but at higher connectivity they are dynamically not stable (i.e., they do not produce globally ordered dynamics) (Gardner and Ashby 1970; May 1972; Meyer and Brown 1998).

If the architecture of the interaction network meets some specific criteria, as discussed further in Section XI, ordered behavior with globally stable patterns of gene expression will arise. To better understand this, we will review the model class of Boolean networks (Section VII), which was introduced by Kauffman precisely to study whether a complex network can give rise to ordered behavior. “Complex” denotes here that the network is large (N in the thousands) and that gene-gene interactions are apparently irregularly wired (i.e., do not form a regular lattice) (Strogatz 2001).

Before discussing complex networks using the model of Boolean networks we will, honoring both history and didactical principles, begin with a 2-gene toy system. We will first discuss “traditional” continuous-variable modeling using differential equations based on the formalism of chemical kinetics and show how discrete system states, corresponding to cell fates, can arise in the two-gene circuit given appropriate system structure. We will then move on to illustrate the Boolean idealization using this very same two-gene example before discussing complex networks.

VI. MULTISTABILITY IN A SMALL GENE CIRCUIT

If a discrete cell fate is defined by the activation configuration of a set of genes, and all cells harbor the same set of genes, realization of a cell fate will require the general ability of a system of interacting elements (the gene regulatory network) to display multiple alternative, discrete stable states (defined by a gene activation configuration). The existence of two or more steady states within one system is called bi-stability or multistability, respectively. Delbrück proposed in 1948 bi-stability as a general principle to explain how discontinuous transitions between two stable states arise in biochemical reaction systems (Delbrück 1949) and could explain differentiation—at about the same time Waddington promoted the picture of the epigenetic landscape in development.

Novick and Weiner first showed in *E. coli* the existence of all-or-none transitions between cell states with respect to lactose metabolism, and that such states can be maintained across generations in the absence of the chemical inducer and genomic mutation (Novick and Weiner 1957). Such endurance of a non-genetical trait (i.e., in the absence of mutation) is at the very core of differentiation into multiple cell types in multicellular organisms (Rubin 1990). Monod and Jacob proposed a gene regulation circuit exhibiting bi-stability to explain differentiation (Monod and Jacob 1961). Thomas showed that the minimal element in a system required

for multistability and switch-like behavior is a positive feedback loop (Cinquin and Demongeot 2002; Thomas 1978), as epitomized by the classical widely studied one-gene system that exerts positive autoregulation (autocatalytic activation of gene expression) and is inactivated with first-order kinetics (Laurent and Kellershohn 1999). The very notion of multistability has been sidelined in mainstream molecular biology until recently.

Here we present a slightly extended system as a pedagogical tool to illustrate the basic principle, motivated by recent findings from the study of genes that govern cell fate decisions with respect to lineage commitment in mammalian hematopoietic progenitor cells. We discuss how a simple continuous value model based on nonlinear differential equations can capture the essential macroscopic dynamics of cell fate transitions of a progenitor cell C that can differentiate into the two cell types A and B .

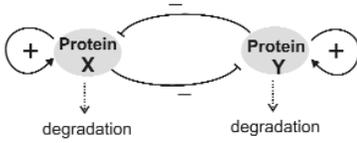
Figure 14.2 shows the circuitry consisting of two genes, X and Y , which exhibit a basal activity but inhibit each other. They also activate their own transcription. Each gene (more precisely, their encoded protein) is also subjected to a first-order kinetic inactivation (degradation). Their activation state (e.g., expression level of the active protein) can change continuously over time.

An essential point in treating such gene regulatory systems as depicted in Figure 14.2 is to model a regulatory influence on the change of activity (dX/dt or dY/dt , by X or Y , respectively) as a sigmoidal input/output (stimulus/response) relationship. Such sigmoidal “transfer functions” reflect *sensitivity amplification* leading to ultrasensitivity (Koshland et al. 1982). It is justified, even in the absence of cooperativity (the best known cause of such sigmoidality), on grounds of the particular physicochemical conditions of the intracellular milieu, which departs from that of ideal well-stirred macroscopic solutions. These conditions include the presence of molecular noise (stochastic focusing), crowded environment, and reaction on surfaces (Paulsson et al. 2000; Savageau 1995; reviewed in Huang 2001).

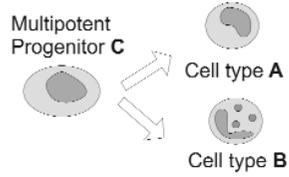
In fact, experimental measurements of many biochemical signaling events suggest that steep sigmoidal transfer functions with a threshold, resulting in an approximately all-or-none response, are ubiquitous. The apparently smooth behavior of variables, evident from the biochemical analysis of bulk cell cultures and the primary motivation for using continuous models, is in great part due to the averaging over asynchronous and noisy cell populations in which gene activity in individual cells behaves in a discontinuous manner (Figure 14.3). Nevertheless, low-dimensional 1- or 2-gene systems exhibiting multistability described by a set of differential equations that assume sigmoidal regulation characteristics have recently gained much interest, both at the theoretical and experimental level (Cherry and Adler 2000; Gardner et al. 2000; Becskei et al. 2001; Bhalla et al. 2002; Sha et al. 2003; Tyson et al. 2003; Xiong and Ferrell 2003; Angeli et al. 2004; Ozbudak et al. 2004).

Here we discuss the particular system with the two genes, X and Y (as shown in Figure 14.2, left-hand column). Such or similar circuitry constellations are widely seen in the genetic pathways underlying control of cell differentiation (Zingg et al.

REGULATORY NETWORK



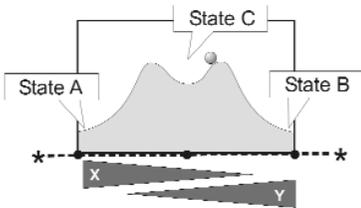
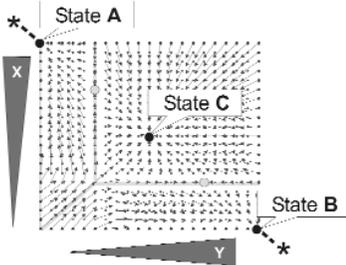
MACROSCOPIC CELL BEHAVIOR



CONTINUOUS MODEL

$$\frac{dX}{dt} = k_1 \frac{S^n}{S^n + Y^n} + k_2 \frac{X^n}{S^n + X^n} - k_4 X$$

$$\frac{dY}{dt} = k_1 \frac{S^n}{S^n + X^n} + k_5 \frac{Y^n}{S^n + Y^n} - k_6 Y$$



DISCRETE MODEL

$$X(t+1) = B_x[S(t)] \quad S(t) = \{X(t), Y(t)\}$$

$$Y(t+1) = B_y[S(t)]$$

B_x :		B_y :	
INPUT	OUTPUT	INPUT	OUTPUT
$\{X(t), Y(t)\}$	$\rightarrow X(t+1)$	$\{X(t), Y(t)\}$	$\rightarrow Y(t+1)$
{0, 0}	\rightarrow 1	{0, 0}	\rightarrow 1
{0, 1}	\rightarrow 0	{0, 1}	\rightarrow 1
{1, 0}	\rightarrow 1	{1, 0}	\rightarrow 0
{1, 1}	\rightarrow 1	{1, 1}	\rightarrow 1

State transition table

$S(t)$	$\rightarrow S(t+1)$
{0, 0}	\rightarrow {1, 1}
{0, 1}	\rightarrow {0, 1}
{1, 0}	\rightarrow {1, 0}
{1, 1}	\rightarrow {1, 1}

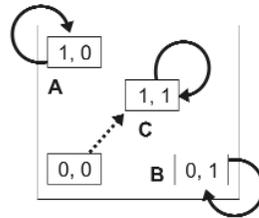


Figure 14.2. Pedagogical example of a simple gene regulatory circuit consisting of two genes/proteins. The genes X and Y inhibit each other and activate themselves (box). The cell fate behavior regulated by this circuit is shown on the right: commitment of a multipotent precursor cell C into two differentiated cells, A and B . Two models of the underlying gene regulatory network are shown.

LEFT COLUMN. Continuous variable model, described by a system of differential equations for the variables $X(t)$ and $Y(t)$ representing the activity of the two genes/proteins. The first and second terms describe mutual inhibition and autoactivation, respectively, which follow a sigmoidal regulatory characteristic captured by a Hill function. The constants S (for simplicity assumed to be identical for all four terms) represent the threshold of the sigmoidal curve, n is the Hill coefficient ($n > 2$), and k_i are rate constants. The last term represents first-order decay. Below, the XY phase plane. Arrows in the vector field indicate movements of states $S(X, Y)$ along their trajectories (enforced by the network interactions) at the given points (X, Y) during a time interval Δt . At the bottom of the figure is shown a hypothetical “potential landscape” along the diagonal (---) through the stable fixed points A , B , and C .

RIGHT COLUMN. Discrete variable (Boolean network) model. This model is quite artificial and only presented to illustrate some limited equivalency to the continuous model on the left. The values of the variables X and Y at discrete time $t + 1$ are Boolean functions, B_X and B_Y , respectively, of the corresponding set of *input genes* at time t , which in this example of an $N = 2$ “network” is equivalent to the entire state of the network, $S(t)$. Below, the two *truth tables* for the Boolean functions, B_X and B_Y , which represent two examples of the function *IMPLICATION*. The entire dynamics can be represented in the *state transition table*, which can be depicted as a *state transition map* (bottom). In this particular case, the basins of attraction consist of only the attractor state itself. Arrows represent the “trajectories” indicating which transition a given state (rectangular box) undergoes when updating the network by executing the Boolean functions. In this example, the Boolean network produces the same set of stable states as the continuous model, which is by far not a general property. Stable states (fixed-point attractors) update into themselves. The state $\{00\}$ is unstable, as in the continuous model. Note that the transition $\{00\}$ to $\{11\}$ (dashed arrow), despite compliance with the Boolean functions, is an artifact of the synchronized updating and goes through unstable regions in the continuous model.



1994; Chen et al. 1995; Nerlov et al. 2000; Zhang et al. 2000; Ohneda and Yamamoto 2002; Grass et al. 2003). The system equations of our circuit are as shown in Figure 14.2.

Note that using the Hill function (= sigmoidal function as used in the equations in Figure 14.2) to capture negative regulation (suppression) assumes a baseline activity of the target gene in the absence of the repressor. The numerical solution of this system of nonlinear differential equations shows that for a wide range of parameter values the system will exhibit three stable fixed points, or *attractor states*. This is illustrated in the X - Y phase plane, the 2D state space that represents the possible states $S(X, Y)$ of the system (Figure 14.2). The two attractor states $A = (X_A, Y_A)$ (with high X and low Y values) and $B = (X_B, Y_B)$ (low X and high Y) correspond to stable states of the system where the activity of either one gene, X or Y , dominates and suppresses the other. Given the mutual repression, this is intuitively plausible.

The states A and B are “attracting” because they are stable. Neighboring states (e.g., with X' and Y' values close to X_A and Y_A) are “attracted” back to state A . A third attractor state C with $X_C \sim Y_C$ represents the configuration in which X and Y are equally active. C is located between A and B in the state space and its *basin of attraction* (the regions in the phase plane in which all system states will end up in the respective attractor state) is bordered by those of the attractors A and B . In

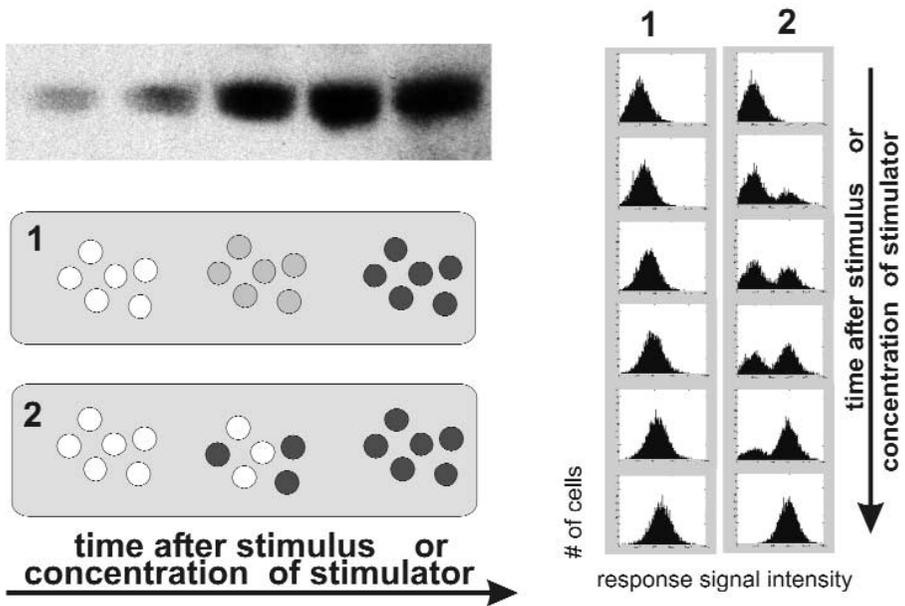


Figure 14.3. Discrete behavior of individual expression of a gene X at the single-cell level. The measurement of protein level (e.g., by immunoblot analysis, top left) for increasing expression (band intensity) in five samples is a population average of millions of cells. Such an increase can arise in principle in two different ways when expression in individual cells is considered. This is schematically illustrated in the gray boxes 1 and 2, where each dot represents one cell and the color intensity the level of X . On the right-hand side, simulated flow cytometry histograms for 10,000 cells are shown, corresponding to the two scenarios 1 and 2. Note the bimodality in the case of 2, and the effect of “noise” that leads to dispersion of expression levels around a mean value.

contrast to these attractors, the existence of attractor C depends on the presence of the auto-regulatory loops. What is interesting is the macroscopic dynamics of the system with three attractors. With the central tenet being that each attractor state corresponds to a stable phenotypic state or a cell fate, it is evident from the state space structure in Figure 14.2 (bottom) that the cell can easily be “kicked out” of state C by a perturbation (i.e., a transient change in the value of X , Y , or both to assume either one of the other attractor states). Thus, C represents the precursor cell that can decide to commit to either the A or B cell fate.

VII. CELL FATES AS ATTRACTORS

At this point we can summarize the dynamic feature of a “network” and its relation to the macroscopic cell fate dynamics. The central idea is that a *cell fate is an attractor state of the dynamic system established by the underlying gene regulatory network*. Attractors are discrete stable states. Intermediary configurations between the attractor states are not stable. Which state a given cell (the network) occupies

depends on the initial condition (position in state space, relative to the basins of attraction). Thus, the state space has a characteristic *substructure* that imposes dynamic constraints onto the global dynamics of the network. It is compartmentalized by the *separatrices* (basin boundaries) into basins of attraction and drives the dynamics of the network from unstable toward stable states. In other words, the biological features of cell fate dynamics presented earlier and captured in Waddington's epigenetic landscape (Figure 14.1b) are quite remarkably reflected in the dynamics of the network.

Attractors of course need not be fixed points. They can be periodic (limit cycles) or even chaotic oscillators. However, in the following we will discuss gene regulatory networks that are high-dimensional systems of $N =$ thousands of variables in which chaos in the classical sense (positive Lyapunov exponents, as typically described for low-dimensional systems) is not as well studied, and the relevance of which for cells is not well known (Bagley and Glass 1996). In contrast, a limit-cycle attractor has been interpreted as the cell division cycle, representing the fate of cell proliferation, a state in which the cell undergoes repeated rounds of cell division by passing through a recurring sequence of biochemical states (Huang 1999; Huang and Ingber 2000; Kauffman 1993)—although many genes are expected to oscillate independently of the cell cycle (see Section X). The chapter by Goldbeter discusses in more detail oscillatory behavior in cell biology.

If cell fates are attractors, a *cell fate switch* (for example, in response to a hormone) that triggers the commitment of a proliferative progenitor cell to differentiate into a particular lineage is a perturbation of the gene regulatory network that causes the system (network) to “jump” to another attractor state. A *perturbation* is formally a transient externally imposed change of the activity values of one or a set of genes so that the state $\mathbf{S}(t)$ of the cell is placed somewhere else in the state space, such as into the basin of attraction of another attractor.

The previously discussed two-gene network reproduces even a more specific behavior of a multipotent progenitor cell that would be represented by cell fate C in Figure 14.2. This cell undergoes a cell fate decision to commit to either one of two cell fates, A or B. In the model, the gene X would be a differentiation marker for the state A, and Y the differentiation marker for B. In fact, hematopoietic cell differentiation—which generates the various blood cells, starting from one multipotent stem cell, the so-called hematopoietic stem cell (HSC) in the bone marrow—occurs in a sequence of multiple “bifurcations” of the differentiation path into two alternative lineages characterized by reciprocal (mutually exclusive) transcription factor activities.

The relative levels of these two factors translate into the all-or-none fate decision. For instance, in the common multipotent precursor cell (CMP) experimental overexpression of transcription factor PU.1 leads to suppression of the transcription factor GATA-1 and to the differentiation of the macrophage/monocyte lineage, whereas overexpression of GATA-1 causes suppression of PU.1 and promotes differentiation into megakaryocyte/erythrocyte precursor cells (Graf 2002). Newer studies that employ the suppression of GATA-1 confirm this integrated dynamic behavior, in that loss of GATA-1 in zebra fish itself was sufficient to push the fate

decision toward the macrophage/monocyte lineage (Galloway et al. 2005). In fact, at the molecular level GATA-1 and PU.1 inhibit each other in a circuitry that corresponds to that between X and Y in Figure 14.2 (Chen et al. 1995; Nerlov et al. 2000; Zhang et al. 2000; Graf 2002).

Interestingly, progenitor cells have been shown to exhibit a promiscuous expression pattern that contains “a bit of everything”; that is, express at low level but simultaneously the genes that define the alternative cell fates and are mutually exclusive in the mature cells (Akashi et al. 2003; Bruno et al. 2004; Enver and Greaves 1998; Hu et al. 1997). This promiscuous “sneak-preview” gene behavior (Graf 2002) is also predicted by the model: the reciprocally behaving “fate-specific marker” genes X and Y are both present in the progenitor C at lower level. Thus, multipotency itself is an attractor state, an entity that arises primarily from the nonlinear dynamics of a network. There is in principle no epistemological need to invoke a “stemness gene” to explain multipotency.

VIII. THE BOOLEAN NETWORK FORMALISM

The small-circuit model is an arbitrarily cut-out fragment of the genome-wide regulatory network, and there is no formal justification yet to treat it as an isolated module. How can we deal with the entire network of 25,000 genes and many more proteins? The lack of detailed knowledge about most interaction modalities, including the parameter values, as well as the rising computational cost when modeling large genome-wide networks has motivated a coarse-graining of the model using discrete-valued networks.

Such discretization is justified because (1) the previously discussed sigmoidal shape of transfer functions in the gene influences can be approximated by a step function (Section VI) and/or (2) the local dynamics produced by such small gene circuit modules is characterized by discontinuous transitions between discrete states (Figure 14.3). An important aspect is that because of computational efficiency discrete networks offer the possibility of studying statistical ensembles of networks (i.e., entire classes of network architectures) and of addressing the question of how a particular architecture maps into particular types of dynamic behavior. Stuart Kauffman championed this network ensemble approach in the 1960s using random Boolean networks, in which gene activity values are binary (1 = ON, and 0 = OFF). (Kauffman 1969). Time is also discretized. Thus, a Boolean network is a generalized form of cellular automata but without the aspect of physical space and the particular neighborhood relations.

Then, as encountered earlier, a network of N elements g_i ($i = 1, 2, \dots, N$), where g_i is the activity state (1 or 0) of gene i , defines a network state \mathbf{S} at any given discrete time step t : $S(t) = (g_1(t), g_2(t), \dots, g_N(t))$ (Figure 14.2, right-hand column). Gene regulation is modeled by the Boolean functions B_i associated with each gene i and map the configuration of the activity states (1 or 0) of its input genes (upstream regulators of gene i) into the new value of g_i for the next time point. Thus, the argu-

INPUT {lac, cAMP}	OUTPUT b-Gal
■ ◆	
{ 0, 0 }	→ 0
{ 0, 1 }	→ 0
{ 1, 0 }	→ 0
{ 1, 1 }	→ 1

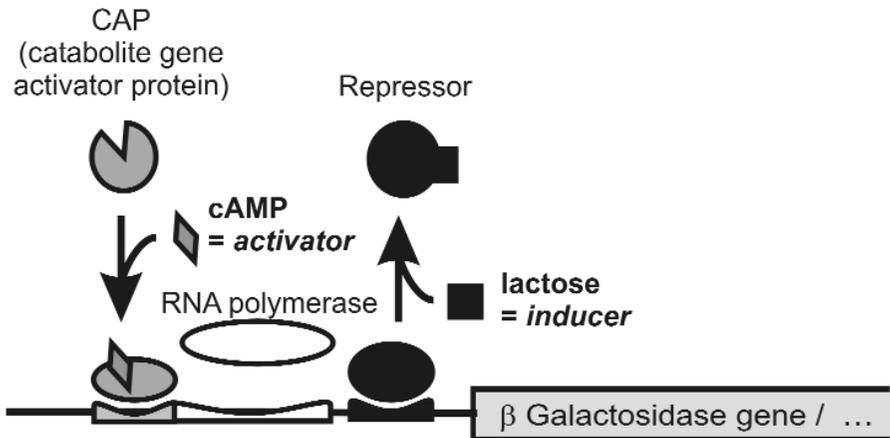


Figure 14.4. Using a Boolean function (*AND*) for capturing a gene regulatory mechanism. This example uses a simplified version of the bacterial *lac operon* controlling the expression of β -galactosidase gene (b-Gal). The gene is turned ON if and only if *cAMP* (which accumulates when cellular glucose is low and activates the transcription factor CAP) and *lactose* (which is the alternative nutrient and deactivates the repressor) are both present. Note that in more complex promoters the temporal order of arrival of the input molecules may play a role. This aspect is ignored in Boolean functions.

ment of the Boolean function B_i is the input vector $\mathbf{R}_i(t) = (g_1(t), g_2(t), \dots, g_{k_i}(t))$, where k_i is the number of inputs the gene i receives. At each time step, the value of each gene is updated: $g_i(t+1) = B_i(\mathbf{R}_i(t))$. The logical functions B_i can be depicted as a “truth table” convenient for large k_s (Figure 14.2, right).

Figure 14.4 shows the example of the well-studied *lac operon* and how its regulatory characteristics can be captured as an *AND* Boolean function for the two inputs. One advantage of using Boolean functions is that they capture the qualitative regulatory rules obtained from biological experiments (such as *AND* and *OR* functions in “promoter logics”) much more efficiently than a set of chemical reactions. To illustrate the basic elements of a Boolean network, we now translate the

two-gene example used previously into a Boolean network formalism (Figure 14.2, left versus right). In the small network of $N = 2$, each gene has $k = 2$ inputs. (The result may appear a bit caricatural, in that the strength of the discrete network model is the study of large networks, but this example shall serve as an illustration of the principles.)

Most investigators have studied the simple case of synchronous homogenous networks, where all the N genes have a fixed number of k inputs and are updated in every time step. The synchrony of updating in particular is artificial. However, it appears that essential features of the global dynamics that are relevant for cell biology are preserved in such synchronously updating networks. In the widely studied case in which each gene has $k = 2$ inputs the Boolean function can be one of the set of $2^{(2 \times k)} = 16$ classical Boolean operators, such as *AND*, *OR*, *NOT*, *IF*, and so on.

Alternatively, for synchronous networks the entire network state \mathbf{S} can be viewed as being updated by the updating function U : $\mathbf{S}(t + 1) = U(\mathbf{S}(t))$, where U summarizes all the N Boolean functions B_i and is represented in the *state transition table*. (In the example of Figure 14.2, because $N = k$, U follows directly from the set of Boolean functions B_i .) The state transition table can be depicted as a *state transition map*, where each state is represented as a box (showing the gene expression profile). Such diagrams were proposed by Wuensche and are particularly illustrative for N up to 10, in that they display all states of the state space (Wuensche 1998). The states are connected by the arrows, which represent individual state transitions and collectively depict the trajectories in the state space (Figure 14.2, bottom right).

Because the Boolean functions are deterministic, a state $\mathbf{S}(t)$ unambiguously transitions into its successor state $\mathbf{S}(t + 1)$. In contrast, a state can have multiple predecessors, in that two or more different states can converge into the same state. Hence, the movement of the state vector in the N dimensional state space defines trajectories that can converge but not diverge. This property of “losing information about the ancestry” is essential to the robustness of the dynamics of networks.

In the example of Figure 14.2, the Boolean functions used to capture the qualitative descriptions of the interactions in this case represent the function *IMPLICATION* for both genes. The resulting dynamics is equivalent to that explained for the continuous model. There are three stable states (attractors): $A = \{10\}$, $B = \{01\}$, and $C = \{11\}$. The state $\{00\}$ flows into the attractor state $\{11\}$, whereas the other two states— $\{10\}$ and $\{01\}$ —update onto themselves.

This example is somewhat artificial because N is small and $N = k$. Glass and colleagues have compared a continuous variable model of networks with the Boolean formalism of discrete networks and investigated the equivalency between these two models (Bagley and Glass 1996; Glass and Kauffman 1973). Overall, the global qualitative dynamics of continuous variable networks and discrete variable networks are in principle similar. However, there are specific differences. Notably due to the synchronized updating scheme in classical Boolean networks, for instance, not all cycling attractors found in the discrete networks have their equivalent in the continuous model. Specifically, to suppress artifacts stemming from synchrony one

would allow only for state transitions in the state transition map in which at every step only one gene changes its value (see legend of Figure 14.2 for details). The general idea that emerges from both of the previously cited models, however, is that the interactions between the network elements introduce constraints to the global dynamics and given a certain class of network architecture can produce ordered system behavior.

IX. SMALL CIRCUITS VERSUS GENOME-WIDE NETWORKS: DIFFERENT PHILOSOPHIES

The philosophy behind modeling *entire networks* using a simple model is fundamentally different from most current attempts to model a *local circuit* (pathway) in accurate detail. Such circuits are implicitly assumed to represent a functional “module” cut out from the genome-wide network. Hence, such circuits ignore many inputs. One focus then is on the detailed kinetics of the molecular interactions, typically formalized on the basis of mass action kinetics, or using even finer-grained models that capture stochastic behavior of individual models. One goal is the simulation of reality as faithfully as possible.

In contrast, the simplification achieved by using the model class of discrete networks, such as Boolean networks, accepts the trade-off of abstracting away local biochemical details for embracing an entire closed system. The ambition of whole-network models is to “see the forest, not just the trees” and to study some fundamental aspects of inherent entirety of a network, in the true spirit of systems biology. This coarse-graining is sometimes necessary because as many scientists and philosophers have articulated in diverse ways there is no understanding without simplification (Picht 1969). The challenge is to choose the appropriate simplification and abstraction and to not forget them when interpreting the results.

The comprehensive set of variables g_i also implies that in the idealized case all influences are covered by the N variables of the system such that physiological perturbations (i.e., externally imposed changes onto the system) are embodied by the change in the values of one or a set of the variables and hence reset the initial state S to another initial state in state space. This is different from the low-dimensional models describing local pathway modules. There, one often tends to view real-time external influences as a tuning of *control parameters* that would change some properties of the low-dimensional state space (e.g., shift of attractor boundaries or sudden disappearance of attractors).

For the Boolean network ensemble approach, such variation of control parameters (such as reaction rate constants) can be thought of as a response to other components of the system or to external perturbations and hence are captured by the changes in the value of the variables. In other words, in this comprehensive view a particular genomic network architecture maps onto a fixed state space “landscape.” Thus, the substructure of the state space is hard-wired in the genome and provides a stage on which developmental and homeostatic processes of the cell

and organism take place. Changes in conditions affect the initial states (i.e., displace the state $S(t)$ within the state space) but do not affect the structure of the landscape shape. Conversely, the tunable control parameters in the study of large networks affect the selection of the subclass of architectures in the space of all possible network architectures (see Section XI).

The actual strength of the Boolean network formalism is the study of the generic qualitative dynamics of classes of network architectures using statistical ensembles of randomly generated networks, whose architecture can be constrained in a controllable way. This approach underlies a philosophy that is diametrically different from that embraced by most current biologists. Thus, it shall be in order to briefly discuss it at this point so that readers of this book who may come from a variety of backgrounds see these two approaches from a broader perspective. We distinguish here between two opposite mind-sets in systems biology (Huang 2004).

(A) *Particularists*: Most of post-genomics and now systems biology is essentially an extension of the old *modus operandi* of classical biology, often ridiculed as “stamp collecting”: the description of a system by enumerating its parts and describing their properties and relationships in accurate detail, be it in a qualitative or quantitative manner. Such a descriptive collecting approach, rooted in the tradition of zoology and botany, has found its counterpart in molecular biology and is most prosaically epitomized in the cloning, characterization, categorizing, and classifying of individual proteins.

High-throughput discovery and mathematical modeling of molecular pathways in current systems biology represent a quantitative extension of this approach but not a qualitative departure (Bray 2003; Endy and Brent 2001). The common goal of these approaches has been to understand a particular (idiosyncratic) instance of a system in as precise terms as possible, using both qualitative description and quantitative “predictive” models. Often, existing principles from the field of engineering are applied to solve problems. The question of generalizability is typically not explicitly asked, taken for granted, or postponed to a later time point.

(B) *Universalists*: An entirely different mind-set of investigation, perhaps more prevalent among theoretical physicists, has as its primary aim (if not *raison d'être*) the discovery of universal properties, and ideally of new universality classes (Bar-Yam 2000). Here, the description of the particulars, which may obscure general principles, gives way to the search for the universal rule. Thus, abstraction and simplification are common methods. Universalists typically do not bother about results that represent new instances but not new principles.

Because of the intrinsic heterogeneity of biological systems and the nature of biology as a historical science, the two approaches—the analysis of specific instances (A) and the quest for general principles (B)—complement each other and are equally necessary. The study of molecular networks in biology must be seen in the light of this dualism of approaches (Huang 2004). Hence, for the particularists (currently the vast majority of systems biologists), when it comes to characterizing gene regulatory networks the most natural ultimate goal is to reverse engineer the network architecture for a particular case; that is, to solve the inverse problem

based on experimental observations of the gene expression dynamics using a variety of algorithms (Liang et al. 1998; Akutsu et al. 1999; D’Haeseleer et al. 2000; Friedman et al. 2000; Yeung et al. 2002; Ehrenberg et al. 2003; Gardner et al. 2003).

In performing this task, the choice of the model class (e.g., linear versus nonlinear, continuous versus discrete, deterministic versus probabilistic, and so on) will affect the formidability of the inverse problem. In contrast, for the universalist approach the choice of the model will affect simulation costs and ideally must not affect the conclusions for biology.

X. DYNAMICS OF LARGE NETWORKS AND THE ENSEMBLE APPROACH

The simplicity and tractability of the Boolean network formalism has triggered wide investigations among statistical physicists and theoretical biologists who have studied the dynamics of large generic networks with the mind-set of universalists— independent of the progress in genomics. These efforts have produced interesting results—even before the explicit incipience of “systems biology” (Kauffman 2004). If the particular details of the instance of real networks can be abstracted away appropriately (so goes the idea), one can study the generic dynamics by analyzing representative samples from the ensemble of “all possible” network architectures defined, for instance, by the previously cited formalism of Boolean networks.

In the case of Boolean networks, a given instance of a network *architecture* A consists of the *topology* of the wiring diagram (the structure of the network displayed as a directed graph of N nodes) and the set of N Boolean functions B_i assigned to each node $i = 1, \dots, N$ of the graph. An often-studied set of architectures comprises all networks in which all nodes have a fixed number k_i of inputs and one of the $2^{\binom{k_i}{k}}$ Boolean functions. The number G of such networks is large:

$$G = \left[\binom{N}{k} \cdot (2^{2^k}) \right]^N$$

When studying the fundamental question of how an architecture maps into a dynamic behavior (structure of the state space) by randomly sampling in a given architecture space, it is important to note that multiple different architectures A can map into the one same dynamic behavior; that is, have the same state space structure (but not vice versa) if all possible Boolean functions (truth tables) are allowed. This redundancy of architectures stems from the fact that in a subset of Boolean functions some regulatory inputs have no influence on the output, so that the number of effective inputs $k_{i,eff}$ of a node i is not always equal to the “nominal” input $k_{i,nom}$ defined by the network topology: $k_{i,eff} \leq k_{i,nom}$. The number of effective dynamic structures H for the class of fixed- k input networks is given by (Myers 2001)

$$H = \left[\binom{N}{k} \cdot (N-k) \cdot \left\{ \sum_{m=0}^k (-1)^{k-m} \cdot \binom{k}{m} \cdot \frac{(2^{2^m})}{N-m} \right\} \right]^N$$

The ratio $M = G/H > 1$ is called *multiplicity* and is a function of N and k . M can become quite large. It is important for ensemble studies, notably of network robustness to rewiring, that one operate in (N, k) regimes with $M \sim 1$.

Using Boolean networks and the ensemble approach, Kauffman determined the architectural parameters that influence the global long-term behavior of complex networks with N up to 100,000 (Kauffman 1993). The most striking result from the ensemble studies is that for a reasonably broad class of architectures even a complex irregular (randomly wired) network can produce ordered dynamics with globally “coherent” patterns, such as convergence to stable states, as discussed in material following.

The global behavior of Boolean networks can be divided into three broad regimes: ordered, chaotic, and critical (Kauffman 1993). Networks in the *ordered* regime, when placed into any random initial state in the state space, will quickly settle down in one of the fixed-point attractors or limit cycle attractors that have a small period T compared to N and thus produce macroscopically stable behaviors. They also have a small number of attractors (with large basins). In contrast, networks in the *chaotic* regime will apparently “wander” aimlessly in state space. Because the latter is discrete and finite, the network will eventually encounter a previously occupied state and repeat its trajectory. However, the period T of this limit cycle is very long, and can be in the order of 2^N .

Given the astronomic size of this number, this “limit cycle” will appear as an aperiodic and endless trajectory (permanent transient). Thus, networks in the chaotic regime are not stable, and their behavior is sensitive to the initial state. This definition of chaos is distinct from that of (deterministic) chaos in continuous systems, where the time evolution of infinitesimally closed initial states can be monitored. Nevertheless, the degree of chaos in discrete networks is well defined and can be quantified based on the slope of the curve in the so-called Derrida plot, which assesses how a large number of pairs of initial states evolves in one time step (Derrida and Pomeau 1986). A network in the *critical* regime is at the *edge between order and chaos*. More recently, it was shown that it is possible to determine the behavior class from the architecture, without simulation and determining the Derrida plot, by calculating the expected *average sensitivity* of all Boolean functions (Shmulevich and Kauffman 2004).

One remarkable result of Kauffman’s early studies is that for a large class of architectures (specified following) ordered behavior (i.e., higher-order pattern at the scale of the *entire* network dynamics) emerges, such as compact attractor states with large basins of attraction. If cell fates correspond to robust attractors, genomic networks of real cells are expected to be in the ordered regime that would contain such attractor states. Thus, we require that the large $N = 25,000$ gene network exhibits relatively few ($<N$) attractors that are either limit cycles with small period or fixed-point attractors, and that they have large basins. Such attractors would be *high-dimensional attractors* (i.e., cell fate would be robust with respect to a large number of state space dimensions). As discussed in Section XII, this appears to be the case in real mammalian cells.

XI. ARCHITECTURAL FEATURES OF LARGE NETWORKS AND THEIR DYNAMICS

The architecture of the Kauffman Boolean networks appears as a caricature of real networks, but they can teach us a great deal about the generic dynamics of the class of systems comprised of a complex network of interacting elements such as the gene regulatory network. In an independent development, the recent availability of data of real networks in biology and technology has triggered the study of complex irregular network topologies as static graphs, stimulated in part by earlier studies of social networks (Amaral et al. 2000; Barabasi and Albert 1999; Watts and Strogatz 1998). These fields of study are now beginning to merge. In the following we summarize some of the interesting architecture features and their significance for ordered global dynamics.

(1) *The average connectivity per node k* : Initial studies on Boolean networks by Kauffman assumed a homogenous distribution of k . It was found that $k = 2$ networks are in the ordered/critical regime (Kauffman 1993). Above a critical k_c value (which depends on other parameters, as described following) networks behave chaotically. Analysis of continuous linearized models also suggests that in general sparsity of connections is more likely to promote ordered dynamics, or stability (May 1972).

(2) *The distribution of k_i over the individual network genes, i* : Recent work on the topology of complex networks in general and of molecular networks in particular has revealed that many "evolved" networks (i.e., where N grew over time by addition of new elements and connections, such as the protein-to-protein interaction networks) appear to have a connectivity distribution that approximates a power law (Barabasi and Albert 1999). Such networks have no characteristic average value of k (sampling of larger number of nodes N will lead to larger "average" k values) and are hence said to be "scale-free" (see following). Analysis of scale-free Boolean networks suggests that this property favors a behavior in the ordered regime for a given value of the parameters k and p (see Section XIII) (Aldana and Cluzel 2003; Fox and Hill 2001).

(3) The nature of Boolean functions is also an important aspect of the network architecture that can influence global dynamics. In his early works, Kauffman characterized Boolean functions with respect to the following two important features.

(a) *Internal homogeneity p* : The parameter p ($0.5 \leq p \leq 1$) is the proportion of either 1s or 0s in the output column of the Boolean function. Thus, a function with $p = 0.5$ has equal numbers of 1s and 0s. Boolean functions with p -values close to 1 or 0 are said to exhibit high internal homogeneity.

(b) *Canalizing function*: A Boolean function B_i of gene i is said to be canalizing if at least one of its inputs has one value (1 or 0) that imposes one value onto the output of gene i , independently of the values of the other components of the input vector. If an input determines *both* output values (a "fully canalizing" function), the other inputs have no influence on the output at all, and $k_{i,eff}$ is reduced. For instance, for Boolean functions with $k = 2$ only two of the 16 possible functions, XOR and

XNOR, are not-canalizing. Four functions are fully canalizing (i.e., are effectively $k = 1$ functions). Both a high internal homogeneity ρ and a high proportion of canalizing functions contribute to ordered behavior (Kauffman 1993).

(4) There are many global and local topological features of complex networks defined in graph theory terms that appear to be interesting with regard to the behavioral regime of the network dynamics because they were found to be enriched in genome-wide molecular networks when compared to a set of randomly generated null-hypothesis networks that were constrained to exhibit more elementary topological features. These features include the following.

- Proportion of genes with high “betweenness” value
- Small-worldness, cliquishness, modularity, and hierarchy of network (Amaral et al. 2000; Ravasz et al. 2002; Watts and Strogatz 1998)
- Frequency and distribution of local network motives that are enriched in real networks, such as feed-forward loops in bacteria (Milo et al. 2004; Shen-Orr et al. 2002)

The influence of these topology features on the global dynamic behavior remains to be studied.

XII. REALITY CHECK: GENOME-SCALE NETWORK TOPOLOGY

The most salient question to be clarified before studying complex networks and their global dynamics that has been taken for granted is whether the 25,000 genes of the human genome indeed form one *connected* network of regulatory influences at all or whether it is broken down into f independent network fractions of size Q_i . This question cannot be answered until the architecture of the gene regulatory network for mammals is available. However, incomplete data from the gene regulatory network in *E. coli*, and partial data from yeast, suggest the existence of a giant connected component (i.e., a largest connected set of genes that covers substantial portions of the genome) (Lee et al. 2002; Salgado et al. 2001).

Moreover, coarse but genome-wide network topologies based on information from undirected pair-wise protein-to-protein interactions have recently been available for baker’s yeast (*S. cerevisiae*), the roundworm (*C. elegans*), and the fruit fly (*D. melanogaster*) at the resolution of a nondirected graph, with some error rate (notably false positive interactions). Because gene regulation is mediated by proteins, and a protein-to-protein interaction will also indirectly connect two genes, such protein interaction data can be used to obtain a lower bound approximation for functional connectivity between genes across the genome. These network topologies suggest that for yeast, worm, and fly a giant component that covers approximately 90% of the proteome does exist, thus justifying the previously discussed studies of the dynamics of complex networks (Giot et al. 2003; Li et al. 2004; Mewes et al. 2002).

This is in agreement with theoretical studies of “phase transitions” of the size of the giant component in evolving networks that suggest that if the probability of a newly added gene to be connected to the network is higher than the critical value $c = 1/8$ —which is lower than the average connectivity $k \sim 2-4$ (from the previously cited protein interaction network data)—a phase transition occurs with respect to the formation of a giant component (Callaway et al. 2001). Given this result, if the genome in fact consisted of multiple independent modules there would have to have been a strong evolutionary pressure to maintain these independent modules. Then one could imagine a definition of global cell phenotype by a combinatorial code of the set of attractor states in each of the network fragments that will have independent dynamics.

Nevertheless, despite the large size of the giant component the effective regulatory network whose dynamics arises from the mutual regulation of genes may still be smaller than the genome-scale network, in that many genes have only inputs but no outputs. Such peripheral “effector” genes do not drive the dynamics but may be markers of cell fates that exert particular nonregulatory tissue-specific functions. Taking such genes into account, the genome-wide network would have a “medusa” architecture (Kauffman 2004; Lee et al. 2002) with a small core of bidirectionally interconnected genes and a number of “downstream” tentacles that have no influence on the dynamics. Future analysis of gene network architectures will reveal the proportion of the core regulatory genes.

The topology data—at least for protein interaction networks, but perhaps also for gene regulatory networks—suggest that biomolecular networks approximate a scale-free network architecture, or at least have a broad scale with an overrepresentation of extremely highly connected genes (Amaral et al. 2000). However, the relevance of scale-freeness per se is not at all clear. It may be a trivial manifestation of fundamental statistics. Nevertheless, any natural property (evolved or statistically unavoidable) has dynamic consequences, and as mentioned previously the effect on the dynamics of scale-freeness as opposed to purely randomly wired networks has been studied and found to increase the parameter regime for ordered behavior (Aldana and Cluzel 2003).

As for the use of Boolean functions, Harris et al. (2002) have found an enrichment for canalizing functions in a set of 150 experimentally verified regulatory mechanisms of well-studied gene promoters—again, in accordance with the architecture criteria associated with ordered global dynamics (Harris et al. 2002). Similarly, when canalizing functions were randomly imposed onto the published yeast protein-to-protein interaction network topology (to create an architecture whose dynamics was then simulated), ordered behavior was observed (Kauffman et al. 2003). These studies of global network dynamics are incomplete and at best sketchy due to the poverty of the data of gene network architectures and the simplicity of Boolean network models. However, they represent a first step in a new direction of research (given the now recognized necessity for “integration”) and support the idea that genome-wide gene regulatory networks act as entities with ordered global dynamics with high-dimensional attractors that mirror cell fate behavior.

An interesting question is how the network architecture has evolved so as to produce a well-behaved global dynamics. Is natural selection for ordered behavior strong enough to explain the architecture of genomic networks? Some topological features that favor ordered behavior may be so fundamental that they are inherently unavoidable, such as the apparent scale-freeness. Beyond that, it is important to note that the physical process of network evolution puts intrinsic constraints onto its architecture, in that the increase in genome size by addition of new genes is ontologically associated with and hence constrained by the physicochemical events of DNA rearrangement (Huang 2004; Taylor and Raes 2004).

For instance, it has been suggested that gene duplication and rewiring of regulatory connectivity (the most likely mechanisms by which the genome increases its gene number) will produce the scale-free architecture (Berg et al. 2004). Other topological features, such as local feed-forward loops (Milo et al. 2002), may also be an inevitable by-product of the historical and constructive constraints of network evolution that happen to have a functional advantage instead of being the result of sculpturing “from scratch” by natural selection for optimal dynamic functionality (Huang 2004). Most likely, both mechanisms (intrinsic constraints of network architecture evolution and natural selection for optimal functionality) may have acted synergistically.

XIII. EXPERIMENTAL EVIDENCE FOR HIGH-DIMENSIONAL ATTRACTORS

Many observations of “macroscopic” cell behavior, some made before the notion of gene regulatory networks, suggest *eis ipsius* that cell fates correspond to robust high-dimensional attractors. However, these observations did not fit in the paradigm of molecular “pathways” dedicated to induce a particular cell fate and hence were not pursued. A long-held view is that differentiation is caused by a “differentiation factor” acting on a cell, proliferation induced by a “growth factor,” and so on. In other words, specific messenger molecules that carry an instructive information trigger a cytoplasmic signal transduction pathway that leads to the activation or suppression of the appropriate genes.

Such a system would not be robust to perturbations, versatile, or evolvable, and could not explain the mutual exclusivity of cell fates (Goss 1967; Waddington 1956). This latter property itself suggests that the action of individual regulatory pathways and local modules must somehow be globally coordinated throughout the cell, lending further support to the notion that gene regulation is coordinated across virtually the entire genome via a genome-wide regulatory network. In fact, the genome-wide pattern of gene activation characteristic for each particular cell phenotype appears to be quite stable, in agreement with the idea that cell types are high-dimensional attractor states (Hsiao et al. 2001; Perou et al. 2000). The high-dimensional robustness is best reflected in the observation that despite the astronomical number of theoretically possible gene activation configurations cells reliably integrate multiple simultaneous and often conflicting signals that affect

genes across the genome (such as cytokines, extracellular matrix, cell-cell contact, and so on) and respond by selecting one of just a few possible cell fates (Evan et al. 1995; Huang and Ingber 2000; Raff 1992).

Often, the very same fate can be triggered by an amazingly broad variety of unrelated signals, including those that lack molecular specificity, such as mechanical forces, cell shape, or non-biological chemicals (Huang and Ingber 2000). Specifically, as elaborate a cellular “program” as differentiation can be induced in many cell lines by nonphysiological chemicals, such as solvents or alcohols, in agreement with the idea that the differentiated state is a stable attractor state (Huang 2002). Similarly, induction of neural tissue in the gastrula stage in early development of vertebrates can experimentally be triggered by many physicochemical perturbations (pH, temperature, dyes), as well as by specific mRNA encoding a broad variety of proteins (De Robertis et al. 2000). Hence, neural differentiation has been dubbed the “default cell fate.”

Even if the ability of diverse and apparently “nonspecific” stimuli to trigger a specific cell fate could be explained by them sharing one common molecular target (e.g., release of a particular cytokine that would be responsible for triggering the cell fate switch), still the other disparate “nonspecific effects” on other target genes would have to be “buffered” away. This response behavior is most simply and naturally explained by a high-dimensional attractor state with a broad basin of attraction.

The advent of massively parallel technologies has now opened the door for actually monitoring the genome-scale dynamics of the intracellular regulatory network. Recent experimental work from our laboratory used DNA microarrays to monitor the dynamics of gene expression profiles (as a surrogate measure of the state vector $S(t)$) in human promyelocytic HL60 cell differentiation (Huang et al. 2005). These cells can undergo neutrophil differentiation in response to a wide array of chemicals, some of them not physiologic and lacking molecular specificity (Collins 1987). We exploited this finding to obtain evidence for a high-dimensional attractor at the level of individual genes by monitoring the actual trajectories in gene expression state space.

By using two biochemically unrelated stimuli, all-trans-retinoic acid and the non-specific solvent dimethylsulfoxide (which both trigger differentiation of HL60 precursor cells into mature neutrophils), the cells could be brought to entirely different states in gene expression state space. However, as the cells differentiate the two high-dimensional state space trajectories converged. This convergence of trajectories is a necessary defining feature of a high-dimensional attractor state. Indeed, we were able to show that the mature differentiated state in HL60 cells was approached from two different directions of the state space with respect to at least 2,000 gene dimensions (Huang et al. 2005). Thus, the neutrophil state appears to be a stable state with respect to 2,000 state space dimensions.

In regard to individual genes, this means that during differentiation the expression levels of a large number of genes are not required to change in a unique manner, but that they have some degree of freedom in their temporal response to

the differentiating stimulus. A gene may be up- or down-regulated initially, but will be attracted to the final expression level characteristic of the differentiated state as the cell reaches the new stable cell fate. This observation defies the traditional notion of unique, specific and dedicated “differentiation pathways.”

XIV. BROADER BIOLOGICAL IMPLICATIONS

The biological ramifications of the presence of a gene expression state space substructured into attractors that collectively form a sort of “epigenetic landscape” (much as in Waddington’s visionary metaphor) are far reaching and may give the conventional paradigm of “signal transduction pathway” new meaning. Some of the implications are quite obvious, which the disciplined biologist with a long history of observing tissue homeostasis and development may immediately appreciate. We can only superficially touch on some points here. In a general sense, the compartmentalization of the gene expression state space into attractors allows the cell to categorize its repertoire of response to external perturbations.

On the basis of this property, cells are like programmable agents that serve as units for a rule-based behavior for the self-assembly of larger multicellular entities: tissues, organs, and organisms. In fact, a portion of the genes expressed in the context of a stable state **S** are cell-cell communication proteins, such as cytokines and extracellular matrix components, which mediate the interaction between cells to form a higher-level cell-cell interaction network. This cellular network in turn has a state space in which attractor states would represent tissue states, such as inflammatory states, regenerative states, and “tumor states.”

This is significant because in cancer biology research evidence is accumulating that confirms the old picture that cancer is not simply a cell-autonomous disease in which a cell clone proliferates in an uncontrolled manner but a tissue disease in which multiple cell types (stroma, epithelial, endothelial and inflammatory cells, and so on) jointly establish a form of unfortunately very stable pathological tissue state.

At the level of an individual cancer cell, the compartmentalization of a (finite) state space into attractor basins also permits a more global view of the effect of mutations in tumorigenesis—beyond the current focus on its effect on a pathway. A mutation then may increase the basin of an attractor at the cost of a neighboring attractor because of the finiteness of the state space. In fact, it appears that tumor cells have an enlarged basin of attraction for the proliferative fate, at the cost of that of the apoptosis of differentiation state. This may explain the puzzling finding that some cytokines, such as TGF- β and GM-CSF which normally cause cell cycle arrest or stimulate differentiation, can turn into a promoter of proliferation in tumors cells (Schmetzer et al. 1999; Tang et al. 2003). In other words, tumor cells may not intrinsically proliferate more rapidly (shorter cell cycle) but rather appear to have a more robust proliferative state (in agreement with an enlarged basin for

the proliferative state attractor) and hence interpret more environmental stimuli as mitogenic.

The stability with respect to multiple state space dimensions displayed by a high-dimensional attractor does not mean that cell states are absolutely robust. Life is the interplay between stability and flexibility. Indeed, the attractors in real cells must be such that despite being inherently stable to many perturbations they allow certain influences to cause a defined transition between attractor states. An important corollary of the high-dimensionality of stability of attractor states is that a change in a large set of genes will be necessary for a transition from one attractor to another (e.g., the switch from proliferative stem cell to the terminally differentiated state during development, or from a quiescent to the proliferative state in tissue regeneration (Huang 2002)).

In contrast, random fluctuations in individual genes or sets of genes would perturb an attractor state to a state within its own basin of attraction, from which the cell can relax back to the original attractor state. This safeguard against unwanted transitions between attractors could explain the pleiotropy and promiscuity in signal transduction pathways, in that signaling proteins affect hundreds to thousands of downstream targets (Fambrough et al. 1999; Menssen and Hermeking 2002). Such fanning-out molecular regulation schemes may have been wired by evolution precisely to produce orchestrated changes in a large but highly distinct set of genes that could both “encode” the conditions for a fate transition and induce a specific transition between different attractor states.

A related fundamental aspect of the concept of an attractor landscape that is hard-wired in the genome is that it provides a deterministic guiding structure, so to speak—a stage on which the inherently noisy molecular processes of regulated gene expression takes place (Paulsson 2004). The stability of high-dimensional attractors ensures that the stochastic fluctuations in the levels of expression and activation of multiple interacting proteins will in general not affect the global cell phenotype but will be limited to a small volume in state space, around the attractor. Thus, a stable phenotype of an apparently homogeneous cell population may actually represent a “swarm” of points in the gene expression state space rather than a single point at the bottom of an attractor.

Accordingly, signal transduction may be viewed as a process that orchestrates these random fluctuations in the expression of thousands of individual genes so that they (metaphorically speaking) “add up their perturbing energy” to increase the probability that an “outlier” cell in the swarm would jump over a crest and enter the basin of attraction of a neighboring state. Because the high dimensionality requires that multiple genes change their expression to produce an attractor switch, a combinatorial scheme in which the fluctuations in the individual genes are biased by the external signal would allow fine-tuning of the cell fate transition probability. Such (epigenetic) cell population heterogeneity and probabilistic response in fact is readily observed. For instance, it has long been known that an increase in cell growth rate reflects an increase in the likelihood that individual cells will enter the cell cycle and progress through the late G1 checkpoint (as measured by an

increase in percentage of cells that enter the DNA synthesis phase of the cell division cycle).

Similarly, the decision of a multipotent progenitor cell to commit to one differentiated cell lineage (Figure 14.1) often appears to be stochastic (Enver et al. 1998; Mayani et al. 1993). There is accumulating evidence from processes monitored at the single-cell level that many other forms of cell regulation are based on probabilistic digital events (Hume 2000; Lahav et al. 2004; Levsky and Singer 2003). In summary, the deterministic structure of the state space allows for molecular noise to be translated into a macroscopically observable stochasticity, such as in the apparently random choice of cell fates in multipotent stem cells. The advantage of stochasticity and heterogeneity of cell phenotypes at the level of cell populations (perhaps robustness in tissue homeostasis) is easily envisioned but remains to be studied more systematically.

XV. CONCLUSIONS

In this chapter, we have discussed the importance of taking a coarser look at the gene regulatory network in order to capture the macroscopic dynamics of cell fate regulation. We have reviewed Boolean networks as a model class suited for such a perspective on gene networks. The trade-off of abstracting away biochemical details is offset by the insight gained into “system-level” qualities of the dynamics of the genome-wide network. The global dynamics of the network exhibits higher-order patterns that map into the observable whole-cell behavior of cell fate regulation. Its characteristic system-level features may be overlooked when solely focusing on characterizing molecular details of individual pathway modules.

Because of the heterogeneity at multiple size scales and the unique blend of universality and idiosyncrasy in living systems, in addition to new technology mental flexibility will also be paramount for the future systems biologist. Thus, the challenge will not only be the development of desirable technologies such as more accurate multiplexed real-time measurement methods for determining gene activity in context (resolved to the single cell level) but a new intellectual structure among computational biologists that embraces both careful quantitative modeling as well as abstraction and simplification.

REFERENCES

- Akashi, K., He, X., Chen, J., Iwasaki, H., Niu, C., Steenhard, B., Zhang, J., Haug, J., and Li, L. (2003). Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis. *Blood* **101**:383–389.
- Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 17–28.

- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (1994). *Molecular Biology of the Cell*. (3d ed.). New York: Garland Publishing.
- Aldana, M., and Cluzel, P. (2003). A natural class of robust networks. *Proc. Natl. Acad. Sci. USA* **100**:8710–8714.
- Amaral, L. A., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**:11149–11152.
- Angeli, D., Ferrell, J. E. Jr., and Sontag, E. D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl. Acad. Sci. USA* **101**:1822–1827.
- Arney, K. L., and Fisher, A. G. (2004). Epigenetic aspects of differentiation. *J. Cell Sci.* **117**:4355–4363.
- Bagley, R. J., and Glass, L. (1996). Counting and classifying attractors in high dimensional dynamical systems. *J. Theo. Biol.* **183**:269–284.
- Barabasi, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**:509–512.
- Bar-Yam, Y. (2000). *Unifying Themes in Complex Systems: Proceedings of the International Conference on Complex Systems*. Philadelphia: Perseus Press.
- Becskei, A., Seraphin, B., and Serrano, L. (2001). Positive feedback in eukaryotic gene networks: Cell differentiation by graded to binary response conversion. *EMBO J.* **20**:2528–2535.
- Berg, J., Lassig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **4**:51.
- Bhalla, U. S., Ram, P. T., and Iyengar, R. (2002). MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297**:1018–1023.
- Bouvier, M. (1990). Cross-talk between second messengers. *Ann. NY Acad. Sci.* **594**:120–129.
- Bray, D. (2003). Molecular networks: The top-down view. *Science* **301**:1864–1865.
- Bruno, L., Hoffmann, R., McBlane, F., Brown, J., Gupta, R., Joshi, C., Pearson, S., Seidl, T., Heyworth, C., and Enver, T. (2004). Molecular signatures of self-renewal, differentiation, and lineage choice in multipotential hemopoietic progenitor cells *in vitro*. *Mol. Cell Biol.* **24**:741–756.
- Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E., and Strogatz, S. H. (2001). Are randomly grown graphs really random? *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **64**:041902.
- Chang, H. Y., Chi, J. T., Dudoit, S., Bondre, C., van de Rijn, M., Botstein, D., and Brown, P. O. (2002). Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proc. Natl. Acad. Sci. USA* **99**:12877–12882.
- Chen, H., Ray-Gallet, D., Zhang, P., Hetherington, C. J., Gonzalez, D. A., Zhang, D. E., Moreau-Gachelin, F., and Tenen, D. G. (1995). PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**:1549–1560.
- Cherry, J. L., and Adler, F. R. (2000). How to make a biological switch. *J. Theo. Biol.* **203**:117–133.
- Chi, J. T., Chang, H. Y., Haraldsen, G., Jahnsen, F. L., Troyanskaya, O. G., Chang, D. S., Wang, Z., Rockson, S. G., van de Rijn, M., Botstein, D., et al. (2003). Endothelial cell diversity revealed by global expression profiling. *Proc. Natl. Acad. Sci. USA* **100**:10623–10628.
- Cinquin, O., and Demongeot, J. (2002). Positive and negative feedback: striking a balance between necessary antagonists. *J. Theo. Biol.* **216**:229–241.
- Collins, S. J. (1987). The HL-60 promyelocytic leukemia cell line: Proliferation, differentiation, and cellular oncogene expression. *Blood* **70**:1233–1244.

- Delbrück, M. (1949). Discussion. In *Unités biologiques douées de continuité génétique Colloques Internationaux du Centre National de la Recherche Scientifique*. Paris: CNRS.
- De Robertis, E. M., Larrain, J., Oelgeschlager, M., and Wessely, O. (2000). The establishment of Spemann's organizer and patterning of the vertebrate embryo. *Nat. Rev. Genet.* **1**:171–181.
- Derrida, B., and Pomeau, Y. (1986). Random networks of automata: A simple annealed approximation. *Europhys. Lett.* **1**:45–49.
- D'Haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **16**:707–726.
- Ehrenberg, M., Elf, J., Aurell, E., Sandberg, R., and Tegner, J. (2003). Systems biology is taking off. *Genome Res.* **13**:2377–2380.
- Eisenberg, E., and Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends Genet.* **19**:362–365.
- Endy, D., and Brent, R. (2001). Modelling cellular behaviour. *Nature* **409**:391–395.
- Enver, T., and Greaves, M. (1998). Loops, lineage, and leukemia. *Cell* **94**:9–12.
- Enver, T., Heyworth, C. M., and Dexter, T. M. (1998). Do stem cells play dice? *Blood* **92**:348–351.
- Evan, G. I., Brown, L., Whyte, M., and Harrington, E. (1995). Apoptosis and the cell cycle. *Curr. Opin. Cell Biol.* **7**:825–834.
- Fambrough, D., McClure, K., Kazlauskas, A., and Lander, E. S. (1999). Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97**:727–741.
- Fox, J. J., and Hill, C. C. (2001). From topology to dynamics in biochemical networks. *Chaos* **11**:809–815.
- Friedman, N., Linal, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**:601–620.
- Galloway, J. L., Wingert, R. A., Thisse, C., Thisse, B., and Zon, L. I. (2005). Loss of *gata1* but not *gata2* converts erythropoiesis to myelopoiesis in zebrafish embryos. *Dev. Cell* **8**:109–116.
- Gardner, M. R., and Ashby, W. R. (1970). Connectance of large dynamic (cybernetic) systems: Critical values for stability. *Nature* **228**:784.
- Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**:339–342.
- Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**:102–105.
- Georgopoulos, K. (2002). Haematopoietic cell-fate decisions, chromatin regulation and ikaros. *Nat. Rev. Immunol.* **2**:162–174.
- Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., and Bickmore, W. A. (2004). Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell* **118**:555–566.
- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**:1727–1736.
- Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *J. Theo. Biol.* **39**:103–129.
- Goss, R. J. (1967). The strategy of growth. In H. Teir and T. Ryttaumaa (eds.). *Control of Cellular Growth in the Adult Organism*, pp. 3–27, London: Academic Press. pp. 3–27.
- Graf, T. (2002). Differentiation plasticity of hematopoietic cells. *Blood* **99**:3089–3101.

- Grass, J. A., Boyer, M. E., Pal, S., Wu, J., Weiss, M. J., and Bresnick, E. H. (2003). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl. Acad. Sci. USA* **100**:8811–8816.
- Harris, S. E., Sawhill, B. K., Wuensche, A., and Kauffman, S. A. (2002). A model of transcriptional regulatory networks based on biases in the observed regulation rules. *Complexity* **7**:23–40.
- Hsiao, L. L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P., et al. (2001). A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7**:97–104.
- Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* **11**:774–785.
- Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states: An integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.* **77**:469–480.
- Huang, S. (2001). Genomics, complexity and drug discovery: Insights from Boolean network models of cellular regulation. *Pharmacogenomics* **2**:203–222.
- Huang, S. (2002). Regulation of cellular states in mammalian cells from a genome-wide view. In J. Collado-Vides and R. C. Hofstadter (eds.), *Gene Regulation and Metabolism: Post-Genomic Computational Approach*, pp. 181–220, Cambridge, MA: MIT Press.
- Huang, S. (2004). Back to the biology in systems biology: What can we learn from biomolecular networks. *Brief Funct. Genomics Proteomics* **2**:279–297.
- Huang, S., and Ingber, D. E. (2000). Shape-dependent control of cell growth, differentiation, and apoptosis: Switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261**:91–103.
- Huang, S., Eichler, G. S., Bar-Yam, Y., and Ingber, D. E. (2005). Cell fate as high-dimensional attractor of a complex gene regulatory network. *Phys. Rev. Lett.* (in press).
- Hume, D. A. (2000). Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. *Blood* **96**:2323–2328.
- Jablonka, E., and Lamb, M. J. (2002). The changing concept of epigenetics. *Ann. N. Y. Acad. Sci.* **981**:82–96.
- Kauffman, S. (2004). A proposal for using the ensemble approach to understand genetic regulatory networks. *J. Theo. Biol.* **230**:581–590.
- Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2003). Random Boolean network models and the yeast transcriptional network. *Proc. Natl. Acad. Sci. USA* **100**:14796–14799.
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theo. Biol.* **22**:437–467.
- Kauffman, S. A. (1993). *The Origins of Order*. New York: Oxford University Press.
- Khorasanizadeh, S. (2004). The nucleosome: From genomic organization to genomic regulation. *Cell* **116**:259–272.
- Koshland, D. E., Jr., Goldbeter, A., and Stock, J. B. (1982). Amplification and adaptation in regulatory and sensory systems. *Science* **217**:220–225.
- Kubicek, S., and Jenuwein, T. (2004). A crack in histone lysine methylation. *Cell* **119**:903–906.
- Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A. J., Elowitz, M. B., and Alon, U. (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nat. Genet.* **36**:147–150.
- Laurent, M., and Kellershohn, N. (1999). Multistability: A major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.* **24**:418–422.

- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**:799–804.
- Levsky, J. M., and Singer, R. H. (2003). Gene expression and the myth of the average cell. *Trends Cell Biol.* **13**:4–6.
- Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**:540–543.
- Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 18–29.
- Maliackal, P. J., Brock, A., Ingber, D. E., and Huang, S. (2005). High “betweenness” proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* (in press).
- Marcotte, E. M. (2001). The path not taken. *Nat. Biotechnol.* **19**:626–627.
- May, R. M. (1972). Will a large complex system be stable? *Nature* **238**:413–414.
- Mayani, H., Dragowska, W., and Lansdorp, P. M. (1993). Lineage commitment in human hemopoiesis involves asymmetric cell division of multipotent progenitors and does not appear to be influenced by cytokines. *J. Cell Physiol.* **157**:579–586.
- Menssen, A., and Hermeking, H. (2002). Characterization of the c-MYC-regulated transcriptome by SAGE: Identification and analysis of c-MYC target genes. *Proc. Natl. Acad. Sci. USA* **99**:6274–6279.
- Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30**:31–34.
- Meyer, D. A., and Brown, T. A. (1998). Statistical mechanics of voting. *Phys. Rev.* **81**:1718–1721.
- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science* **303**:1538–1542.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298**:824–827.
- Monod, J., and Jacob, F. (1961). Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26**:389–401.
- Muller, F., Bernard, V., and Tobler, H. (1996). Chromatin diminution in nematodes. *Bioessays* **18**:133–138.
- Myer, J. (2001). Personal Communication.
- Nerlov, C., Querfurth, E., Kulesa, H., and Graf, T. (2000). GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* **95**:2543–2551.
- Novick, A., and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA* **43**:553–566.
- Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., et al. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**:1378–1381.
- Ohneda, K., and Yamamoto, M. (2002). Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta Haematol.* **108**:237–245.
- Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and Van Oudenaarden, A. (2004). Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**:737–740.
- Paulsson, J. (2004). Summing up the noise in gene networks. *Nature* **427**:415–418.

- Paulsson, J., Berg, O. G., and Ehrenberg, M. (2000). Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl. Acad. Sci. USA* **97**:7148–7153.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**:747–752.
- Picht, P. (1969). *Mut zur Utopie*. München: Piper.
- Raff, M. C. (1992). Social controls on cell survival and cell death. *Nature* **356**:397–400.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**:1551–1555.
- Rubin, H. (1990). On the nature of enduring modifications induced in cells and organisms. *Am. J. Physiol.* **258**:L19–L24.
- Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J. (2001). RegulonDB (version 3.2): Transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**:72–74.
- Savageau, M. A. (1995). Michaelis-Menten mechanism reconsidered: Implications of fractal kinetics. *J. Theo. Biol.* **176**:115–124.
- Schmetzer, H. M., Gerhartz, H. H., and Wilmanns, W. (1999). GM-CSF stimulates proliferation of clonal leukemic bone marrow cells in acute myeloid leukemia (AML) *in vitro*. *Ann. Hematol.* **78**:449–455.
- Sha, W., Moore, J., Chen, K., Lassaletta, A. D., Yi, C. S., Tyson, J. J., and Sible, J. C. (2003). Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts. *Proc. Natl. Acad. Sci. USA* **100**:975–980.
- Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**:64–68.
- Shmulevich, I., and Kauffman, S. A. (2004). Activities and sensitivities in Boolean network models. *Phys. Rev. Lett.* (in press).
- Southan, C. (2004). Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics* **4**:1712–1726.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature* **410**:268–276.
- Tang, B., Vu, M., Booker, T., Santner, S. J., Miller, F. R., Anver, M. R., and Wakefield, L. M. (2003). TGF-beta switches from tumor suppressor to prometastatic factor in a model of breast cancer progression. *J. Clin. Invest.* **112**:1116–1124.
- Taylor, J. S., and Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Ann. Rev. Genet.* **38**:615–643.
- Thomas, R. (1978). Logical analysis of systems comprising feedback loops. *J. Theo. Biol.* **73**:631–656.
- Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* **15**:221–231.
- Waddington, C. H. (1956). *Principles of Embryology*. London: Allen & Unwin.
- Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of “small-world” networks. *Nature* **393**:440–442.
- Wuensche, A. (1998). Genomic regulation modeled as a network with basins of attraction. *Pac. Symp. Biocomput.* 89–102.
- Xiong, W., and Ferrell, J. E., Jr. (2003). A positive-feedback-based bistable “memory module” that governs a cell fate decision. *Nature* **426**:460–465.

- Yeung, M. K., Tegner, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**: 6163–6168.
- Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K. A., Mueller, B. U., Narravula, S., Torbett, B. E., Orkin, S. H., and Tenen, D. G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood* **96**:2641–2648.
- Zingg, J. M., Pedraza-Alva, G., and Jost, J. P. (1994). MyoD1 promoter autoregulation is mediated by two proximal E-boxes. *Nucleic Acids Res.* **22**:2234–2241.

Spatiotemporal Systems Biology

Avijit Ghosh*, **David Miller***, **Rui Zou*[†]**,
Bahrad Sokhansanj[†], and **Andres Kriete[†]**

*Department of Physics**, *School of Biomedical Engineering[†]*, *Science and Health Systems, Drexel University, Philadelphia, Pennsylvania, USA*

Chapter 15

ABSTRACT

Computational and theoretical considerations for the extension of systems biology into the spatiotemporal realm will be discussed. Both limitations and extensions of current approaches within the research community will be investigated, along with the approach taken by our group in a newly developed software package, CellSim. Application of all computational aspects described rely on cellular assays imaged by fluorescence confocal microscopy. Taken together, this extension to systems biology may answer questions about complex protein networks and the role spatial heterogeneity may play in such processes.

I. INTRODUCTION

This chapter provides an introduction into spatiotemporal (ST) systems biology. The chapter is organized as follows. First, the theoretical foundations (Section II), consisting of both biophysical and related numerical aspects, are introduced. Based on these fundamentals, the cellular simulator CellSim and examples of simulations are given (Section III). Data that feeds this simulator may be based on kinetic imaging, described in Section IV. The chapter concludes with a listing of recommended resources.

A. Cell compartmentalization and heterogeneity

Systems biology has, until recently, considered the cellular activity to be fully described as simply a set of complex coupled chemical reactions that occur

concurrently to bring about the disparate and multifaceted behavior exhibited by cells. In this sense, perhaps one of the most important aspects of systems biology is the very real emphasis on describing the cell (and its chief component, protein) as being intimately part of this complexity, manifested in networks of protein and messenger molecule cascades described in detail in previous sections of this volume. In this view, the complexity of cellular function is manifest through these cascades, which may exhibit—through feedback, feed-forward, amplification, and other signaling processes—important biological regulatory and functional mechanisms controlling all aspects of cellular function, from metabolism to cellular growth.

It is a testament to systems biology's recent coming of age that even this immense complexity belies the true nature of cells. A cursory glimpse into real living cells gives rise to the notion that cells are immense, heterogeneous, complex machines with a hierarchy of macroscopic (approximately 10^{-6} m) to microscopic (approximately 10^{-9} m) features acting in unison. Furthermore, cells are organized in multi-cellular systems on a much larger scale into an array of specialized and differentiated groups forming organs and other structures that encompass a viable living creature.

A host of compartments—such as the mitochondria, endoplasmic reticulum (ER), nucleus, Golgi apparatus, lysosomes, and peroxisomes—play important and *localized* roles in cellular function. The nucleus serves as a repository for the genome and is the chief location of regulatory processes controlling gene expression as well as DNA and RNA synthesis. Synthesis of the integral membrane and secretory proteins occur within the ER and are later trafficked to their appropriate locations. The Golgi apparatus is not only a major site of carbohydrate synthesis but a provider of the conduit for trafficked proteins exported from the ER. Mitochondria, which represent the energy factories of the cellular machinery, are the sources for ATP synthesis.

Defunct macromolecules are degraded in lysosomes. Specific oxidative reactions that would be harmful if occurring in the cytosol are confined within peroxisomes. Although the complexity of cells is inherently inscribed by the wide array of interacting protein and molecular networks and systems, the heterogeneous nature of these compartments and their interactions play a large role in regulating the protein networks thus far described. Thus, cellular complexity is inherently spatiotemporal—described more fully as not only sets of complex protein networks *within* organelles and the cytosol, but as a set of interactions *between* compartments and the cytosol.

Protein motility within cells is guided by both passive and active transport, with protein localization controlled by specialized sorting signals (either peptides or patches). Gated transport regulates trafficking between the nucleus and the cytosol, whereas transmembrane protein complexes can directly transport proteins through the complex into a neighboring compartment. In addition, a large amount of soluble protein is also transported by vesicular transport. In this mechanism, a vesicle is formed in a source compartment containing the proteins to be transported and is subsequently ejected and then localized to the destination compartment.

In all three cases protein transport may be described as a combination of random motion and localized recognition via binding events. The recognition occurs

through specific signal peptides or patches that may bind to a complementary recognition complex. In gated transport and transmembrane protein complexes, the complementary recognition complex is itself directly part of the transmembrane protein or the nuclear pore complex. On the other hand, vesicular transport is controlled specifically by SNAREs and targeting GTPases, which serve a similar function but will localize the entire vesicle rather than a single complex. Transport of a protein to a nuclear pore complex or to a transmembrane complex is chiefly governed by random thermal motions within the organelle itself. Similarly, localization of a vesicle to a target organelle may be considered random diffusion of the vesicle coupled to SNAREs or GTPases (which provide localization to the targeted organelle).

B. Diffusion

Diffusion, the natural random motion of objects through a medium, plays a vital role in cell functioning in many processes such as calcium transport, transcription, and non-equilibrium dynamics (Brown and Kholodenko 1999; Kholodenko et al. 2000; Kholodenko 2003; Peletier et al. 2003). As described previously, nature has given cells numerous mechanisms for transporting materials into and out of the cell, as well as moving materials to different locations within the cell itself—notably, transporter proteins, motor proteins, and transport via potential differences and ion gradients.

Typically, diffusion is neglected in most systems biology models. The model cell is instead treated as a single point in space possessing instant dilution, often called the “well-stirred” approximation. This is due to the added complexity of modeling diffusion and lack of straightforward experimental techniques to provide the necessary measurements needed to fully describe a spatiotemporal model. If the time resolution of the system is large enough, this approximation is valid for many materials with fast diffusion rates and/or small volumes. Furthermore, in many cases the diffusion constant may be folded into the effective association or disassociation rate constants in Michaelis-Menten reactions. In this approximation, diffusion acts simply as a mechanism to slow down the apparent associative or disassociative rate constant, and transport between compartments may be effectively treated as gradients between spatially averaged concentrations of the transported species.

Concentration gradients of enzymes within cells that modulate signal transduction belie this simplicity (Khurana et al. 1996; Holdaway-Clarke et al. 1997; Lam et al. 2003; Belenkaya et al. 2004). With experimental and computational technological advancements allowing finer temporal and spatial resolution, the development of spatiotemporal extensions to traditional systems biology has become much more tractable. Unless the timescale of interest is fast enough to neglect intra-compartmental concentration gradients or the concentration gradient is essentially flat, diffusion is likely to play a critical role in governing the time evolution of the system and should not be ignored.

II. SPATIOTEMPORAL SYSTEMS BIOLOGY: THEORY

A. The mathematics of the diffusion equation

Diffusion is based on the fact that random Brownian movement (Brown 1827) is statistically likely to cause particles in areas of higher concentration to move to areas of lower concentration. One may view this phenomenon as a mathematical consequence of the fact that particles are more likely to move to a lower concentrated area simply because there are more particles in the high-concentration area that can randomly move to the low-concentration area than particles in the low-concentration area that can do the reverse. The mathematical equation describing diffusion is, aptly, the diffusion equation

$$\frac{\partial C}{\partial t} = D\nabla^2 C \quad (15.1)$$

This describes how the time rate of change of the amount of a substance C at a location is proportional to the second spatial derivative at the same location. The expression can be derived for the case of one spatial dimension simply using elementary arguments on a Cartesian grid, and can easily be expanded to higher dimensions by superposition. We shall do this here for illustrative purposes.

Assume that on a 1D grid a single particle takes a random right or left step of length dx in each time span of dt . Each step is taken to be independent of all previous steps, and the total number of particles involved is high enough to validate our probability assumptions. On average, the change in number of particles Δn at a position x in a time step dt is given by

$$\Delta n = n_x^{t+1} - n_x^t \quad (15.2)$$

with the subscript x representing position and the superscript t representing time. For readability, n_x^{t+1} should be interpreted as the number of particles at position $x - dx$ at time $t + dt$. (that is, ± 1 represents plus or minus one infinitesimal in the appropriate units). Consider the following discretization: If the particles make steps of dx each and every time step dt and the particles have a probability p_l of moving to the left and probability p_r of moving to the right, Δn is

$$\Delta n = p_r n_{x-1}^t - (p_l + p_r) n_x^t + p_l n_{x+1}^t \quad (15.3)$$

which simplifies to

$$\begin{aligned} \Delta n &= \frac{1}{2} n_{x-1}^t - n_x^t + \frac{1}{2} n_{x+1}^t \\ &= \frac{1}{2} (n_{x-1}^t - 2n_x^t + n_{x+1}^t) \end{aligned} \quad (15.4)$$

Multiplying through by the identity $\frac{dx^2}{dt} \frac{dt}{dx^2}$ gives

$$\frac{dn_x^t}{dt} = D \frac{n_{x-1}^t - 2n_x^t + n_{x+1}^t}{dx^2} \quad (15.5)$$

where $D = \frac{dx^2}{2dt}$ equals the diffusion constant. Letting the infinitesimals go to zero while keeping D constant results in the original diffusion equation (Borman et al. 2004).

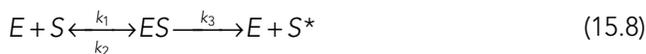
It is worth noting that the diffusion constant itself is dependent on a variety of factors, such as the size/shape of the diffusing particles and the viscosity/density of the diffusive medium, and thus must be derived experimentally (Arrio-Dupont et al. 2000). For certain simple cases, the diffusion equation may be analytically integrated. However, in general such analytic solutions do not exist for diffusion problems, and certainly not for problems pertaining to cells, where cell geometry, kinetics, and non-trivial initial conditions complicate the problem. Before embarking on this more complicated problem, we will first provide a cursory review of the coupled problem in the reaction-diffusion equation: namely, the reaction portion.

1. The mathematics of chemical kinetics

The framework of the reaction part of the reaction-diffusion equation is grounded in kinetic rate theory (Purich 2004). Every interaction between members of the signal cascade is expressed as a set of basic chemical reactions between species, such as:



where (6) represents an aggregation event between species A and B , and (7) represents a chemical reaction between A and B forming products C and D . k_f and k_b are the forward and backward rate constants to be determined from an analysis of the response of mammalian cell assays to various perturbations. Enzymatic reactions such as phosphorylation or acetylation are represented using the Michaelis-Menten formulation:



Such an enzymatic process is the product of two sequential processes. The catalytic step is irreversible with a rate constant of k_3 , and the association is reversible with forward and backward rate constants of k_1 and k_2 , respectively. The system of kinetic reactions represented by (6) and (7) can be rewritten as a series of ordinary differential equations (ODEs). These equations describe a contribution to the rate of change in concentration of a particular species as a function of time:

$$\frac{d[A]}{dt} = k_b[AB] - k_f[A][B] \quad (15.9)$$

$$\frac{d[A]}{dt} = k_b[C][D] - k_f[A][B] \quad (15.10)$$

with corresponding ODEs for each of the other species expressed in (6) and (7). The entire pathway is represented as a system of differential equations that describes the change in concentration of any particular species as a function of rate constants. Modeling protein interactions using only equations of type (9) and (10) is referred to as the well-stirred approximation. The cell is assumed to be “infinitely mixed” or homogenous.

2. Stochastic models

Simulating differential equations to model reaction-diffusion processes will accurately predict the average behavior of (1) large numbers of molecules within cells and (2) the average outcome of a cell process over a large number of cells. However, in many cases deterministic and continuous approaches cannot accurately simulate biological phenomena that arise from stochastic effects. For example, in the case of cancer random molecular and cellular effects with low individual probability accumulate, eventually causing dramatic physiological effects.

Biological systems, particularly those involved with genetic regulation, are very noisy—and distinct phenotypic outcomes directly result from that noise (McAdams and Arkin 1997, 1999; Elowitz and Leibler 2000). The problem of noise is exacerbated by the low cellular concentrations typical of many key regulatory proteins. If one speaks of nanomolar concentrations of a protein, that corresponds to just a few to tens of individual protein molecules. For example, in gene regulation there are only a few sites on DNA (which can be thought of as individual “molecules” or reaction sites) where transcription factors can bind and mRNA be produced. Therefore, stochastic and discrete simulations may be necessary to develop accurate reaction-diffusion models for such processes. Recently, an extensive review focusing on simulation in bacterial cells was conducted by McAdams and Arkin (1998).

Because biological processes involve a large number of molecules and protein species, the state space is too large for an exact solution of stochastic differential equations describing a reaction. Gillespie (1976, 1977) proposed a Monte Carlo method to exactly simulate the stochastic time evolution of a reaction system. The probability of each reaction occurring is a function of its rate constant (measured experimentally) and the number of available reactants in the simulation. At each point in time, there exists a joint probability distribution function for both the reaction and the time at which it can occur.

This generates a random trajectory through the state space that converges in the mean to the solution of the continuum model. Similarly, an average over an appropriate set of repeated experiments is expected to lead to the solution from a continuum model, and in this context one may view deterministic spatiotemporal models as the expected solutions from an appropriate ensemble average of experiments. This is convenient in that these ensemble averages are the simplest experimentally reproducible observables.

Arkin et al. (1998) applied the Gillespie method to a fully stochastic model of *E. coli* infected by the λ phage virus, with two outcomes: lysogeny (integration of the

phage into the bacterial DNA and “quiet” replication) and lysis (explosion of the cell and virus release). The simulation incorporated transcription and translation of genes, protein-protein and DNA-protein reactions responsible for replication, and proteases, for a total of 32 chemical reactions (including transcription and translation, which were modeled as hundreds of individual reaction events for each base). The simulation was implemented using parallel supercomputers. However, subsequent algorithmic improvements (Gibson and Bruck 2000) have made it much faster without changing any physical assumptions.

Whereas most applications of the Gillespie approach to stochastic reaction simulation have been only for a homogenous volume (i.e., “1D”—reaction systems), it has recently been applied to non-biological surface chemistry (Lukkien et al. 1998). A significant drawback is scalability, in that the number of time steps that must be computed increases with the total number of protein molecules to an intractable point for eukaryotic cells. Thus, much recent work has been devoted to developing accelerated and adaptive methods that integrate stochastic-discrete and deterministic-continuum methods at appropriate time scales.

Stochastically-induced spatiotemporal patterns of Jung and Mayer-Kress have biological applications (Jung and Mayer-Kress 1995a, 1995b) to evolution (Dunkel et al. 2004), electrochemical oscillators (Kiss et al. 2004), neuronal models (Doiron et al. 2004), and calcium signaling (Coombes et al. 2004). Turner et al. provide an excellent review of the state-of-the-art in stochastic biochemical simulation (Turner et al. 2004).

B. The mathematics and numerical analysis of the reaction-diffusion equation

In the spatiotemporal extension of this classical model, transport is treated explicitly. Active transport is modeled using elementary reactions that couple to transporter proteins and may be represented by differential equations of the type *AND*. Passive transport can be represented with the diffusion equation for each species, as

$$\frac{\partial C}{\partial t} = D\nabla^2 C \quad (15.11)$$

where D is the diffusion constant for that particular species. Active transport along actin filaments, for instance, may be modeled directly as part of a system of ATP-driven chemical reactions.

As rate parameters need to be derived by the appropriate experimental approaches, diffusion constants may be estimated by experimental techniques such as using modulated fringe pattern photo-bleaching (Arrio-Dupont et al. 2000). The key to building a quantitative model of the dynamical behavior of the chromatin network of the spatio-temporal system is coupling the system of ODEs representing the enzymatic kinetics (equations 15.9 and 15.10) with a system of partial differential equations (equation 15.11) representing the diffusive behavior of each species within the nucleus or on the membrane. The total contribution to the rate of change in concentration of any species at position \vec{r} is the sum total of the con-

tributions to the rate of change from all relevant reactions and transport equations. The coupling between transport and molecular kinetics may then be rewritten in a mixed finite-difference format as follows:

$$\frac{\Delta[X_r^i]}{\Delta\tau} = D_i \left(\frac{[C_{r+1}^i] - 2[C_r^i] + [C_{r-1}^i]}{\Delta x^2} \right) + \sum_j k_{ij} [C_r^i] + \sum_{l,m} k_{ilm} [C_r^l] [C_r^m] + \sum_j \frac{P_{ij}}{V_i} ([C_r^j] - [C_r^i]) \quad (15.12)$$

The term (C_i) represents the concentration of species i at point \vec{r} in the nuclear matrix. D_i is the diffusion constant for species i . The first sum tallies all unimolecular reactions involving species i , the second sum tallies bimolecular reactions, and the final sum represents passive diffusive transport of species i between compartments. The parameter P_{ij} is the permittivity of channel ij and V_i is the volume of the destination compartment. Higher-order reactions may be included in the obvious generalized fashion. In the previous equation, Δx is the spatial separation between two consecutive points, and Δt represents the temporal resolution of the numerical analysis. Using this formulation, the time evolution of each species in the protein network may be followed both spatially and temporally.

1. Operator splitting

For the combined reaction-diffusion system, one may use operator splitting to propagate the total operator. Given

$$\frac{dC(t)}{dt} = L_{RD}C(t) = (L_R + L_D)C(t) \quad (15.13)$$

where L_{RD} is a reaction diffusion operator, L_R and L_D are the individual reaction and diffusion operators with corresponding propagators $U_r(t)$ and $U_d(t)$:

$$\begin{aligned} C^{t+1} &= U_R(\delta t)C^t \\ C^{t+1} &= U_D(\delta t)C^t \end{aligned} \quad (15.14)$$

The second-order Strang splitting method (Strang 1968) may be written as

$$C(t + \delta t) = U_R\left(\frac{\delta t}{2}\right)U_D(\delta t)U_R\left(\frac{\delta t}{2}\right)C(t) \quad (15.15)$$

In a software package CellSim, described in more detail later in this document, we have implemented the reaction-diffusion-reaction ordering for the splitting as recommended by Sportisse (2000) and implemented by others (Singer and Pope 2004). For reaction-limited models, CellSim implements an adaptive time step algorithm that uses the second-order Rosenbrock method to propagate the first operator a half step.

The time step determined by the reaction operator is then used to propagate the diffusion operator and then the second half of the reaction operator. It must be emphasized that this adaptive scheme is only valid for stiff reaction-limited reaction-diffusion models. A more general approach, currently being implemented,

uses both the error of each operator as well as the splitting error to estimate the time step (Miller and Ghosh, in prep.).

2. The diffusion operator

Many schemes exist for integration of diffusion. The most straightforward implementation of the diffusion operator is the forward time-centered space algorithm (FTCS). Using reduced units by setting the constant $a = \frac{Ddt}{dx^2}$, the FTCS method calculates the concentration n at the next time step as follows:

$$n_x^{t+1} = (1 - 2a)n_x^t + a(n_{x-1}^t + n_{x+1}^t) \quad (15.16)$$

Stability analysis of this algorithm reveals a stability condition of $2a < 1$ for the method to be stable. Although this method is simple and stable for small time steps, it is generally inefficient and undesirable. To remove the stability condition, one could use a first-order implicit scheme in which we apply the Laplacian a step dt ahead of the current time,

$$\frac{n_x^{t+1} - n_x^{t-1}}{dt} = D \left(\frac{n_{x-1}^{t+1} - 2n_x^{t+1} + n_{x+1}^{t+1}}{dx^2} \right) \quad (15.17)$$

If spatial boundary conditions (Dirichlet or von Neumann) are known, the set of equations produced by the previous equation can be solved iteratively. Such solution by recursion is typical of implicit methods wherein concentrations at a forward time step appear on the right-hand side of the equation. Related to this approach are second-order schemes such as Crank-Nicolson (1947), which has a simple description as the average of the previous two methods:

$$\frac{n_x^{t+1} - n_x^t}{dt} = \frac{1}{2} \left(D \frac{n_{x-1}^{t+1} - 2n_x^{t+1} + n_{x+1}^{t+1}}{dx^2} + D \frac{n_{x-1}^t - 2n_x^t + n_{x+1}^t}{dx^2} \right) \quad (15.18)$$

Crank-Nicolson is unconditionally stable for dt and dx , and yields second-order accuracy in time and space. Implicit methods have the main advantage of being unconditionally stable, but they also require a matrix inversion. For 1D problems, this method requires the diagonalization of a tridiagonal matrix at each time step. Whereas the 1D case is relatively inexpensive, 2D and especially 3D problems require solutions of considerably more complex (although still sparse) matrices. To alleviate this unwieldy structure, further operator splitting of the diffusion operator into three 1D operators may be used.

This involves splitting the multi-dimensional diffusion into appropriate time intervals and applying a 1D step for each direction. In two dimensions, using two steps of $\frac{\delta t}{2}$ the scheme's stability properties are maintained, but this is lost in three dimensions (with three $\frac{\delta t}{3}$ time steps) and the scheme becomes only conditionally stable (Press 1992). For problems in higher dimensions, an Alternating Direction Implicit

(ADI) introduced by Douglas (Douglas 1962), maintains unconditional stability, is second order accurate in both space and time, and is generalizable for solving diffusion problems of arbitrary dimensionality. In 3D, it may be schematically written out as:

$$\begin{aligned}\frac{w^* - w_n}{\Delta t} &= \frac{\alpha}{2} \Delta_x^2 (w^* - w_n) + \alpha \Delta_y^2 w_n + \alpha \Delta_x^2 w_n \\ \frac{w^{**} - w_n}{\Delta t} &= \frac{\alpha}{2} \Delta_x^2 (w^* - w_n) + \frac{\alpha}{2} \Delta_y^2 (w^{**} - w_n) + \alpha \Delta_x^2 w_n \\ \frac{w_{n+1} - w_n}{\delta t} &= \frac{\alpha}{2} \Delta_x^2 (w^* - w_n) + \frac{\alpha}{2} \Delta_y^2 (w^{**} - w_n) + \alpha \Delta_x^2 (w_{n+1} - w_n)\end{aligned}\quad (15.19)$$

where $\alpha = \frac{D}{\delta x^2}$ and Δ_{dir}^2 is a simple second order finite difference along a strip of space in the direction of the subscript:

$$\Delta_x^2 w_n = w_n(x - \Delta x) - 2w_n x + w_n(x + \Delta x) \quad (15.20)$$

Subtracting (19a) from (19b) and (19b) from (19c), reduces the scheme to three tridiagonal systems of equations, each of which can be solved efficiently using elementary linear algebra. The method. In 1D, the scheme reduces to the standard Crank-Nicolson diffusion scheme (Crank-Nicolson 1947).

3. The reaction operator

The reaction operator may be integrated using a host of standard methods. Currently, CellSim has the following integrators.

- Euler
- Exponential Euler
- Second- and fourth-order Runge-Kutta
- Adaptive fourth-order Runge-Kutta
- Second- and fourth-order adaptive Rosenbrock

Although Euler ($(y(t + \delta t) = \dot{y}(t, y(t))\delta t)$) is perhaps the simplest of numerical integrators, it is neither particularly stable nor accurate. For problems in chemistry and biology, exponential Euler takes advantage of the fact that simple kinetic interactions often give rise to exponential decay functions. That is, for kinetics one frequently encounters equations of the form

$$\frac{dy}{dt} = A - By \quad (15.21)$$

Schematically, the exponential Euler method may be written as

$$y(t + \delta t) = y(t)e^{-B\delta t} + \frac{A}{B}(1 - e^{-B\delta t}) \quad (15.22)$$

Although this scheme allows for the use of larger time steps, at low concentrations this scheme suffers from some inaccuracy that will propagate through the system. Therefore, although popular this method should be used with some caution. The workhorses of ODE solvers, Runge-Kutta methods have been implemented within CellSim to address this problem. The commonly used fourth-order formulation uses four strategically placed evaluations of the function's derivative within a given time step δt , and a weighted average of these derivatives is used to propagate the system a full time step:

$$\begin{aligned}
 k_1 &= \dot{y}(t, y(t))\delta t \\
 k_2 &= \dot{y}\left(t + \frac{\delta t}{2}, y(t) + \frac{k_1}{2}\right)\delta t \\
 k_3 &= \dot{y}\left(t + \frac{\delta t}{2}, y(t) + \frac{k_2}{2}\right)\delta t \\
 k_4 &= \dot{y}(t + \delta t, y(t) + k_3)\delta t \\
 y(t + \delta t) &= y(t) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)
 \end{aligned} \tag{15.23}$$

For highly coupled stiff systems of nonlinear ODEs, explicit Runge-Kutta methods become less desirable. For these systems, implicit generalizations of Runge-Kutta methods such as Rosenbrock methods are recommended. The Rosenbrock scheme uses the Jacobian matrix of the equations to propagate the system forward in time. In other words, the sensitivity of the solution's slope to changes in other species is considered rather than just the slope of the solution itself. One such second-order Rosenbrock method has been implemented in CellSim (Dekker and Verwer 2003):

$$\begin{aligned}
 (\mathbf{I} - \lambda \delta t \mathbf{J})\bar{k}_1 &= \dot{y}(y(t)) \\
 (\mathbf{I} - \lambda \delta t \mathbf{J})\bar{k}_1 &= \dot{y}\left(y(t) + \frac{k_1}{2}\delta t\right) - 2k_1 \\
 y(t + \delta t) &= y(t) + \frac{3}{2}k_1\delta t + \frac{1}{2}k_2\delta t
 \end{aligned} \tag{15.24}$$

with constant λ , identity matrix \mathbf{I} , and Jacobian \mathbf{J} .

4. Adaptive algorithms and error analysis

As many chemical systems exist as transients that rapidly equilibrate to steady state, it is natural to seek adaptive time-step algorithms. Stiff integrators such as fourth-order Rosenbrock methods have been highly successful in integrating purely kinetic systems. Inherent in such schemes is the need to calculate the Jacobian at each time step. For a single grid point with no diffusion, the Jacobian is a $N \times N$ square matrix, where N is the number of reactants in the simulation. Expanding this to an extended grid of multiple points (say, n grid points in any arrangement) and includ-

ing diffusion will generate a Jacobian consisting largely in the form of a $n \times n$ block matrix, each block itself an $N \times N$ matrix.

This matrix is highly sparse. The second-order diffusion operator would only involve nearest-neighbor grid points, leaving most matrix elements empty (as non-adjacent grid points are uncorrelated). Unfortunately, the size of the matrix is still cost prohibitive for performing the necessary LU decomposition required by adaptive Rosenbrock methods.

5. Spatiotemporal sensitivity analysis

Consider the following spatiotemporal biochemical system:

$$\frac{\partial \bar{\mathbf{C}}}{\partial t} = \bar{\mathbf{f}}(\bar{\mathbf{C}}, \bar{\mathbf{k}}, t) + \bar{\mathbf{D}} \cdot \nabla^2 \bar{\mathbf{C}} \quad (15.25)$$

where \mathbf{C} denotes N time dependent species concentrations, the kinetics component of the system is $\mathbf{f}(\mathbf{C}, \mathbf{k}, t)$ with parameters \mathbf{k} , and the corresponding diffusion component is $D\nabla^2\mathbf{C}$. The generalized sensitivity parameter

$$S_{i,j} = \frac{dC_i}{dk_j} \quad \begin{array}{l} i = 1, \dots, N \\ j = 1, \dots, M \end{array} \quad (15.26)$$

is then

$$\frac{d}{dt} \frac{dC_i}{dk_j} = \frac{dS_{i,j}}{dt} = \sum_{l=1}^N \frac{\partial f_l}{\partial k_l} \frac{\partial C_l}{\partial k_j} + \frac{\partial f_i}{\partial k_j} + \nabla^2 \frac{dC_i}{dk_j} \quad (15.27)$$

Applying operator splitting, it is clear from this equation that applying the diffusion propagator to the sensitivity parameters is sufficient to account for diffusion. The final term in equation 15.27 can be determined through simple finite differencing of the sensitivity parameters. However, the first two terms are more complicated. Our implementation is a Rosenbrock-based method that allows adaptive time steps to be incorporated into sensitivity calculations. In practice, two types of sensitivity parameters may be calculated within CellSim: sensitivity parameters with respect to k and parameters with respect to certain initial concentrations. As the latter is simpler to evaluate, we shall focus this discussion on fast evaluations of parameter-based sensitivities.

The reaction part of the previous equation may be rewritten as

$$\sum_l \mathbf{J}(i, l) \frac{\partial C_l}{\partial k_j} + \frac{\partial f_i}{\partial k_j} \quad (15.28)$$

where \mathbf{J} is the $N \times N$ Jacobian matrix. To propagate both the sensitivity parameters $S_{i,j}$ and the original set of species C_i , consider an extended biochemical system of $(M + 1)N$ equations:

$$\begin{pmatrix} \frac{dC_1}{dt} \\ \frac{dC_N}{dt} \\ \frac{dS_{1,1}}{dt} \\ \frac{dS_{N,1}}{dt} \\ \frac{dS_{1,M}}{dt} \\ \frac{dS_{N,M}}{dt} \end{pmatrix} \quad (15.29)$$

CellSim will integrate the coupled system using both standard integrators as well as Rosenbrock methods. For standard integrators, the propagator is reasonably simple to define. One needs only to generate the appropriate equations and consider the extended system. The Rosenbrock method requires the generation of an extended Jacobian of the new model system. This requires the automatic generation of the Hessian (second-order concentration derivatives) along with several other terms in the original system.

CellSim automatically generates these higher-order terms, and the computational expense of evaluating the extended Jacobian is mitigated by its sparsity and the ability to use large adaptive time steps, which reduce the number of required steps. To both illustrate this procedure and describe its implementation within CellSim, a small sample system is introduced. Consider a simple system with five species $C_1..C_5$:



CellSim will first automatically generate the following differential equations:

$$\begin{aligned} f_1 &= \frac{dC_1}{dt} = -k_1 C_1^2 C_2 + k_2 C_3 \\ f_2 &= \frac{dC_2}{dt} = -k_1 C_1^2 C_2 + k_2 C_3 \\ f_3 &= \frac{dC_3}{dt} = k_1 C_1^2 C_2 - k_2 C_3 \\ f_4 &= \frac{dC_4}{dt} = -k_3 C_4 + k_4 C_5 \\ f_5 &= \frac{dC_5}{dt} = k_3 C_4 - k_4 C_5 \end{aligned} \quad (15.31)$$

Because of the degeneracy of terms appearing in the differential equations, only four unique terms are generated and will be used during each elementary step of the reaction propagator. In this example, the terms are as follows:

$$\begin{aligned}t_1 &= k_1 C_1^2 C_2 \\t_2 &= k_2 C_3 \\t_3 &= k_3 C_4 \\t_4 &= k_4 C_5\end{aligned}\tag{15.32}$$

Hence, a system of differential equations may be considered simply a summation over precalculated terms. Two types of terms currently exist: one for passive transport channels (described in a later section) and one for mass action kinetics (termed a `kineticTerm`). A species is indexed by two integers: one for the compartment number r (row) and one for the species c (column) in that compartment. Internally, kinetic species are stored simply as:

```
class kineticSpeciesClass {
Public:
Int r,c;
...
}
```

The indexing of this term (r,c) is used to evaluate data structures to get information about the species—perhaps most importantly the current value (concentration) of that species. The actual concentrations are packed into a large contiguous memory array to minimize cache misses. Within `CellSim`, a kinetic term class has the following structure:

```
Class kineticTermClass: public genericTermClass {
Public:
svector <kineticTermClass> species;
svector <firstderivativeClass> firstderivativesforC;
svector <secondderivativeClass> secondderivativesforC;
firstderivativeforKClass firstderivativeforK;
secondderivativeforKClass;
secondderivativeforK;
double k;
svector <double> jacobianMultiplier;
}
```

A `svector` may be considered simply a standard STL vector that has been optimized for the purposes of `CellSim`. When using methods that require the Jacobian, the partials are all pre-generated and calculated once, minimizing the number of evaluations as well as taking advantage of the sparsity of the extended Jacobian (and other objects) that need to be built. The definitions for each term in the class are as follows.

- *species*: The species in a particular term. For instance, for the term t_1 the species list contains C_1, C_1, C_2 . C_1 is stored twice in the structure because C_1 exists as C_1^2 .
- *firstderivativesforC*: Stores the partial of this term with respect to all species within this term. For the term t_1 , this vector stores $\frac{\partial t_1}{\partial C_1}$ and $\frac{\partial t_1}{\partial C_2}$.
- *secondderivativesforC*: Contains all Hessian terms that are non-zero for this term. For the term t_1 , three terms are stored: $\frac{\partial^2 t_1}{\partial C_1 \partial C_1}, \frac{\partial^2 t_1}{\partial C_1 \partial C_2}, \frac{\partial^2 t_1}{\partial C_2 \partial C_2}$.
- *firstderivativeforK*: This partial stores the partial derivative of this term with respect to its own parameter. For the term t_1 , the only evaluated derivative is $\frac{\partial t_1}{\partial k_1}$.
- *secondderivativeforK*: This final partial derivative stores a vector of all mixed terms of the form $\frac{\partial^2 t}{\partial C_i \partial k}$. For t_1 , the following two terms are stored: $\frac{\partial^2 t}{\partial k_1 \partial C_1}$ and $\frac{\partial^2 t}{\partial k_1 \partial C_2}$.
- *k-rate constant of the kinetic term*: For the term t_1 , the kinetic constant is k_1 .
- *jacobianMultiplier*: A precalculated coefficient for partial derivatives with respect to C_i . Two values are stored for the term t_1 .

6. Evaluating the original jacobian

Before extending the system to the sensitivity parameters, we perform fast evaluation of the extended Jacobian of the fully coupled system. In our example, the original Jacobian is

$$\begin{pmatrix} -2k_1 C_1 C_2 & -k_1 C_1^2 & k_2 & 0 & 0 \\ -2k_1 C_1 C_2 & -k_1 C_1^2 & k_2 & 0 & 0 \\ 2k_1 C_1 C_2 & k_1 C_1^2 & -k_2 & 0 & 0 \\ 0 & 0 & 0 & -k_3 & k_4 \\ 0 & 0 & 0 & k_3 & -k_4 \end{pmatrix} \quad (15.33)$$

By pre-generating the appropriate terms by first evaluating *firstDerivativesForC* at a given time step, the Jacobian may be evaluated directly by taking the appropriate summation of the derivatives

$$\frac{\partial f_i}{\partial C_j} = \sum_k^L \frac{\partial t_{ik}}{\partial C_j} \quad (15.34)$$

where L is the number of terms for that equation i . The Jacobian itself is stored in a special sparse matrix class that only stores the non-zero elements for the calculation. For non-Rosenbrock integrators, the sparse matrix class allows CellSim to use

fast sparse matrix multiplies to evaluate the first term in the equation. The second part of the equation is precalculated in *firstderivativeforK*. By precalculating these terms, only derivatives requested by the user script are actually calculated. Once these terms have been evaluated, the right-hand side of the equation may be evaluated to propagate fully the reaction portion of the sensitivity parameters.

The propagation of sensitivity parameters using the method thus described works for classes of integrators such as Euler and Runge-Kutta but is not particularly suitable for stiff systems. For this reason, considerable time has been spent in implementing stiff integrators such as Rosenbrock for sensitivity parameters within CellSim.

7. Calculation of the extended Jacobian

From a computational standpoint, one may consider the propagation of the extended system as simply a new system with its own corresponding Jacobian, J_c . The structure of this Jacobian has a relatively simple block matrix form:

$$J_c = \begin{pmatrix} J & 0 & \dots & 0 \\ \mathbf{S}(1, 1 \dots M)_{1 \dots N} & J & & 0 \\ \mathbf{S}(N, 1 \dots M)_{1 \dots N} & 0 & & J \end{pmatrix} \quad (15.35)$$

$\mathbf{S}(i,j)_q$ is defined as $\frac{\partial S_{i,j}}{\partial C_q}$. $\mathbf{S}(i,j)_{1 \dots N}$ is defined as $\{\mathbf{S}(i,j)_1, \mathbf{S}(i,j)_2, \dots, \mathbf{S}(i,j)_N\}$ and $\mathbf{S}(i, 1 \dots M)_{1 \dots N}$ is defined $\{\mathbf{S}(i,1)_{1 \dots N}, \mathbf{S}(i,2)_{1 \dots N}, \dots, \mathbf{S}(i,M)_{1 \dots N}\}^T$. As J is already calculated, the only new terms that need to be calculated are the bottom-left-hand portion of J_c .

8. Numerical evaluation of J_c

An individual term in this portion of the block matrix may be written out in the following form:

$$\frac{\partial \left(\frac{dS_{i,j}}{dt} \right)}{\partial C_q} = \sum \frac{\partial^2 f_i}{\partial C_i \partial C_q} \frac{\partial C_i}{\partial k_j} + \frac{\partial^2 f_i}{\partial C_q \partial k_j} \quad (15.36)$$

The form of this equation is exactly like Equation 15.28 except that $\frac{\partial f_i}{\partial C_i}$ has been changed to $\frac{\partial^2 f_i}{\partial C_i \partial C_q}$ and there now exists a second term $\frac{\partial^2 f_i}{\partial C_q \partial k_j}$. Hence, the procedure is exactly the same as was used in calculating the original Jacobian J , except that now the previously defined *secondderivativeClass* is also precalculated

before each time-step. As this second class is also internal to the term, only non-zero terms are precalculated and used to fill in J_C . The $\frac{\partial^2 f_i}{\partial C_q \partial k_j}$ term is also precalculated in the same manner as before. In this case, only non-zero terms and requested terms by the user script are precalculated and used to fill in the final matrix. Throughout the calculation, sparse matrix classes are used to remove unnecessary matrix multiplies throughout the calculation.

III. CELLSIM: A CELLULAR SIMULATOR

The mathematics described has recently been implemented in the freely available software package CellSim, developed by our group under the Gnu Public License (GPL). This package is highly optimized for high-performance distributed computing platforms that use the message-passing interface (MPI) parallel-programming library (Pacheco 1996).

The distributed computing platform is particularly efficient for transport-coupled kinetics. The kinetic terms are essentially communication independent, as they depend only on the local concentrations of each species. Furthermore, as the computational cost of transport is much lower than the kinetic components, the system is essentially immune to communication overhead and may therefore be parallelized with near linear efficiency.

A. Compartmentalization

Currently, a finite difference scheme is used to determine cell geometry. The geometry is explicitly defined by the Cartesian grid. The set of compartments defined at a grid site determines which species may exist (or overlap) in a certain region of 3D space. The set of appropriate chemical reactions that may be defined at a grid point is automatically generated from the set of all possible chemical reactions and which species exist in which compartments.

From this information, the complete set of appropriate differential equations is automatically generated over the entire grid, which is optimized for each localized grid point in terms of storage and calculation. The natural boundary conditions for the system are periodic, but both Dirichlet and von Neumann boundary conditions (as well as more complicated boundary conditions) may be implemented through the appropriate use of localized chemical reactions.

B. Mpi parallelization

The explicit schemes described in the previous section allow CellSim to be parallelized with linear efficiency. A large simulation may be split into evenly sized blocks. As the reaction operator is communication independent, the only communication cost is on the surface of the blocks. As the computational cost of a block will scale as the number of grid points within the block (proportional to the volume of the

block) and the communication cost will go as the surface area of the block, a regime may always be found in which the communication cost of the system is negligible in comparison to the computational cost. Within this regime, one may naturally move to larger systems with near linear computational cost.

C. Downloading and compiling

A current version of the software may be downloaded via anonymous cvs from *sodium.physics.drexel.edu*. In a UNIX environment, first set the CVSROOT environmental variable as follows:

```
export CVSROOT = :pserver:anonymous@bio.physics.drexel.edu: \
/usr/local/cvs-repository
cvs login
```

Then execute the following command to retrieve the source code.

```
cvs co cellsim-src
```

The code may be compiled using standard make. We have developed a script titled *setup.sh*, which sets up standard compilers and optimization options in the source directory. Alternatively, one may further customize a build by predefining the following environmental variables in *setup.sh*. The machine-dependent compiler flags are all grouped under the optimization section for the compiler in *setup.sh* and clearly marked.

- *OPT*: Sets any optimization flags
- *CXX*: C++ compiler
- *CFLAGS*: Any compiler flags
- *LFLAGS*: Link flags
- *LD*: Linker
- *INCDIR*: Include directory

The CellSim code base is platform independent and has been compiled on Mac OS, Linux, and SunOS under a variety of different compilers, including both the Gnu compilers and the Intel high-performance compilers. It is necessary to have the freely available Gnu Scientific Library (GSL) installed on your system. If a parallel-enabled version of CellSim is desired, MPI must also be installed. The default environmental variables are set by running:

```
source setup.sh
```

An MPI-enabled version may be compiled by using:

```
source setup.sh 1
```

After the environmental variables have been set, the code may be compiled using the following commands:

```
make depend
make
```

To illustrate the use of CellSim, several canonical examples have been included that illustrate some of the features in the current version as well their use in spatiotemporal modeling.

D. Examples

The use of CellSim can be best described through the use of several biologically relevant examples that highlight some of the more salient features of the software suite. CellSim uses a command-line-based scripting interface that is executed as *cellsim file.input*, where *file.input* is the input script.

1. 2D Gray-Scott model of glycolysis

The first biological example is the celebrated Gray-Scott autocatalytic model of glycolysis, originally developed by Selkov (1968). All input file contents (including file names) are set to be case insensitive within CellSim. Comments may be incorporated into any input file using C, C++, or Perl comment styles. A typical input script looks as follows.

```
useReactions reactions.input;
useGrid grid.input;
useInitConcentrations initconc.input;
printOutput output1 {
  printinfo 1;
  printgrid 10 plot/U.10.plot U;
  printgrid 10 plot/V.10.plot V;
}
diffusionConstant all 1e-4;
diffusionConstant species U 2E-5;
diffusionConstant species V 1E-5;
printSysTime;
integrate Euler {
  dt = 1;
  dx = 0.009765625;
  runtime = 2000;
  runDiffusion;
  use output1;
}
printSysTime;
exit;
```

The main input file provides the names of all other necessary initialization files needed for a CellSim run. These files are required for the initialization of the 3D

spatial geometry of the simulation, initialization of species concentrations, and description of all appropriate chemical reactions possible among the species. The main input file (in this example, *file.input*) defines the simulation itself, providing definitions and instructions for printing and for integration.

The first three lines of *file.input* direct CellSim as to where to find definitions of the simulation reactions (*useReactions*), grid geometry (*useGrid*), and initial concentrations (*useInitConcentration*) of the reactants. These accessory files are described in more detail in the next section. Following the definition of the accessory files is a bracketed *printOutput* block that defines a print object named *output1*, which may be used for printing during any integration run. Multiple print objects may be defined and will only be executed with a corresponding *use* command within the integrator.

The print object command defined previously instructs CellSim to print a time stamp to the screen at each step via the *printinfo* command, and to print the entire grid content of species U and V to files every 10 steps via the *printgrid* command. After the print output command, the diffusion constants are defined for the reactants of interest via the *diffusionConstant* command. The first command uses the keyword *all* to set the default diffusion constant for all species, and the second and third commands set the diffusion constant individually for species U and V.

Following the definition of the diffusion constants, CellSim is instructed to integrate using the *integrate* command. In this particular case, CellSim is using the Euler method with a fixed time step $dt = 1$, a spatial resolution of $dx = 0.009765625$, and a runtime of 2000. Within the *integrate* command, *runDiffusion* ensures that diffusion is enabled. In addition, *use output1* instructs CellSim to use the print commands defined previously in the *printOutput* command. The *runtime* command sets the simulation time, which for this example is 2000 seconds.

The final instructions to CellSim are to again print out the actual system time (*printSysTime*) used for the simulation and then exit (*exit*). The user should make sure the units are all self-consistent. The chemical reactions defined within the model are located in the file *reactions.input* defined by the *useReactions* command.

In this simulation, the reactions are the Gray–Scott reactions (Gray and Scott 1983) (see Figure 15.1). A variant of the autocatalytic Selkov model of glycolysis, the Gray–Scott reactions are



This simple system produces a wide variety of spatiotemporal patterns sensitive to the reaction rates and diffusion constants. The reactions file for this simulation reads as follows.

```
locationlist {
location cytosol 1;
default cytosol;
}
```

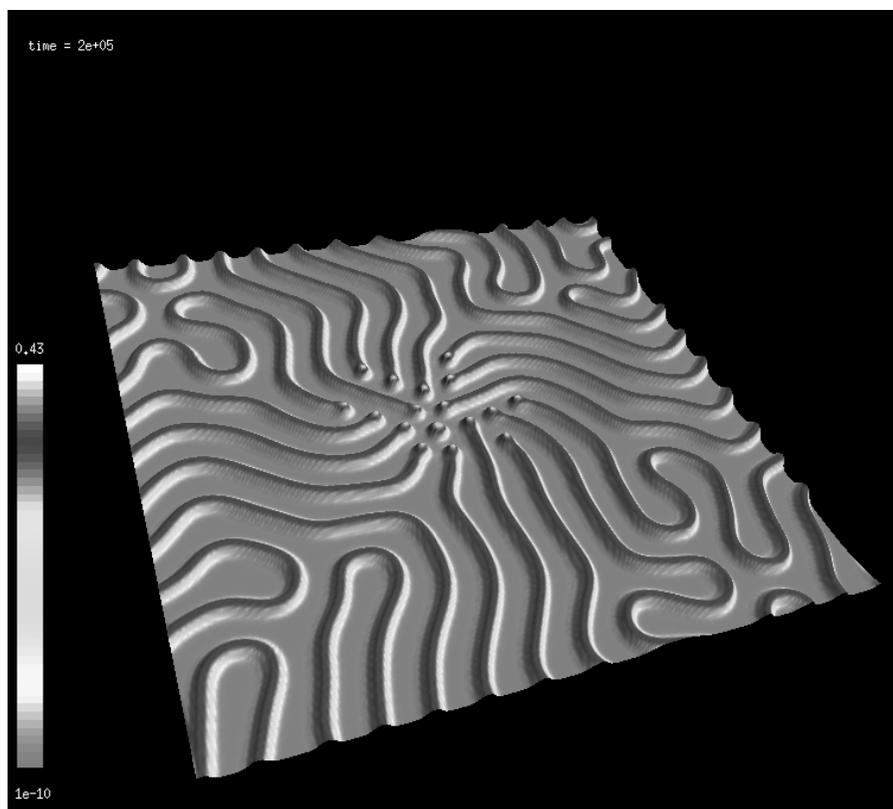


Figure 15.1. Gray-Scott model.

```
reactionlist {  
U + V + V -> V + V + V :: 1;  
V -> P :: 0.06;  
U -> bath :: 0.05;  
bath -> U :: 0.05;  
V -> bath :: 0.05;  
P -> bath :: 0.05;  
}
```

The *locationlist* block defines a single compartment *cytosol* and defines that compartment to have a total volume per unit grid of 1 liter. For spatiotemporal models, leave the volume unit as 1. The actual volume is then defined by the spatial geometry. The volume definition for a compartment may be used in mixed volume kinetic models in which the volume is not inherently defined by the spatial arrangement of the grid. The final command, *locationlist*, defines the default compartment as *cytosol* for reactants via the default command. In this model, all species exist in

the default compartment and thus do not need to be explicitly listed within the *locationlist* command.

The next block, *reactionlist*, defines the reactions in the Gray–Scott model. All reactions are nonreversible, with rate constants following double colons. For this simulation, the grid is simply a square plane of a single compartment. The grid file defined by the *useGrid* command takes the following form.

```
grid 48 48 1 ;
0 0 0 cytosol;
0 1 0 cytosol;
0 2 0 cytosol;
.
.
.
0 47 0 cytosol;
.
.
.
47 0 0 cytosol;
.
.
.
47 47 0 cytosol;
```

Here, the grid has dimensions of $48 \times 48 \times 1$ and all species are defined to exist in the compartment *cytosol*. A grid point may be defined to have any number of compartments. Grid points with multiple compartments may be considered interface regions, and reactions involving species of different compartments may additionally react within this interface region.

The initial conditions in this example consist of a grid containing two areas: a central square and the area surrounding it. The two species U and V initially exist in both areas at different concentrations. Their initial concentrations are perturbed randomly about some value. For CellSim, the initial concentrations can be specified for a species throughout all of its compartments or individually specified at each grid point. Using perturbed concentrations, the *initconc.input* file contains the following commands.

```
P = 0.0;
fixed bath = 1.0;
point 0 0 0 U = 1.00312899386658;
point 0 0 0 V = 0;
point 0 1 0 U = 1.00259570383182;
point 0 1 0 V = 0;
point 0 2 0 U = 0.996343013072222;
point 0 2 0 V = 0;
point 0 3 0 U = 0.990340706493643;
. . .
```

The first command defines the concentration of P to be 0 everywhere. The next command sets the bath to a concentration of 1.0 unit and fixes it so as not to change. Species U and V are perturbed at about 1.0 and 0.0, respectively, in the outer region (and 0.5 and 0.25, respectively, in the inner). All are listed individually at individual grid points. As with all CellSim files, the scripting language will override previous commands with any subsequent commands, and thus a default concentration may be set and then altered at specific grid points with the *point* command.

2. 3D Kinase phosphatase model

As an example of a fully 3D multi-compartment model, the next simulation models a simple signal transduction model of a plasma-membrane-bound receptor, cytosolic phosphatase, and a cytosolic kinase, which is activated at the cell surface in a spherical cell. This model was originally developed by Brown and co-workers (Brown and Kholodenko 1999). Extracellular stimulant S reacts with membrane-bound receptor R to produce S.R, which in turn phosphorylates kinase K to K* at the membrane. The species K* then diffuses inward to react with P inside the cell. After an initial transient stage, K* reaches a steady state of exponentially decreasing radial concentration (see Figure 15.2). The *reactions.input* script for this simulation is as follows.

```
locationList {
location extracellular 1 S;
location imembrane 1 R S.R S.R.K;
location cytosol 1;
default cytosol;
}

numberReactionList {
S + R <> S.R :: 4.2 0.25;
S.R + K <> S.R.K :: 1.2 0.8;
S.R.K -> K* + S.R :: 0.2;
K* + P <> K*.P :: 1.98 25 ;
K*.P -> K + P :: 6;
}
```

The *locationlist* command defines the compartment of each species. The stimulus S exists only in the extracellular region. The receptor and its intermediates are all on the intracellular membrane, and all other species are within the cytosol.

The *numberReactionList* block tells CellSim to read the contained equations in terms of quantity (in our case, micromoles) instead of quantity/volume (micromolar) concentration (used in *reactionList*). This option is useful when the rate constants are in terms of quantity and not concentration. The compartments for this simulation consist of a sphere of cytosol, with membrane overlapping the outermost edge of the cytosol region. At an edge of the cytosolic region, the grid input file for this simulation reads as follows.

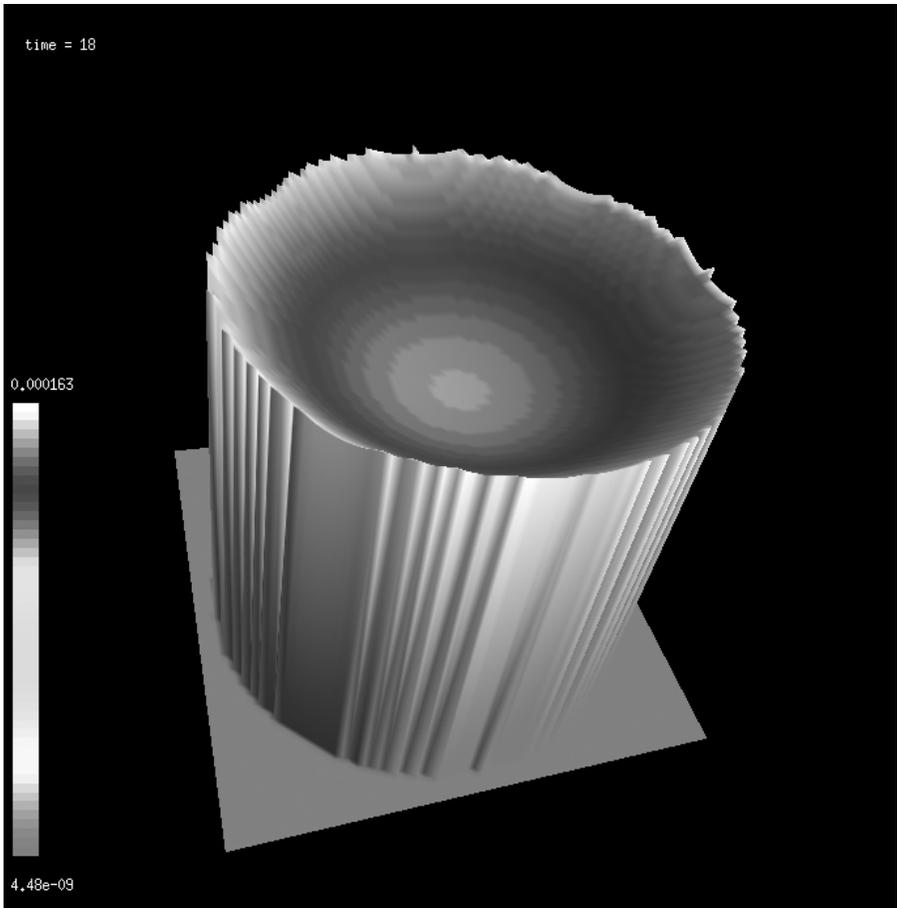


Figure 15.2. Application of CellSim, showing a 2D slice through center of 3D simple signal transduction cell model. The Z axis represents predicted concentration of a single cytosolic kinase (see color plate 9).

```

grid 81 81 81;
.
.
.
40 40 2 extracellular;
40 40 3 imembrane cytosol extracellular;
40 40 4 cytosol;
.
.
.
80 80 80 extracellular;

```

In the overlapping region, the cytosol, intracellular membrane, and extracellular area coexist. In this region, reactions involving the stimuli S and the receptor will occur, as well as reactions involving the stimulated receptor and the cytosolic kinase K . As an example of the adaptive integrators in CellSim, this simulation uses a second-order Strang split reaction-diffusion integration scheme, as follows.

```
useReactions reactions.input;
useGrid grid.input;
useInitConcentrations initialconc.input;
printOutput output1 {
  printinfo 10;
  printplane 10 plot/K.plane.plot K 40 * *;
  printplane 10 plot/K*.plane.plot K* 40 * *;
  printgrid 10 plot/K*.grid.plot K*;
  printgrid 10 plot/K.grid.plot K;
}

integrate arb2 {
  dtguess = 1e-2;
  dt = 1.0;
  dx = 0.009765625;
  runtime = 100;
  useStrang;
  runDiffusion;
  diffusionTolerance 1e-2;
  safety = 0.9;
  tolerance = 1e-2;
  dtmin = 1E-10;
  use output1;
}
```

The new print command *printplane* prints the plane specified by the x - y - z coordinates following the file name, where the integer coordinate specifies the constant plane through the grid and the asterisks define the direction of the plane. In this example, the simulation prints out the 81×81 plane of grid points defined by the equation $x = 40$. The integrate command *integrate arb2* instructs CellSim to integrate using a second-order adaptive Rosenbrock method. The previously unseen commands within the integrate block are specific to the adaptive integrator as follows.

- *dtguess*: The initial time step for the adaptive integrator.
- *dtmin*: The minimum time step allowed.
- *safety*: The maximum increase of a time step is internally set to 50%. This value sets the fraction (0–1) of the maximum increase that should be used.
- *tolerance*: Directs the adaptive method to choose the maximum time step that still achieves a given relative accuracy for the kinetics calculation. In our example, a relative accuracy of 0.01 is required.

- *diffusionTolerance*: Similar to the *tolerance* command, this command directs the adaptive step-doubling diffusion calculation to achieve the given relative accuracy.

3. Sensitivity analysis

Consider the system defined previously (equation 15.30). Suppose one would like to calculate $\frac{\partial C_1}{\partial k_1}$ and $\frac{\partial C_4}{\partial k_5}$. The reaction file for this system is defined in the same format as before, as follows.

```
locationList {
location cytosol 1;
default cytosol;
}

volumeToLiters = 1.0;
reactionList {
C1 + C1 + C2 <> C3 :: 1e-5 2e-5 k1 k2;
C4 <> C5 :: 1E-2 1E-3 k3 k4;
}
```

This definition has the optional labels k_1 , k_2 , k_3 , and k_4 appended to each reaction. In addition, CellSim has the ability to take the same label as part of multiple reactions. This is sometimes useful for parameter optimization in which several parameters are tied together and optimized as a single identity. Similarly, the initial concentration file has the same format as before, as follows.

```
C1 = 100;
C2 = 10;
C3 = 5;
C4 = 1;
C5 = 5;
```

In this example, a purely kinetic model is being used, and thus our grid file consists only of the following.

```
Grid 1 1 1;
0 0 0 cytosol;
```

Finally, the main scripting file must carry new instructions to define, calculate, and print the sensitivity parameters.

```
Use Reactions reactions.input;
useGrid grid.input;
useInitConcentrations initconc.input;
defineAnalyticalDerivative {
```

```
numerator = C1 C5;  
denominator = k1 k4;  
}  
printOutput output1 {  
printkinetics 10 screen C1;  
printsensitivity 1 screen C1 k1;  
printsensitivity 1 screen C5 k4;  
printsensitivity 1 dc1dk1 C1 k1;  
outputappendstring = analyticalDerivative;  
outputprependstring = plot/;  
}  
integrate arb4_sa {  
use output1;  
dt = 1E-5;  
safety = 0.9;  
dtmin = 1E-10;  
dtguess = 1E-5;  
tolerance = 1E-4;  
}  
runAnalyticalDerivative 100;
```

The new commands not previously seen begin with the `defineAnalyticalDerivative` command. Two required subcommands are *numerator* and *denominator*. The numerator must be followed by a list of defined species, whereas the denominator may be species or labeled rate constants. If species are used, sensitivity analysis with regard to the initial concentration of that species is carried out. All combinations of derivatives of the numerator and denominator are analytically evaluated throughout the sensitivity run.

The *outputappendstring* command optionally appends any printed files of derivatives by the argument string. Similarly, *outputprependstring* prepends the file name. In the example file, only a single derivative is being printed to a file whose name will be *plot/dc1dk1.analyticalderivative*. The *printOutput* command has a single new command named *printsensitivity*. Its format is similar to that of *printgrid* except that two parameters (*numerator*, *denominator*) are used to define the derivative to be printed.

In this case, two derivatives $\frac{\partial C_1}{\partial k_1}$, $\frac{\partial C_5}{\partial k_4}$ are printed to the screen every step, and simultaneously $\frac{\partial C_1}{\partial k_1}$ is being printed to a file named *plot/dc1dk1.analyticalderivative*. The next step defines a Rosenbrock integrator designed especially for sensitivity analysis. This new definition makes sure that enough memory is allocated for the extended Jacobian J_C . Finally, the *runAnalyticalDerivative* command tells CellSim to perform the calculation for 500 units of time. A full list of commands available in CellSim follows.

4. Parameter optimization

Experiments described in detail in the next section may be used to develop parameter sets for the chemical reactions, as well as diffusion constants within CellSim. These parameters are estimated by fitting the developed model to the available experimental data using a cost function, which represents the deviation of the model to data:

$$\sum_{\vec{x}, t} (C_{\text{exp}}(\vec{x}, t) - C_{\text{mdl}}(\vec{x}, \vec{k}, t))^2 \quad (15.38)$$

with adjustable parameters \vec{k} . Spatially resolved experiments $C_{\text{exp}}(\vec{x}, t)$ are taken at time shots t . One may choose to optimize the initial concentrations, rate constants, or any other free parameter in the model during the optimization process. A combined simulation–optimization approach has been implemented using fast Rosenbrock integrators. To illustrate this method, a simple example of a purely kinetic system will be given. As with all CellSim simulations, a standard command script file is needed, as follows.

```
useGrid grid.input;
useInitConcentrations initconc.all;
printOutput output1 {
  ...
}
integrate arb4 {
  ...
}
runOptimizer {
readData "1.1um.C1";
readData "2.1um.C1";
readData "3.1um.C1";
readData "4.10um.C1";
readData "5.10um.C1";
readData "6.10um.C1";
addParameterGlobal k1 1E-8 1E-3;
addParameterSingle C2 0.8 1.2 1 3;
addParameterSingle C2 0.8 1.2 4 6;
costType = sumsquare;
statisticsfile = mystats;
weighting 0.9 0.1;
anneal 10000 100 0.1 1.0 1E-3 1.1 0.0;
saveparameters = parameters.final;
bestfitsfile = bestfits.file;
}
```

We have skipped all standard sections and included only the new parts of the input file not previously seen. The new command *runOptimizer* tells CellSim to

perform the optimization procedure. First, each experiment is read in using the `readData` command. Each experiment has the same input format as initial concentration files, each prefaced with a `t = <number>` put together sequentially in a single file. Each experiment is internally read in and numbered ($1 \dots \text{Maximum Experiments}$). The new command `t = <number>` sets the time for the experimental data that follows that command. The `addParameterGlobal <k> <min> <max>` command tells the global optimization procedure to use all experiments to fit k between min and max , where k may be either a species or a labeled rate constant.

Similarly, the `addParameterSingle <k> <min> <max> <minexp> <maxexp>` command also fits k between min and max but will only use a subset of the experiments defined as those numbered between $minexp$ and $maxexp$. Several cost functions are available. In this case, sum of squares is set with the `costType` command. The `weighting` command gives the weighting of random trial moves. The first number dictates the fraction of times a random move involves initial concentrations, whereas the second number dictates the fraction of times a random number involves rate constants.

Statistics during the run are stored in the file defined by the `statisticsfile` command, and the best-fit parameters are saved in the file defined by the `bestfitsfile` command. The simulated annealing parameters are set through the `anneal` command. The parameters set the total number of moves, number of moves at each temperature, step size of random move, Boltzmann constant, initial temperature, factor by which the temperature decreases, and minimum temperature.

E. CellSim visualization: cellSimvis

The visualization component of CellSim is a separate program developed as a client/server package. This module should be considered beta at this time but is currently functional as a monitoring tool for large jobs as well as for visualization of generated data. CellSimVis is based on the freely available GPL licensed QT widget set from Trolltech (used to develop the popular UNIX environment KDE on the GNU/Linux platform). To compile the GUI, OpenGL and the QT development libraries must be installed. CellSimVis (see Figure 15.3) may be downloaded and compiled as follows.

Step 1: Log in to the anonymous cvs server as before, using:

```
export CVSROOT = :pserver:anonymous@bio.physics.drexel.  
edu:/usr/local/cvs-repository  
cvs login
```

Step 2: Check out the source code for the GUI with the command:

```
cvs co cellsim-gui
```

Step 3: Run the following commands.

```
qmake  
Make
```

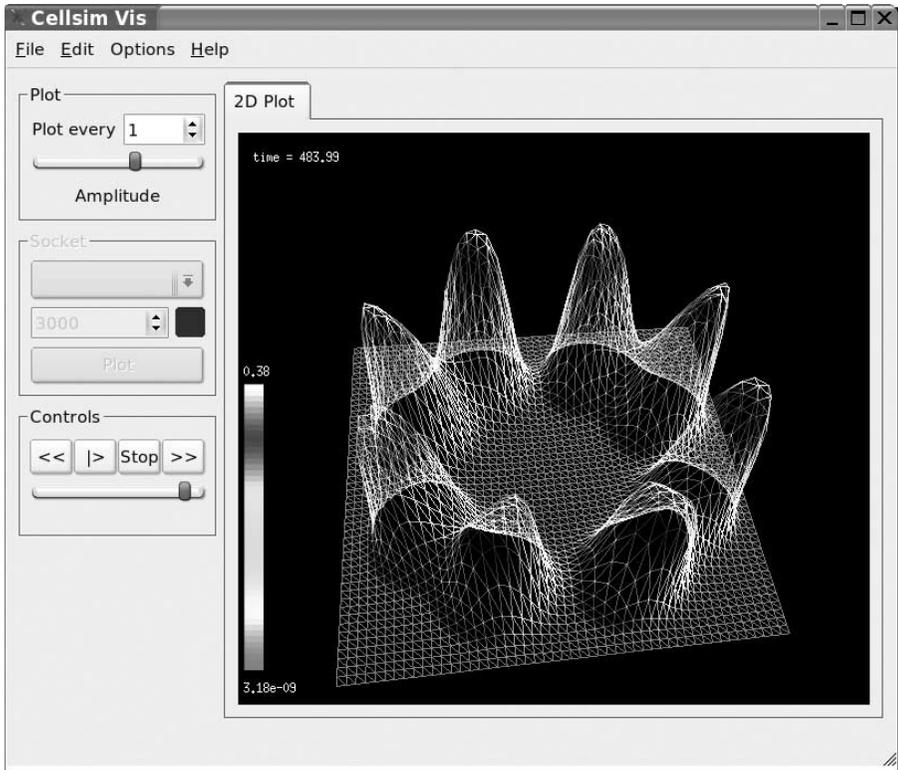


Figure 15.3. CellSimVis visualization engine (see color plate 10).

The current version of CellSimVis has the following features.

- Export of rendered images for publication (PNG format)
- Export of MPG-based movie files
- Import of CellSim Plot files
- Socket-based monitoring of CellSim. Interactive switching of exported species from CellSim
- OpenGL-based 3D contour plots rendered as 2D grid (solid surface, line mesh, or point based).
- Rendered visualization of surface normals
- Automatic scaling of model

Imaging of data in one, two, and three dimensions is available within CellSimVis. Simple plotting (concentration versus time) is hardware rendered using OpenGL primitives. For 2D data (versus time), visualization is implemented both as a simple 2D color contour plot and as a rendered 3D plot of the data with the height representing the concentration on the plane (rendered in real time as a movie). For reading from saved data files, a slider is available for data examination that

enables the user to track the changes in concentration with time. An additional 3D-rendering procedure based on isocontours using marching cubes will be included in a future version of CellSimVis.

Plot files generated by CellSim may easily be imported from the menu (File > Open). For socket-based communication, File > Sockets should be used. Specify the host name and socket for the machine running CellSim. In this mode, CellSimVis will automatically import all available species in the simulation and make them available for rendering. Snapshots of the visualization can be saved as PNG files.

IV. SPATIOTEMPORAL IMAGING

In recent years, high-resolution single-cell imaging has been recognized as a most favorable way of looking at biology (Cole et al. 2003). Cytomics, as described by Valet et al. in this edition, aims to provide cellular information by executing imaging in a high-throughput high-content fashion. This information can be used to classify cells, identify molecular hotspots, and carry out statistical correlation across levels of biological hierarchy. Cytomics approaches can also be applied in functional genomics research to characterize the location of proteins (Murphy 2004) and sub-cellular phenotypes specific to RNAi knockouts in high-throughput assays (Conrad et al. 2004).

These screening technologies provide end-points for a precise description and classification of cells and subcellular phenotypes, and a framework for ST systems biology. As an example, basic morphological properties of cells have been used to increase the realism of computational models (Schoeberl et al. 2002). However, in view of the complex cellular machineries being investigated the goal would be to perform both a time-resolved and multiplexed analysis at high 3D resolution, within spatially distinguishable compartments in single cells. Fluorescence confocal microscopy is ideal for performing these tasks.

Confocal microscopes come in different flavors, but they all have in common the application of non-invasive optical sectioning at low radiation damage, which is ideal for studying structural and functional properties of living cells at full 3D microscopic resolution. Specific experimental perturbations can be introduced and subsequently monitored, and the quantified cellular behavior can be used to classify distinct cellular phenotypes or phenomes (see the chapter by Parvin et al.). However, tagging cellular structures and species with multiple fluorescent dyes are limiting factors, although a wealth of non-invasive fluorescent probes has been developed and is now available for monitoring membranes and cell compartments, as well as specific protein targets in living specimen (DeBernadi et al. 2005).

As a specific example, for the study of signaling pathways one would require quantification of the localization, concentration, dynamics, interaction, and activation (phosphorylation) status of many components involved in the cascade. Newly developed fluorescent technologies move this field forward, such as quantum (Q-dot) dot nanoparticles that are ideal for imaging both localization and concen-

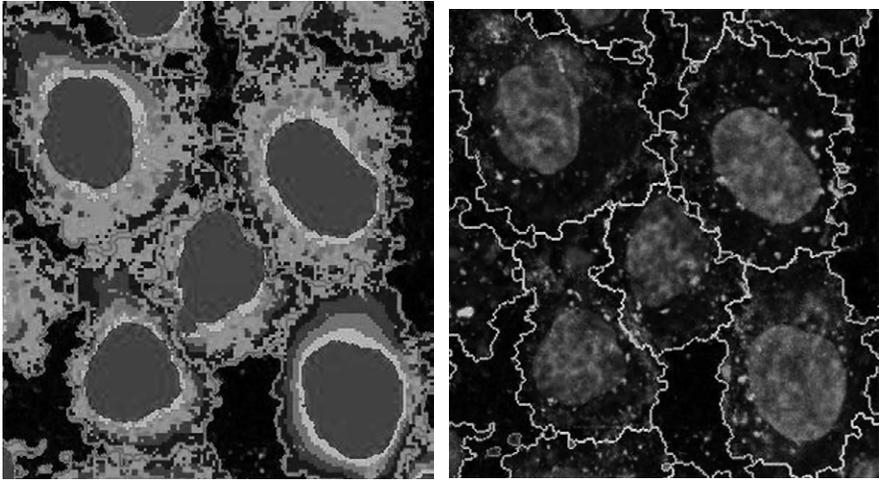


Figure 15.4. Computational imaging delivers quantitative description of the internalization of EGFR, activated by a biotinylated EGF/streptavidin quantum dot complex (green) with A431 cells. Transport routes of internalization can be monitored by *in vivo* imaging, as well as concentration increase over the time of the experiment (left to right). Concentrations of q-dots within equidistant zones of the cytosol (right) of many cells deliver information on averages related to dynamical processes that feed ST-systems biology (see color plate 11).

tration of target proteins (Smith and Nie 2004). In a study of EGFR internalization, an average of 30,000 internalized receptors could be monitored at the single-cell level, and subsequent image analysis provided regional average information about concentration and active endosomal transport (see Figure 15.4).

A distinct advantage of Q-dot assays is in their capability to be multiplexed, but they are currently limited in their ability to provide protein interactions and status of protein activation. Diffusion processes can be measured by fluorescence recovery after photo-bleaching (FRAP) and fluorescence correlation spectroscopy (FCS) (Lippincott-Schwartz et al. 2001). GFP fusion proteins are ideal for FRAP, in that they can be bleached without detectable damage of cells. With these tools, differences in the diffusion constant D due to membrane association, scaffolding, and compartmentalization can be measured.

To detect protein associations in the 1- to 10-nm range, fluorescence resonance energy transfer (FRET) is the preferred imaging technique (Jares-Erijman and Jovin 2003). In conjunction with radiometric sensors such as EGFR-ECFP and PTB-EYFP in one molecule, FRET can be used to monitor phosphorylation dynamics (Offterdinger et al. 2004). Both FRAP- and FRET-related technologies are currently limited in monitoring multiple species simultaneously. As these fields progress, they will determine the realism of comprehensive spatiotemporal models of regulation in signaling networks, nuclear processes, and morphogenesis.

V. CONCLUSIONS

A biological cell is a complex environment for chemical reactions, with a vast and diverse collection of active and passive transport mechanisms, membrane surfaces, and compartments. A new generation of microscopic-imaging techniques capable of the real-time tracking of single molecules in living cells provides visible evidence of the biological significance of process dynamically evolving in both space and time. Computer simulation of physics-based models, coupled with quantitative spatiotemporal data, will allow cell biologists to rigorously develop and test complex hypotheses. Although methods of simulating reaction-diffusion systems have been successfully applied to complex physical systems such as the atmosphere, oceans, and engine combustion, cells present an unprecedented complexity of significant molecular species and transport mechanisms, and a continuing challenge for experimental measurement.

The relatively small size of the cell also presents a challenge, as many relevant processes occur on atomistic scales that are unsuitable for the continuous deterministic approach described in this chapter. However, remarkably, cell-imaging data suggest that a variety of cell processes are amenable to a reaction-transport model, and the number of proteins per cell generally range from several hundred to hundreds of thousands of each species, supporting the use of molecular concentrations. To address biological problems for which discrete stochastic approaches are more suitable, several stochastic simulation methods have been proposed (reviewed in Turner et al. (2004)). Regardless of the algorithm used, it is necessary to develop tools to interpret simulation results, including efficient sensitivity analysis and interactive simple interfaces. The emergence of quantitative techniques in cell biology is ushering in an era of “predictive” biology and medicine, when experiments and computer simulation will be blended to help study disease mechanisms and identify therapeutic targets.

ACKNOWLEDGMENTS

We would like to acknowledge the many fruitful discussions with E. Papazoglou and B. Kholodenko, who were instrumental in the development of CellSim.

RECOMMENDED RESOURCES

We highly recommend the review article by Slepchenko et al. (2002), which offers an extensive review of spatiotemporal systems with an emphasis on Virtual Cell. We also highly recommend Dekker and Verwer’s recently published book on numerical methods for advection-reaction-diffusion equations, which has been tremendously useful in developing the theory and computational aspects behind CellSim (Dekker and Verwer 2003). Stochastic approaches for cell simulation are

comprehensively reviewed by Turner et al. (2004). The following are selected web resources for cell simulation.

<http://systemsbiology.physics.drexel.edu> CellSim is currently hosted at this address. A full set of resources (including help files, source code, updates and other utilities) will be made available as the tools are developed.

www.vcell.org The Virtual Cell site is a freely usable tool for spatiotemporal modeling. This framework for reaction-diffusion modeling has been developed as a national resource. Computational facilities are available directly through the National Resource for Cell Analysis and Modeling (NRCAM) to allow remote simulation through Virtual Cell to academic groups.

<http://biodynamics.indiana.edu/CellModeling/AboutCellX.html> CellX is a cell dynamics simulator based on 2D and 3D reaction transport simulation currently under development.

www.sbml.org There are efforts underway to provide unified computational coding platforms, including the Systems Biology Markup Language at this address. This standard specification has a freely available Lesser Gnu Public License (LGPL) library available for systems biology software projects. SBML specifications levels 1 and 2 currently treat compartments kinetically and thus cannot be used for explicit spatiotemporal modeling discussed in this chapter. However, explicit spatial geometry and diffusive properties are currently under consideration for the forthcoming SBML level 3 specification. Standardization of a unified markup language for spatiotemporal models through efforts such as SBML is needed to further develop this field.

<http://icb.med.cornell.edu/crt/SigPath/index.xml> SigPath is an XML-based information system designed for signaling pathways and networks within cells.

www.mcell.psc.edu/ MCell is a Monte Carlo spatiotemporal simulation on the cellular scale. SmartCell is a general framework for the modeling and simulation of diffusion-reaction networks on a mesoscopic scale using stochastic reaction models (Coombes et al. 2004). Similar models include Stochsim and MesoRD.

REFERENCES

- Arkin, A., Ross, J., et al. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**(4):1633–1648.
- Arrio-Dupont, M., Foucault, G., et al. (2000). Translational diffusion of globular proteins in the cytoplasm of cultured muscle cells. *Biophys. J.* **78**(2):901–907.
- Belenkaya, T. Y., Han, C., et al. (2004). *Drosophila* Dpp morphogen movement is independent of dynamin-mediated endocytosis but regulated by the glypican members of heparan sulfate proteoglycans. *Cell* **119**(2):231–244.
- Borman, G., Brosens, F., and DeSchutter, E. (2004). Modeling Molecular Diffusion *Computational Modeling of Genetic Biochemical Networks* 189 MIT Press NY.
- Brown, G. C., and Kholodenko, B. N. (1999). Spatial gradients of cellular phospho-proteins. *FEBS Lett.* **457**(3):452–454.
- Brown, R. (1827). A brief account of microscopical observations. (unpublished).

- Cole, M. J., Pirity, M., et al. (2003). Shedding light on bioscience. Symposium on Optical Imaging: Applications to Biology and Medicine. *EMBO Rep.* **4**(9):838–843.
- Coombes, S., Hinch, R., et al. (2004). Receptors, sparks and waves in a fire-diffuse-fire framework for calcium release. *Prog. Biophys. Mol. Biol.* **85**(2/-3):197–216.
- Crank, J., and Nicolson, P. (1947). A Practical Method for Numerical Integration of Solution of Differential Equations of Heat-Conduction Type. *Proc. Camb. Philos. Soc.* **43**(50):51–67.
- DeBernardi, M. A., and Brooker, G. (1998). "Simultaneous fluorescence ratio imaging of cyclic AMP and calcium kinetics in single living cells." *Adv. Second Messenger Phosphoprotein Res.* **32**:195–213.
- Dekker, K., and Verwer, J. (2003). *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. New York: Springer-Verlag.
- Doiron, B., Lindner, B., et al. (2004). Oscillatory activity in electrosensory neurons increases with the spatial correlation of the stochastic input stimulus. *Phys. Rev. Lett.* **93**(4):48–101.
- Douglas, J. (1962). Alternating Direction Methods for Three Space Variables. *Numerische Mathematik*, **4**(41):41–63.
- Dove, A. (2003). Screening for content the evolution of high throughput. *Nature Biotech.* **21**(8):859–864.
- Dunkel, J., Hilbert, S., et al. (2004). Stochastic resonance in biological nonlinear evolution models. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **69**(5/2):56–118.
- Elowitz, M. B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* **403**:335–338.
- Gibson, M. A., and Bruck, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A* **104**(9):1876–1889.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Computat. Phys.* **22**:403–434.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Genetics* **81**:2340–2361.
- Gray, P., and Scott, S. K. (1983). Autocatalytic reactions in the isothermal continuous stirred tank reactor: Isolas and other forms of multistability. *Chem. Eng. Sci.* **38**(1):29–43.
- Holdaway-Clarke, T. L., Feijo, J. A., et al. (1997). Pollen tube growth and the intracellular cytosolic calcium gradient oscillate in phase while extracellular calcium influx is delayed. *Plant Cell* **9**(11):1999–2010.
- Jares-Erijman, E. A., and Jovin, T. M. (2003). "FRET imaging." *Nat. Biotech.* **21**(11):1387–1396.
- Jung, P. J., and Mayer-Kress, G. (1995a). Noise controlled spiral growth in excitable media. *Chaos* **5**(2):458–462.
- Jung, P. J., and Mayer-Kress, G. (1995b). Spatiotemporal stochastic resonance in excitable media. *Physical Review Letters* **74**(11):2130–2133.
- Kholodenko, B. N. (2003). Four-dimensional organization of protein kinase signaling cascades: The roles of diffusion, endocytosis and molecular motors. *J. Exp. Biol.* **206**(12):2073–2082.
- Kholodenko, B. N., Brown, G. C., et al. (2000). Diffusion control of protein phosphorylation in signal transduction pathways. *Biochem. J.* **350**(3):901–907.
- Khurana, S., Kreydiyyeh, S., et al. (1996). Asymmetric signal transduction in polarized ileal Na(+)-absorbing cells: Carbachol activates brush-border but not basolateral-membrane PIP2-PLC and translocates PLC-gamma 1 only to the brush border. *Biochem. J.* **313**(2):509–518.
- Kiss, I. Z., Hudson, J. L., et al. (2004). Noise-aided synchronization of coupled chaotic electrochemical oscillators. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **70**(2/2):26–210.

- Lam, H., Matroule, J. Y., et al. (2003). The asymmetric spatial distribution of bacterial signal transduction proteins coordinates cell cycle events. *Dev. Cell* **5**(1):149–159.
- Lippincott-Schwartz, J., Snapp, E., and Kenworthy, A. (2001). Studying protein dynamics in living cells. *Nat. Reviews Molecular Cell Biology* **2**:444–456.
- Lukkien, J. J., Segers, J. P. L., et al. (1998). Efficient Monte Carlo methods for the simulation of catalytic surface reactions. *Physical Review E* **58**(2):2598–2610.
- McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94**(3):814–819.
- McAdams, H. H., and Arkin, A. (1998). Simulation of prokaryotic genetic circuits. *Ann. Rev. Biophys. Biomol. Struct.* **27**:199–224.
- McAdams, H. H., and Arkin, A. (1999). It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**(2):65–69.
- Meyer, T., and Tereul, M. N. (2003). Fluorescence imaging of signalling networks. *TRENDS in Cell Biology* **13**:101–106.
- Murphy, R. F. (2004). "Automated interpretation of protein subcellular location patterns: implications for early cancer detection and assessment. *Ann N Y Acad Sci.* May(1020): 124–131.
- Offterdinger, M., Georget, V., Girod, A., and Bastians, P. I. H. (2004). "Imaging phosphorylation dynamics of the epidermal growth factor receptor." *J. Biological Chemistry* **279**(35):36972–36981.
- Pacheco, P. (1996). *Parallel Programming with MPI*. San Francisco: Morgan Kaufmann.
- Peletier, M. A., Westerhoff, H. V., et al. (2003). Control of spatially heterogeneous and time-varying cellular reaction networks: A new summation law. *J. Theo. Biol.* **225**(4):477–487.
- Press, W. H. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, UK/New York: Cambridge University Press.
- Purich, D. (ed.) (2004). *Enzymatic Kinetics and Mechanism: Detection and Characterization of Enzyme Reaction Intermediates*. New York: Academic Press.
- Schoeberl, B., Eichler-Jonsson, C., et al. (2002). Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* **20**(4):370–375.
- Selkov, E. (1968). On the mechanism of single-frequency self-oscillations in glycolysis: A Simple Kinetic Model. *Eur. J. Biochem.* **4**(1):79–86.
- Singer, M. A., and Pope, S. B. (2004). Exploiting ISAT to solve the reaction-diffusion equation. *Combustion Theory and Modeling* **8**(2):361–383.
- Slepchenko, B. M., Schaff, J. C., et al. (2002). Computational cell biology: Spatiotemporal simulation of cellular events. *Ann. Rev. Biophys. Biomol. Struct.* **31**:423–441.
- Smith, A. M., and Nie, S. (2004). "Chemical analysis and cellular imaging with quantum dots." *Analyst.* Aug;**129**(8):672–677.
- Sportisse, B. (2000). An analysis of operator splitting techniques in the stiff case. *Journal of Computational Physics* **161**(1):140–168.
- Strang, G. (1968). On the Construction and Comparison Difference Schemes *SIAM Journal on Numerical Analysis* **5**(506):506–517.
- Taylor, D. L., Woo, E. S., and Giuliano (2001). Real-time molecular and cellular analysis: the new frontier of drug discovery. *Current Opinions in Biotechnology* **12**:75–81.
- Turner, T. E., Schnell, S., et al. (2004). Stochastic approaches for modeling *in vivo* reactions. *Computational Biology and Chemistry* **28**(3):165–178.
- Vesely, F. J. (2001). *Computational Physics: -An Introduction.* (2d ed.). New York/-London: Kluwer Academic/Plenum Publishers.

Cytomics: From Cell States to Predictive Medicine

G. Valet*, **R. F. Murphy****, **J. P. Robinson[†]**,
A. Tarnok[‡], and **A. Kriete[§]**

**Max-Planck-Institut für Biochemie, Martinsried, Germany; **Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; †Purdue University, Lafayette, Indiana, USA; ‡Department of Pediatric Cardiology, Heart Center, University Hospital Leipzig, Germany; §Drexel University, Philadelphia, Pennsylvania and Coriell Institute, Camden, New Jersey, USA*

Chapter 16

ABSTRACT

Cytomics, the systematic study of biological organization and behavior at the cellular level, has developed out of computational imaging and flow cytometry and promises to provide essential data for systems biology. The ability to perform high-content and high-throughput imaging and analysis to reveal complex cellular phenotypes will not only further our understanding of how cells and tissues carry out their functions but also provide insight into the mechanisms by which these functions are disrupted.

Advances in flow, chemical, and tissue cytometry extend the applicability of cytomics to tissues, cytological smears, and blood and other body fluids. As such, cytomics not only provides a new framework for spatiotemporal systems biology but enriches personalized or individualized medicine. This can take the form of individual disease course predictions for therapy selection purposes as well as identification of discriminatory bio-parameter patterns.

I. INTRODUCTION

Systems biology aims at the understanding of the integral functionality of single cells, organs, or organisms by molecular analysis and mathematical modeling (Ideker et al. 2001; Kitano et al. 2002; Hood et al. 2003, 2004). This task is significantly more complex for organisms than for single-cell systems such as bacteria or yeast. Organismic complexity derives from the diversity of genotypes among individuals, via variable exposure histories to environmental influences in numerous specialized organs. Cell states are characterized by significant internal hetero-

geneity according to cell cycle, functional status, size, and molecule content—with a heritable variation in the baseline gene expression (Morley et al. 2004). This inherent variability may be constrained by a limited number of cell states or fates, as described in the contribution of Huang.

Cytomics, the multimolecular quantitative analysis of the heterogeneity of cells and cell systems (cytomes), in combination with exhaustive bioinformatics knowledge extraction from analysis results (Valet 2002; Chitty 2005), aims to provide comprehensive, accurate, and systematic data. These qualities have been defined as the cornerstones for measurement technologies in systems biology (Kitano 2001). High-content and high-throughput methodologies are essential characteristics of cytomics for both single cells and tissues (Ecker et al. 2005, Boyce et al. 2005).

Currently, the concept of cytomics profits from advances in areas such as location proteomics, flow and tissue cytometry, screening assays, and cell and tissue arrays. Such advances move us toward a broad, systematic collection of information for clustering and cataloging cells according to their molecular, organelle, and morphometric phenotypes. Cataloging cell states by assessing a wide spectrum of quantities, which may be seen as state variables, is not necessarily driven by particular hypotheses—a property cytomics shares with other “-omics” methodologies. Realization of this concept have been successfully applied to the generation of profiles of drug activity, using a hypothesis-free molecular cytology (Perlman et al. 2004) and signaling network analysis (Sachs et al. 2005).

This chapter reviews various approaches in basic biological research and medicine for generating quantitative, flow, and image-based data for a comprehensive profiling and structural state space analysis. Analysis of changes of the cellular phenotype due to specific experimental perturbations are reviewed elsewhere in this book. Cytomics-related image analysis of subcellular protein distributions, compartments, cells, and tissues is in many areas specific to the imaging technology employed. However, for data mining statistical tools well known in bioinformatics are employed to classify and subsequently catalog cell states, whereas statistical correlations can span across levels of biological organizations.

As an example, cytomics data may be correlated with gene expression data to identify significant molecular markers, but may also enable creation of a bridge between cellular phenotypes and emerging physiological processes in the sense of an integrated physiology approach. Furthermore, cytomics provides a framework for the development of computational models of cells in support of a spatio-temporal systems biology.

II. COMPUTATIONAL IMAGING IN CYTOMICS

A. Single-cell image analysis

One of the most important outcomes of the Human Genome Project is the realization that there is considerably more biocomplexity in the genome and the pro-

teome than previously appreciated (Herbert 2004). Not only are there many splice variants of each gene system, but some proteins can function in entirely different ways (in different cells and in different locations of the same cell), lending additional importance to the single-cell analysis of laser scanning cytometry and confocal microscopy. These differences would be lost in the mass spectroscopy of heterogeneous cell populations. Hence, cytomics approaches may be critical to the understanding of cellular and tissue functions.

Fluorescence microscopy represents a powerful technology for stoichiometric single-cell-based analysis in smears or tissue sections. Whereas in the past the major goal of microscopy and imaging was to produce high-quality images of cells, in recent years an increasing demand for quantitative and reproducible microscopic analysis has arisen. This demand came largely from the drug discovery companies, but also from clinical laboratories. Slide-based cytometry is an appropriate approach for fulfilling this demand (Tarnok and Gerstner 2002). Laser scanning cytometry (Gerstner et al. 2002; Tarnok and Gerstner 2002; Megason et al. 2003) was the first of this type of instrument to become commercially available, but today several different instruments are on the market (Jager et al. 2003; Molnar et al. 2003; Schilb et al. 2004).

These types of instruments are built around scanning fluorescence microscopes that are equipped with either a laser (Tarnok and Gerstner 2002; Schilb et al. 2004) or a mercury arc lamp as the light source (Bajaj et al. 2000; Molnar et al. 2003). The generated images are processed by appropriate software algorithms to produce data similar to flow cytometry. Slide-based cytometry systems are intended to be high-throughput instruments, although at present they have a lower throughput than flow cytometers. These instruments allow multicolor measurements of high complexity (Gerstner et al. 2002; Ecker and Steiner 2004) comparable to or exceeding that of flow cytometers.

A substantial advantage over flow cytometry is that cells in adherent cell cultures and tissues can be analyzed without prior disintegration (Smolle et al. 2002; Kriete et al. 2003; Ecker et al. 2004; Gerstner et al. 2004). In addition, due to the fixed position of the cells on the slide or in the culture chamber cells can be relocated several times and reanalyzed. Even restaining and subsequent reanalysis of each individual cell is feasible. Because a high information density on the morphological and molecular pattern of single cells can be acquired by slide-based cytometry, it is an ideal technology for cytomics.

Although at present not realized, the information density per cell can be increased further by implementing technologies such as spectral imaging (Ecker et al. 2004), confocal cytometry (Pawley 1995), fluorescence resonance energy transfer (FRET) (Jares-Erijman and Jovin 2003; Ecker et al. 2004; Peter and Ameer-Beg 2004), near-infrared Raman spectroscopy (Crow et al. 2004), fluorescence lifetime imaging (FLIM) (Murata et al. 2000; Peter and Ameer-Beg 2004), optical coherence tomography (Boppart et al. 1998), spectroscopic optical coherence tomography (Xu et al. 2004), and second harmonic imaging (Campagnola et al. 2003). All of these technologies mark the progress in optical bio-imaging.

In the future, developments in imaging resulting from a family of concepts that allows image acquisition far beyond the resolution limit (down to the nm range) are expected. These include multiphoton excitation (Manconi et al. 2003), ultrasensitive fluorescence microscopes (Hesse et al. 2004), stimulated emission depletion (STED) microscopy (Hell 2003), spectral distance microscopy (Esa et al. 2000), atomic force microscopy (AFM) and scanning near-field optical microscopy (SNOM) (Rieti et al. 2004), and image restoration techniques (Holmes and Liu 1992). Using laser ablation in combination with imaging, even thick tissue specimens can be analyzed on a cell-by-cell basis (Tsai et al. 2003).

B. Innovative preparation and labeling techniques

Biomolecular analysis techniques such as bead arrays (Lund-Johansen et al. 2000; Tarnok et al. 2003), layered expression imaging (Englert et al. 2000), single-cell polymerase chain reaction (PCR) (Taylor et al. 2004), tyramide signal amplification (Freedman and Maddox 2001), biomolecule labeling by quantum dots (Parak et al. 2003), magnetic nanobeads (McCloskey et al. 2003), and aptamers (Ulrich 2004) open new horizons of sensitivity, molecular specificity, and multiplexed analysis. With additional tools—such as laser microdissection (Taylor et al. 2004), laser catapulting (Burgemeister et al. 2003), and fast electric single cell lysis (Han et al. 2003)—single cells can be rapidly isolated and further subjected to genomic or proteomic analysis (Burgemeister et al. 2003; McClain et al. 2003; Taylor et al. 2004) or single-cell capillary electrophoresis (Han et al. 2003).

The dimensionality of measured molecular cell data can be substantial, especially when repeated six- or eight-color staining protocols are performed on many different cell populations (Lenz et al. 2003; Ecker et al. 2004; Mittag et al. 2005) and their spatial interrelationships within a tissue are taken into account (Smolle et al. 2002; Ecker and Steiner 2004; Gerstner et al. 2004). The data density is multiplied if high-density single-cell analysis such as SNOM, AFM (Rieti et al. 2004), and STED (Hell 2003)—combined with single-cell genomics (Burgemeister et al. 2003; Taylor et al. 2004) or proteomics—(Han et al. 2003; McClain et al. 2003) is added.

A highly multiplexed yet hypothetical model for cytomic analysis of biological specimens could work as follows. Viable cells may be initially stained for cell functions (e.g., intracellular pH, transmembrane potential, intracellular Ca^{2+}), followed by fixation to remove the functional stains and restaining for specific extra- or intracellular constituents such as antigens, lipids, or carbohydrates, including, specific nucleic acids. Serial optical analysis will permit for every individual cell the 3D-reconstruction of its exact localization within the network of other cells in a tissue, together with the molecular morphology of its cell membrane, nucleus, organelles, and cytoplasm (including the parameterization of 3D shapes).

Serial histological sections taking stereological aspects of tissue architecture into account (Mandarim-de-Lacerda 2003) could serve as a basis for the standardized analysis of proximity and interaction patterns for intracellular structures such as

nucleus and organelles, as well as for different cell types within the tissue architecture (which can even include time as a parameter for 4D intravital microscopy [Mempel et al. 2004]). Microscopic image capture and analysis systems using their spatial relocation capacities will increasingly permit such staining sequences. Further genomic and proteomic characterization of single cells will yield substantial input into our understanding of cell development and function in the histological context, as further outlined in the section following.

C. Location proteomics

Systems biology researchers seek to build accurate predictive models of complex biological systems, typically incorporating information about events involving different types of biological macromolecules and occurring on different length and time scales. This requires the creation of systematic frameworks for representing this information and large-scale projects to acquire it (creating the “parts lists” for building models). A critical requirement for the success of such large-scale projects is being able to automate not only sampling, specimen preparation, and data collection but also data analysis.

1. Automated classification of subcellular location patterns

A particularly important category of information for building systems models is the location of proteins and other biological molecules within cells. Because fluorescence microscopy is the most commonly used method for determining the subcellular location of proteins, an important initial question was whether automated analysis of subcellular patterns in fluorescence microscope images was feasible. This question was answered by the demonstration that five subcellular patterns could be distinguished in 2D images of Chinese hamster ovary cells (Boland et al. 1998) and that 10 subcellular patterns could be distinguished in HeLa cells (Murphy et al. 2000; Boland and Murphy 2001).

The dramatic variation in cell size, shape, and orientation exhibited by cultured cells combined with the extensive variation in position of organelles within cells suggested that approaches involving direct (pixel-by-pixel) comparisons with a library of cell images of known patterns would not provide accurate assignment of new images to one of those patterns. Instead, a feature-based approach was used in which each image is represented by a set of numerical features that capture various aspects of the pattern without being overly sensitive to rotation or translation within the sample plane.

These features have been systematically described and combined into sets of subcellular location features (SLFs). Initial work on distinguishing 10 patterns in HeLa cells achieved an average accuracy of 83% on individual cells using feature set SLF5 and a neural network classifier (Boland and Murphy 2001). Subsequent work has improved this accuracy to 92% using feature set SLF16 and a majority-voting ensemble classifier (Huang and Murphy 2004).

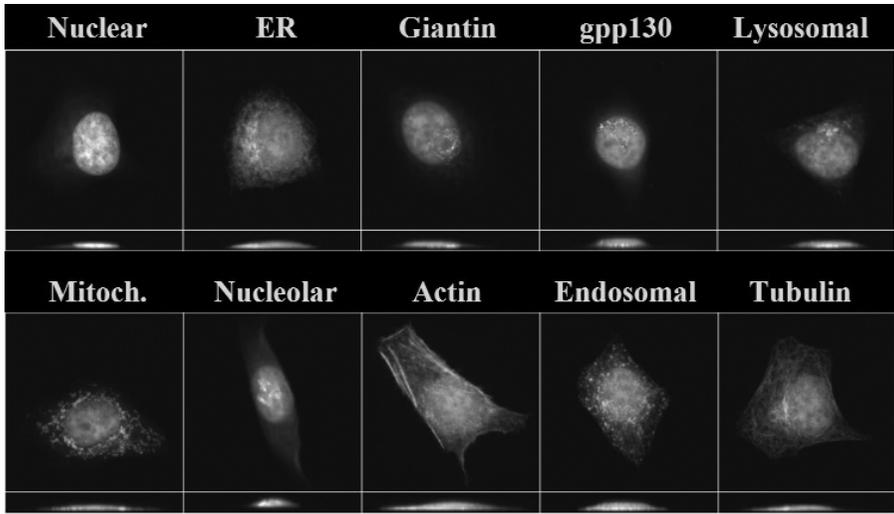


Figure 16.1. Representative 3D images of 10 subcellular patterns that can be distinguished with high accuracy by automated classifiers. The distribution of a DNA probe is shown in red, that of total cell protein in blue, and that of specific organelle markers in green. The paired images are maximum value projections along the z or x axis. (Picture copyright Carnegie Mellon University.) (see color plate 12).

An important conclusion from this work was that patterns that cannot be distinguished by visual examination could be discriminated by the automated systems (Murphy et al. 2003). In particular, two Golgi proteins that cannot be distinguished better than random guessing by visual examination can be recognized with accuracies of 82 to 90% using SLF16 (Huang and Murphy 2004). Discrimination of the similar lysosomal and endosomal patterns by the automated system is also 5 to 6% better than that achieved by visual examination.

Because macromolecules are distributed in three dimensions within cells, not just two, the accuracy of classification of 3D images obtained by confocal microscopy was also investigated. An initial accuracy of 91% using feature set SLF9 and a neural network classifier was obtained for the same 10 patterns previously studied in HeLa cells (Velliste and Murphy 2002), and this accuracy was subsequently improved to 98% using feature set SLF17 (Chen and Murphy 2004). Example images of the 10 patterns are shown in Figure 16.1.

2. Automated microscopy and pattern classification

These results demonstrate that the fundamental problem of recognizing the major subcellular patterns in 2D and 3D images has been solved. However, practical experience shows that the automation of the data acquisition process (including auto-focusing and detection of structurally consistent and homogeneously stained cells) still imposes limitations to achieve highest classification accuracy. As the technol-

ogy evolves, the next step can be taken in applying these methods to characterize entire proteomes, and we have coined the term *location proteomics* to describe this approach (Chen et al. 2003).

One way in which this can be done is to collect images of many different proteins and assign each protein to one of the major classes. An important recent test of an automated approach has been performed using expression in MCF7 cells of 11 GFP-tagged proteins via transfection (Conrad et al. 2004). The average accuracy reported was 82%, but this average included recognition of a separate “artifact” class (created by visual inspection of the training images). The accuracy obtained for assignments to the 11 protein patterns was 73%.

The corresponding higher error rate if compared for similar analysis in HeLa cells could be due to any of a number of differences between the studies, including cell type, the use of overexpressed fusion proteins versus endogenous proteins, the magnification, and the use of different feature sets. But most importantly, the additional challenges of accurate automated autofocusing and cell segmentation must be considered as well. Nonetheless, current results are encouraging for the use of automated microscopy, especially in that the accuracy of classification of a particular protein can be improved by combining results from more than one cell (Boland and Murphy 2001).

3. *Clustering of proteins by location pattern*

An alternative to assigning proteins to “known” subcellular location patterns is to use unsupervised learning methods to identify the statistically significant patterns observed and group proteins by them. The principle is to represent each protein pattern using the SLFs but to use cluster analysis to group them rather than to classify them. This approach was demonstrated using 3D images of a number of proteins in 3T3 cells (Chen et al. 2003). This study used cloned cell lines expressing randomly-chosen proteins fused internally with GFP using CD-tagging (Jarvik et al. 2002).

A recent study of 90 of these clones tested different ways of measuring distance between proteins in the feature space, as well as different clustering approaches (Chen and Murphy 2005). The results indicated that the clones formed 17 distinguishable clusters that provide greater refinement than visual description of the patterns using standard terms. The consensus tree obtained, along with example images from various clusters, is shown in Figure 16.2.

4. *Imaging protein kinetics*

The results summarized previously were all obtained using static images that represent the steady-state distribution of proteins. A major upcoming challenge will be the acquisition and incorporation into location proteomics of information on the kinetics with which proteins move in these steady states, as well as the kinetics with which those states change due to the cell cycle, environmental changes, onset of disease, or addition of drugs. Time-lapse imaging techniques allow consistent pro-

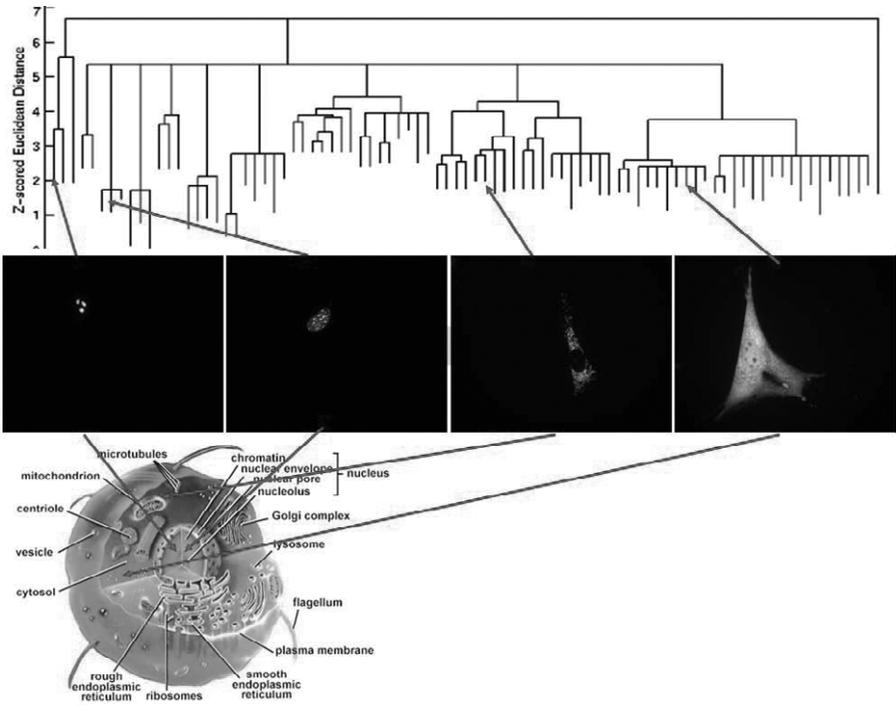


Figure 16.2. Consensus subcellular location tree for 3T3 cell lines obtained by CD-tagging. 3T3 cells were infected with a retroviral construct carrying a GFP coding sequence surrounded by splicing donor and acceptor sites (CD-tagging). Clones expressing randomly-tagged proteins were isolated and the tagged gene identified by RT-PCR (Jarvik et al. 2002). At least 10 3D images of live cells were obtained for each clone using spinning disk microscopy (Chen and Murphy 2003), and the clones were grouped as described in the text and as by Chen and Murphy (2005). The degree of dissimilarity between any pair of clones can be found by measuring the vertical distance from one of them to the highest node along the minimal path to the other clone, plus the vertical distance from that node to the other clone (as discussed in the text, this distance reflects the separation between the two clones in the SLF feature space). The names of the proteins are not shown due to space limitations. Examples of images from various branches of the tree are shown. The full tree with names and images of all clones is available at <http://murphylab.web.cmu.edu/services/PSLID/tree.html> (see color plate 13).

filing of changes in the cell state. FRAP and FRET (and the tracking protein complexes with q-dots) are powerful methods for studying dynamics. Relocation of protein species reveals important functional activities on state-space transients. An example is the release of cytochrome c from the mitochondria into the cytosol that mediates apoptosis (Goldstein et al. 2000), or the dynamics of histone binding to chromatin in living cells (Mistell et al. 2000).

5. Location proteomics and generative models

It is anticipated that the methods described previously will be used for large-scale studies to characterize the subcellular distribution of proteins in a number of cell

types. Important areas of additional research involve enabling patterns that are a mixture of other patterns to be “unmixed” and making it possible to describe the distribution of any protein using generative models that can be incorporated into simulations to create distributions of many proteins within cells *in silico*.

D. Cytomics analysis in tissues

The introduction of robotic microscopy in concert with robust machine vision software that can discern the histomorphology of multicellular arrangements in tissues in a comprehensive fashion has extended cytomics into the tissue domain. A multicellular high-throughput, high-content analysis of tissues (sometimes termed tissomics [Ecker and Tarnok 2005; Kriete and Boyce 2005]) can be used to support and confirm histopathological assessment of tissues, allowing a more complete quantitative evaluation of phenotypical cellular responses and the identification of structural markers of tissue normality, injury, and disease. Specific applications include basic biomedical research, pathobioinformatics, investigative toxicology, drug target development, and tissue engineering.

1. High-throughput imaging

Complete imaging of a histological glass slide (20 × 50 mm) at 20× microscopic magnification can generate up to 3,000 individual digital color images, which new types of ultrafast scanners can image within minutes (Weinstein et al. 2001). The resulting image montages of several GB in size represent entire histological slides or tissue arrays. The enormous amount of data generated by this new class of microscopic scanners is stored in databases. Secondary representations, by subsampling or data compression, mainly serve viewing or control purposes and have been developed as part of adequate solutions for the handling and mining of such large data sets.

2. High-content image analysis

The analytical task of cytomics in tissues is a fully automated analysis of tissue profiles without user intervention, which can be challenging given variations caused by the prevailing methods of tissue preparation and staining. Robust analysis procedures that rely on the topology of cells and tissue structures, and new object-oriented approaches, are preferred solutions that have distinct advantage over the prevailing pixel-oriented methods (Price et al. 2002; Kriete et al. 2003). Understanding tissues as a hierarchy of larger anatomical constructs, consisting of different cell types in different phases of development, that further contain cell organelles and cell nuclei is key for object segmentation (Kriete and Boyce 2005).

These entities, once identified, provide a rich source for a hypothesis-free geometric intensity and field-specific characterization. The identification of significant components that change with disease state or treatment may be found by multivariate statistics in the course of further data analysis. The method is extendable to include specific stains and biomarkers, as well as tissue microarrays or tissue cultures.

III. DATA ANALYSIS

A. Data mining, differential analysis, and discrimination

Because cytomics investigates individual cells, it is possible to separate (or gate) cells into different state or response categories (Boolean classification). Cells are grouped and catalog based on a number of cellular features from one or multiple probes, typically originating from one level of observed granularity or resolution (horizontal analysis according to biological hierarchy). This includes overall fluorescence intensity, rate of rise/fall (for kinetics), area, object pixel statistics (average intensity, min and max), and variation of pixel intensity within ROIs (granulation algorithm).

Classes of cells can be color-coded for easy visualization, and average measurements over a subset of cells can be taken. A “well” classification is then applied based on the number of cells in each well that meet a user-defined threshold. In turn, a response “heat” plate-map that readily highlights cellular trends or compound hits can be generated. This process enables us to identify features that best reflect specific biological responses and that are therefore good screenable parameters.

Multiparametric single-cell analysis by flow or image cytometry can provide significant amounts of data that may seem difficult to distribute and analyze (Hood et al. 2004). Solutions to handle biological image data over the Web have been suggested but have had limited application (Lindek et al. 1999), whereas analytical procedures may relate cytomics data with biomedical literature and bioinformatics databases (Abraham et al. 2004).

Differential data pattern analysis (Valet and Hoeffkes 2004) provides a means of analyzing multiparametric data of various types in parallel in a nonhierarchical way. Such data from flow and image analysis, chip arrays, clinical chemistry, and clinical data can be simultaneously processed in a manner similar to that of predictive medicine by cytomics (Valet 2002). Disease-induced differentials in patients versus normal individuals, stationary disease patients, or survivors are analyzed in this approach instead of differentials from perturbed model systems that may not exactly reflect the human situation (Horrobin 2003).

The algorithmic procedure is summarized as follows. Numeric data columns are transformed into triple matrix characters (–) = decreased for values below a lower percentile threshold, into (0) = unchanged when between lower and upper threshold, and into (+) = above an upper percentile threshold. The resulting triple matrix database is classified in a learning situation for samples of patients from different classification categories (such as healthy versus diseased, progressive versus stationary disease, and survivor versus non-survivor patients). Individual triple matrix columns are temporarily removed from the learning process in a sequential way to assess their individual contribution to the classification result.

At the end of the learning process only data columns having improved the initial classification remain in the discriminatory bio-parameter patterns comprising typically between 10 and 30 parameters. The bio-parameter patterns can be further

used for the exploration of molecular disease pathways and in the search for new drug targets. In this way, single-cell- and single-individuum-oriented analyses provide a maximum of discrimination because no averaging over heterogeneous entities occurs during data acquisition and bioinformatic evaluation (Szaniszlo et al. 2004).

Principal component analysis is another way of reducing the complexity of the data, in particular if cytomics data are merged with other bioinformatic data sets from the same cells and tissues, such as gene expression profiles (Kriete et al. 2003). An alternative is Fisher discriminant analysis (FDA), which was used previously to demonstrate improved differentiation of treatment groups if chemical data are combined with multicellular phenotypical data (Kriete et al. 2005).

Nature-induced bio-parameter perturbations or differentials (such as between diseased versus healthy, progressive versus stationary disease, or survivor versus non-survivor patients) can be directly analyzed instead of generating hypothesis-driven systematic perturbations in model systems. Individualized disease course prediction for patients is possible in this way (Valet 2002), without the prerequisite of fully understanding the entire molecular network of disease-associated cell system changes. Discriminatory data patterns are obtained by multiparameter data analysis.

These data patterns can be further investigated by a molecular reverse-engineering strategy (Valet 2005) to understand disease-inducing molecular pathways or to find new drug targets. It is advantageous for this concept that many data sets are already available as starting material from current or past clinical studies in which patients are routinely followed for diagnostic or therapeutic purposes.

B. System-wide data correlations

Multicellular profiles can be correlated statistically with gene expression profiles. Foundations for this (vertical) analysis crossing different levels of biological hierarchy are multi-sample comparisons, assuming that changes on one level of biological organization consistently alter the phenotype and function on a higher physiological level.

As an example, Spearman's rank order correlations have revealed significant monotonic relationships that illuminate important connections between structural features in tissue composition and gene expression levels (Kriete et al. 2003). Similarly, a hierarchical clustering analysis based on a jackknife correlation demonstrated correlations between groups of genes with tissue cytometric markers (Kriete and Boyce 2005). As such, cytomics-related techniques provide covariants that can be used to enrich gene expression analysis (Boyce et al. 2005).

IV. DISCUSSION

Cells represent elementary building units of cell systems, organs, and organisms, and diseases are caused by molecular changes in cells and cell systems. Consid-

ering the heterogeneity of human cell systems, single-cell analysis (Szaniszlo et al. 2004) is important in resolving a maximum of compartmentalized molecular heterogeneity; for example, to discriminate changes in diseased or disease-associated cells from nonaffected bystander cells. Technical progress broadens the number and quality of available cell state variables, such as cytometry using microfluidic chips (Palkova et al. 2004; Wu et al. 2004) and capillary electrophoresis (Dovichi and Hu 2003; Arkhipov et al. 2005). Cell microgenomics expression profiles (Taylor et al. 2004)—as well as single-cell proteomics (Dovichi and Hu 2003) and metabolomics (Palkova et al. 2004; Wu et al. 2004; Arkhipov et al. 2005)—also become accessible.

An important concept of systems biology subsists in the application of multiple differential perturbations on biological cell systems to observe their molecular reactivity with the aim of mathematical modeling to understand the mechanisms of the observed alterations. The prediction of the reactivity for biological systems under predefined conditions represents a further goal. Cell arrays and microwell infection assays on cultured cells in conjunction with RNAi allows screening of the morphological phenotypical states in a high-throughput fashion. At present, a suggested comprehensive mapping of all proteins in the cell or in cell compartments by using high-resolution electron tomography (Baumeister 2004) is still limited, and the required resolution has to be improved. Light microscopic imaging techniques in conjunction with fluorescence markers, as described here, are therefore the preferred technique and can be more easily applied in a medical environment.

Single-cell techniques overcome the problem of averaged cellular information in cell homogenates or extracts where it cannot be decided whether observed changes derive from all cells or only from a particular cell subpopulation. The analysis of humoral body compartments such as blood plasma or serum, urine, or cerebrospinal fluid as a further alternative provides only secondary information by cell-derived molecules. Metabolites from cellular disease processes may have been altered in the meantime, or they may not become apparent in humoral compartments for lack of secretion or owing to fast renal or biliary excretion.

It may be contended that the single-cell approach will frequently not be feasible because not all cells of a given sample can be analyzed (as, for example, in smears, biopsies, or histological sections). Experience shows that it is not obligatory to analyze all cells of a given sample before one can derive relevant conclusions. It is frequently sufficient to analyze a representative fraction of diseased cells as well as reference cells. This will be shown by the subsequent examples. Mechanical disaggregation of tissues at 0 to 4°C for cell function analysis by flow cytometry destroys between 90 and 95% or more of all cells.

Furthermore, a relative enrichment of epithelial and inflammatory cells occurs because fibroblasts or smooth muscle cells have been largely destroyed. More than 90% of cancer patients are correctly identified from flow-cytometrically identified molecular cell properties (Valet et al. 1984; Liewald et al. 1990). This indicates that a representative fraction of cancer cells and normal epithelial reference cells has

survived despite the fact that the cellular composition of the samples has changed and that the tissue architecture was lost during cell preparation.

The result is not surprising because diseases represent molecular changes in cells and cell systems. The analysis of diseased cells or disease-associated inflammatory and immune cells should therefore by itself contain the relevant molecular information about the actual state (diagnosis) and the future development (prediction) of a disease, irrespective of the original position of the analyzed cells in an organ. A further reservation concerning single-cell analysis is that cell properties may be altered during preparation for analysis (Hood et al. 2004). Deep-freezing of tissues, immediate cell fixation, or cell preparation between 0 and 4°C for functional studies, however, minimizes such risks.

Valuable information is obtained, for example, from the functional analysis of oxidative status or oxidative burst in inflammatory immune cells such as lympho-, mono-, and granulocytes. Such disease-associated cells can be measured in tissues but advantageously also in the peripheral blood, where high-speed multiparameter flow-cytometric single-cell analysis is possible and provides individualized predictions or risk assessments for intensive-care patients (Valet et al. 1998, 2001). We can conclude that molecular alterations by cell preparation or staining steps cannot be generally excluded. They do, however, definitely not impair the determination of clinically relevant molecular cell parameters.

V. CONCLUSIONS

The value of the single-cell/single-individual analysis concept resides in its clinical value for the individual patient as well as in the bio-parameter patterns being of interest for molecular reverse engineering by systems biology. The backward molecular analysis may provide information on specific molecular pathways responsible for disease formation and reveal new drug targets. A specific focus of cytomics is in location proteomics, which uses quantitative readouts for functional genomics.

Detailed knowledge of the location, concentration, and activation of proteins and other biological molecules and valuable information can be obtained by studying cell behaviors in a systematic fashion in parallel with ongoing proteomics projects. Information on specific biomarkers and proteome changes associated with function, disease, and age can be valuable for diagnostics or therapeutics before a complete mapping of the proteome is available.

Diseases are typically diagnosed by clinicians from clinical symptoms or clinical chemistry parameters, or by pathologists from the evaluation of the altered microscopic morphology in tissue sections or in cytological samples. Single-cell image or flow-cytometric analysis extends the diagnostic knowledge level by the description of molecular cell phenotypes and may detect alterations at a stage where no morphological correlate is yet detectable.

Such measurement may also address therapy-related future disease courses of individual patients as a clinically promising new feature (predictive medicine by

cytomics) (Valet et al. 2001; Valet 2002; Valet and Tarnok 2003). A human cytochrome project has recently been proposed (Valet and Tarnok 2004; Valet et al. 2004) to particularly focus on the development and management of clinically complex diseases such as malignancies, infections, diabetes, allergies, rheumatoid diseases, asthma, myocardial infarction, stroke, and others.

The translational research concept laid out is deductive for the selected analytical parameters, but inductive during the data evaluation phase because the information of all quantifiable variables and cells is investigated for its discriminatory potential. In this step, most of the non-differential information in state space is typically eliminated as irrelevant during the algorithmic data-sieving phase.

The remaining discriminatory information may uncover new molecular knowledge otherwise unreachable by traditional hypothesis formulation. It also provides initial focus points for modeling efforts. It is unknown how much molecular knowledge is required to model, for example, disease susceptibility or future disease courses in individual patients. However, cytomics now opens the possibility of constraining bottom-up forward engineering (Collins et al. 2003)—as in network or spatiotemporal modeling—with precise data from a cellular level of biological organization.

REFERENCES

- Abraham, V. C., Taylor, D. L., and Haskins, J. R. (2004). High content screening applied to large-scale cell biology. *Trends in Biotechnology* **22**:15–22.
- Arkipov, S. N., Berezovski, M., Jitkova, J., and Krylov, S. N. (2005). Chemical cytometry for monitoring metabolism of a Ras-mimicking substrate in single cells. *Cytometry* **63A**: 41–47.
- Bajaj, S., Welsh, J. B., Leif, R. C., and Price, J. H. (2000). Ultra-rare-event detection performance of a custom scanning cytometer on a model preparation of fetal nRBCs. *Cytometry* **39**:285–294.
- Baumeister, W. (2004). Mapping molecular landscapes inside cells. *Biol. Chem.* **385**(10):865–872.
- Boyce, K., Kriete, A., Nagatomi, S., Kelder, B., Cosohigano, K., and Kopchick, J. J. (2005). Phenotypical enrichment strategies for microarray data analysis applied in Type II diabetes study. *OMICS* **9**:252–266.
- Boland, M. V., Markey, M. K., and Murphy, R. F. (1998). Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* **33**:366–375.
- Boland, M. V., and Murphy, R. F. (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **17**:1213–1223.
- Boppart, S. A., Bouma, B. E., Pitris, C., Southern, J. F., Brezinski, M. E., and Fujimoto, J. G. (1998). *In vivo* cellular optical coherence tomography imaging. *Nat. Med.* **4**:861–865.
- Burgemeister, R., Gangnus, R., Haar, B., Schutze, K., and Sauer, U. (2003). High-quality RNA retrieved from samples obtained by using LMPC (laser microdissection and pressure catapulting) technology. *Pathol. Res. Pract.* **1**(99):431–436.
- Campagnola, P. J., and Loew, L. M. (2003). Second-harmonic imaging microscopy for visualizing biomolecular arrays in cells, tissues and organisms. *Nature Biotech.* **21**:1356–1360.

- Chen, X., and Murphy, R. F. (2005). Objective clustering of proteins based on subcellular location patterns. *Journal of Biomedicine and Biotechnology* (in press).
- Chen, X., and Murphy, R. F. (2004). Robust classification of subcellular location patterns in high-resolution 3D fluorescence microscope images. In *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* vol. 3: 1632–1635.
- Chen, X., Velliste, M., Weinstein, S., Jarvik, J. W., and Murphy, R. F. (2003). Location proteomics: Building subcellular location trees from high-resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc. SPIE* **4962**:298–306.
- Chitty, M. (2001). -Omes and -omics Glossary. www.genomicglossaries.com/content/omes.asp.
- Conrad, C., Erfle, H., Warnat, P., Daigle, N., Lorch, T., Ellenberg, J., Pepperkok, R., and Eils, R. (2004). Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research* **14**:1130–1136.
- Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003). A vision for the future of genomics research. *Nature* **422**:835–847.
- Crow, P., Uff, J. S., Farmer, J. A., Wright, M. P., and Stone, N. (2004). The use of Raman spectroscopy to identify and characterize transitional cell carcinoma *in vitro*. *BJU Int.* **93**:1232–1236.
- Dovichi, J., and Hu, S. (2003). Chemical cytometry. *Curr. Opin. Chem. Biol.* **7**:603–608.
- Ecker, R. C., and Tarnok, A. (2005). Cytomics goes 3D: Towards tissomics. *Cytometry* (in press).
- Ecker, R. C., De Martin, R., Steiner, G. E., and Schmid, J. A. Application of spectral imaging microscopy in cytomics and fluorescence resonance energy transfer (FRET) analysis. *Cytometry* **59A**:172–181.
- Ecker, R. C., and Steiner, G. E. (2004). Microscopy-based multicolor tissue cytometry at the single-cell level. *Cytometry* **59A**:182–190.
- Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**:14863–14868.
- Englert, C. R., Baibakov, G. V., and Emmert-Buck, M. R. (2000). Layered expression scanning: Rapid molecular profiling of tumor samples. *Cancer Research* **60**:1526–1530.
- Esa, A., Edelmann, P., Kreth, G., et al. (2000). Three-dimensional spectral precision distance microscopy of chromatin nanostructures after triple-colour DNA labeling: A study of the BCR region on chromosome 22 and the Philadelphia chromosome. *J. Microsc.* **199**:96–105.
- Freedman, L. J., and Maddox, M. T. (2001). A comparison of anti-biotin and biotinylated anti-avidin double-bridge and biotinylated tyramide immunohistochemical amplification. *J. Neuroscience Meth.* **112**:43–49.
- Gerstner, A. O., Trumppheller, C., Racz, P., Osmancik, P., Tenner-Racz, K., and Tarnok, A. (2004). Quantitative histology by multicolor slide-based cytometry. *Cytometry* **59A**: 210–219.
- Gerstner, A. O. H., Laffers, W., Lenz, D., Bootz, F., Steinbrecher, M., and Tárnok, A. (2002). Near-infrared dyes for immunophenotyping by LSC. *Cytometry* **48**:115–123.
- Goldstein, J. C., Waterhouse, N. J., Juin, P., Evan, G. I., and Green, D. R. (2000). The coordinate release of cytochrome c during apoptosis is rapid, complete and kinetically invariant. *Nat. Cell Biol.* **2**:156–162.
- Han, F., Wang, Y., Sims, C. E., et al. Fast electrical lysis of cells for capillary electrophoresis. *Anal. Chem.* **75**:3688–3696.
- Hell, S. W. (2003). Towards fluorescence nanoscopy. *Nature Biotech.* **21**:1347–1355.
- Herbert, A. (2004). The four Rs of RNA-directed evolution. *Nat. Genet.* **36**:19–25.

- Hesse, J., Sonnleitner, M., and Schutz, G. J. (2004). Ultra-sensitive fluorescence reader for bioanalysis. *Curr. Pharm. Biotechnol.* **5**:309–319.
- Holmes, T. J., and Liu, Y. H. (1992). Image restoration for 2-D and 3-D fluorescence microscopy. In A. Kriete (ed.), *Visualization in Biomedical Microscopies: 3-D Imaging and Computer Applications*, pp. 283–323. Verlag Chemie: Weinheim.
- Hood, L., Heath, J. R., Phelps, M. E., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**:640–643.
- Hood, L. (2003). Systems biology: Integrating technology, biology and computation. *Mechanisms of Aging and Development* **124**:9–16.
- Horrobin, D. F. (2003). Modern biomedical research: An internally self-consistent universe with little contact with medical reality? *Nature Rev. Drug Discovery* **2**:161–164.
- Huang, K., and Murphy, R. F. (2004). Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics* **5**:78.
- Huang, K., and Murphy, R. F. (2004). From quantitative microscopy to automated image understanding. *J. Biomed. Optics* **9**:893–912.
- Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Ann. Rev. Genomics Hum. Genet.* **2**:343–372.
- Jager, S., Garbow, N., Kirsch, A., Preckel, H., Gandenberger, F. U., Herrenknecht, K., Rudiger, M., Hutchinson, J. P., Bingham, R. P., Ramon, F., Bardera, A., and Martin, J. (2003). A modular, fully integrated ultra-high-throughput screening system based on confocal fluorescence analysis techniques. *J. Biomol. Screen.* **8**:648–659.
- Jares-Erijman, E. A., and Jovin, T. A. (2003). FRET imaging. *Nature Biotechnol.* **21**:1387–1395.
- Jarvik, J. W., Fisher, G. W., Shi, C., Hennen, L., Hauser, C., Adler, S., and Berget, P. B. (2002). *In vivo* functional proteomics: Mammalian genome annotation using CD-tagging. *BioTechniques* **33**:852–867.
- Kitano, H. (2002). Systems biology: A brief overview. *Science* **295**:1662–1664.
- Kitano, H. (2001). *Foundations of Systems Biology*. Cambridge, MA: MIT Press.
- Kriete, A., and Boyce, K. (2005). Automated tissue analysis: A bioinformatics perspective. *Methods of Information in Medicine* **44**:32–37.
- Kriete, A., Freund, J., Anderson, M., et al. (2003). Combined histomorphometric and gene expression profiling applied to toxicology. *Genome Biology* **4**:R32.
- Kriete, A., Anderson, M., Love, B., Caffrey, J., Young, B., Sendera, T., Magnuson, S., and Braughler, M. (2003). Combined histomorphometric and gene expression profiling applied to toxicology. *Genome Biology* **4**:R32.
- Lenz, D., Gerstner, A., Laffers, W., Steinbrecher, M., Bootz, F., and Tárnok, A. (2003). Six and more color immunophenotyping on the slide by laser scanning cytometry (LSC). In D. V. Nicolau, J. Enderlein, R. C. Leif, and D. L. Farkas (eds.), *Proceedings of SPIE* **4962**:364–374.
- Liewald, F., Demmel, N., Wirsching, R., Kahle, H., and Valet, G. (1990). Intracellular pH, esterase activity and DNA measurements of human lung carcinomas by flow-cytometry. *Cytometry* **11**:341–348.
- Lindek, S., Fritsch, R., Machtynger, J., de Alarcon, P. A., and Chagoyen, M. (1999). Design and realization of an on-line database for multidimensional microscopic images of biological specimens. *J. Struct. Biol.* **125**:103–111.
- Lund-Johansen, F., Davis, K., Bishop, J., and de Waal-Malefyt, R. (2000). Flow cytometric analysis of immunoprecipitates: High-throughput analysis of protein phosphorylation and protein-protein interaction. *Cytometry* **39**:250–259.

- Manconi, F., Kable, E., Cox, G., Markham, R., and Fraser, L. S. (2003). Whole-mount sections displaying microvascular and glandular structures in human uterus using multiphoton excitation microscopy. *Micron* **34**:351–358.
- Mandarim-de-Lacerda, C. A. (2003). Stereological tools in biomedical research. *Ann. Acad. Bras. Cienc.* **75**:469–486.
- McClain, M. A., Culbertson, C. T., Jacobson, S. C., Allbritton, N. L., Sims, C. E., and Ramsey, J. M. (2003). Microfluidic devices for the high-throughput chemical analysis of cells. *Anal. Chem.* **75**:5646–5655.
- McCloskey, K. E., Chalmers, J. J., and Zborowski, M. (2003). Magnetic cell separation: Characterization of magnetophoretic mobility. *Anal. Chem.* **75**:6868–6874.
- Megason, S. G., and Fraser, S. E. (2003). Digitizing life at the level of a cell: High-performance laser-scanning microscopy and image analysis for in toto imaging of development. *Mechanisms Development* **120**:1407–1420.
- Mempel, T. R., Henrickson, S. E., and Von Andrian, U. H. (2004). T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature* **427**:154–159.
- Mistell, T., Gunjan, A., Hock, R., Bustin, M., and Brown, D. T. (2000). Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408**:877–881.
- Mittag, A., Lenz, D., Gerstner, A. O. H., Sack, U., Steinbrecher, M., Koksche, M., Raffae, A., Bocs, J., and Tarnok, A. (2005). Polychromatic (eight color) slide-based cytometry for the phenotyping of leukocyte, NK and NKT subsets. *Cytometry* (in press).
- Molnar, B., Berci, L., Diczhazy, C., et al. (2003). Digital slide and virtual microscopy based routine and telepathology evaluation of routine gastrointestinal biopsy specimens. *J. Clin. Pathol.* **56**:433–438.
- Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**:743–747.
- Murata, S., Herman, P., Lin, H. J., and Lakowicz, J. R. (2000). Fluorescence lifetime imaging of nuclear DNA: Effect of fluorescence resonance energy transfer. *Cytometry* **41**:178–185.
- Murphy, R. F., Velliste, M., and Porreca, G. (2003). Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *J. VLSI Sig. Proc.* **35**:311–321.
- Murphy, R. F., Boland, M. V., and Velliste, M. (2000). Towards a systematics for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**:251–259.
- Palkova, Z., Vachova, L., Valer, M., and Preckel, T. (2004). Single-cell analysis of yeast, mammalian cells, and fungal spores with a microfluidic pressure-driven chip-based system. *Cytometry* **59A**:246–253.
- Parak, W. J., Gerion, D., Pellegrino, T., et al. (2003). Biological applications of colloidal nanocrystals. *Nanotechnology* **14**:15–27.
- Pawley, J. (ed.). (1995). *Handbook of Biological Confocal Microscopy*. (2d ed.). New York: Plenum Press.
- Peter, M., and Ameer-Beg, S. M. (2004). Imaging molecular interactions by multiphoton FLIM. *Biol. Cell* **96**:231–236.
- Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science* **306**:1194–1198.

- Price, J. H., Goodacre, A., Hahn K., et al. (2002). Advances in molecular labeling, high-throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J. Cell Biochem. Suppl.* **39**:194–210.
- Rieti, S., Manni, V., Lisi, A., Giuliani, L., Sacco, D., D'Emilia, E., Cricenti, A., Generosi, R., Luce, M., Grimaldi, S. (2004). SNOM and AFM microscopy techniques to study the effect of non-ionizing radiation on the morphological and biochemical properties of human keratinocytes cell line (HaCaT). *J. Microsc.* **213**:20–28.
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. **308**(5721):523–529.
- Schilb, A., Riou, V., Schoepfer, J., Ottl, J., Muller, K., Chene, P., Mayr, L. M., and Filipuzzi, I. (2004). Development and implementation of a highly miniaturized confocal 2D-FIDA-based high-throughput screening assay to search for active site modulators of the human heat shock protein 90beta. *J. Biomol. Screen.* **9**:569–577.
- Smolle, J., Gerger, A., Weger, W., Kutzner, H., and Tronnier, M. (2002). Tissue counter analysis of histologic sections of melanoma: Influence of mask size and shape, feature selection, statistical methods and tissue preparation. *Anal. Cell Pathol.* **24**:59–67.
- Sultan, M., Wigle, D. A., Cumbaa, C. A., Maziarz, M., Glasgow, J., Tsao, M. S., and Jurisica, I. (2002). Binary tree-structured vector quantization approach to clustering and visualizing microarray data. *Bioinformatics* **18**(1):S111–S119.
- Szanişzlo, P., Wang, N., Sinha, M., Reece, L. M., van Hook, J., Luxon, B. A., and Leary, J. F. (2004). Getting the right cells to the array: Gene expression microarray analysis of cell mixtures and sorted cells. *Cytometry* **59A**:191–202.
- Tarnok, A., Hamsch, J., Chen, R., and Varro, R. (2003). Cytometric bead array to measure six cytokines in twenty-five microliters of serum. *Clin. Chem.* **49**:1000–1002.
- Tarnok, A., and Gerstner, A. (2002). Clinical applications of laser scanning cytometry. *Cytometry* **50**:133–143.
- Taylor, T. B., Nambiar, P. R., Raja, R., Cheung, E., Rosenberg, D. W., and Andereg, B. (2004). Microgenomics: Identification of new expression profiles via small and single-cell sample analysis. *Cytometry* **59A**:254–261.
- Tsai, P. S., Friedman, B., Ifarraguerri, A. I., et al. (2003). All-optical histology using ultrashort laser pulses. *Neuron*. **39**:27–41.
- Ulrich, H., Martins, A. H., and Pesquero, J. B. (2004). RNA and DNA aptamers in cytomics analysis. *Cytometry* **59A**:220–231.
- Valet, G. K. (2002). Predictive medicine by cytomics: potential and challenges. *JBRHA* **16**:164–167.
- Valet, G. K. (2005). Human cytome project, cytomics and systems biology: The incentive for new horizons in cytometry. *Cytometry* **64A**:1–2.
- Valet, G. K., and Hoeffkes, H. G. (2004). Data pattern analysis for the individualised pretherapeutic identification of high-risk diffuse large B-cell lymphoma (DLBCL) patients by cytomics. *Cytometry* **59A**:232–236.
- Valet, G. K., and Tarnok, A. (2003). Cytomics in predictive medicine. *Cytometry* **53B**:1–3.
- Valet, G. K., and Tarnok, A. (2004). Potential and challenges of a human cytome project. *JBRHA* **18**:87–91.
- Valet, G. K., Rüssmann, L., and Wirsching, R. (1984). Automated flow-cytometric identification of colo-rectal tumor cells by simultaneous DNA, CEA-antibody and cell volume measurements. *J. Clin. Chem. Clin. Biochem.* **49**:83–90.

- Valet G. K., Roth, G., and Kellermann, W. (1998). Risk assessment for intensive care patients by automated classification of flow cytometry data. In J. P. Robinson and G. F. Babcock (eds.). *Phagocyte Function*, pp 289–306. New York: Wiley-Liss.
- Valet, G. K., Kahle, H., Otto, F., Bräutigam, E., and Kestens, L. (2001). Prediction and precise diagnosis of diseases by data pattern analysis in multiparameter flow cytometry. Melanoma, juvenile asthma and human immunodeficiency virus infection. *Meth. Cell. Biol.* **64**:487–508.
- Valet, G. K., Leary, J. F., and Tarnok, A. (2004). Cytomics-: New technologies towards a human cytome project. *Cytometry* **59A**:167–171.
- Velliste, M., and Murphy, R. F. (2002). Automated determination of protein subcellular locations from 3D fluorescence microscope images. Proceedings 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002), Bethesda, Maryland, USA. 867–870.
- Weinstein, R. S., Descour, M. R., Liang, C., Bhattacharyya, A. K., Graham, A. R., Davis, J. R., Scott, K. M., Richter, L., Krupinski, E. A., Szymus, J., Kayser, K., and Dunn, B. E. (2001). Telepathology overview: From concept to implementation. *Hum. Pathol.* **32**:1283–1299.
- Wu, H., Wheeler, A., and Zare, R. N. (2004). Chemical cytometry on a picoliter-scale integrated microfluidic chip. *PNAS* **101**:12809–12813.
- Xu, C., Ye, J., Marks, D. L., and Boppart, S. A. (2004). Near-infrared dyes as contrast-enhancing agents for spectroscopic optical coherence tomography. *Opt. Lett.* **29**:1647–1649.

The IUPS Physiome Project: Progress and Plans

**Peter Hunter, Kelly Burrowes, Justin
Fernandez, Poul Nielsen, Nic Smith, and
Merryn Tawhai**

*Bioengineering Institute, University of Auckland, New
Zealand*

Chapter 17

ABSTRACT

The IUPS Physiome Project aims to facilitate the understanding of physiological function in healthy and diseased mammalian tissues by developing a multi-scale modeling framework that can link biological structure and function across all spatial scales. To achieve this requires an open-source internationally collaborative effort to design XML standards for encapsulating models, web-accessible model databases, and computational tools for authoring and visualizing models and running model simulations. A brief overview of the project is given in this chapter, with a discussion of the current progress and future plans for three particular organ systems: the heart, lungs, and musculo-skeletal system.

I. INTRODUCTION

The focus of biomedical science over the past few decades has, for good reason, been on molecular biology and the revolution associated with sequencing the human genome, understanding gene transcription and translation, determining the 3D structure of proteins, and using a range of remarkable imaging technologies to investigate cellular processes. With the vast amount of experimental data now available, it is hardly surprising that a “systems biology” (quantitative, model-based framework) has emerged to deal with the overwhelming complexity of molecular and cellular biology. A number of chapters in this book describe many aspects of this systems biology framework.

Another revolution, equally important, has occurred over the same period. This is the development of imaging technologies such as tagged MRI, helical scan CT, high-resolution electrical mapping, and so on, which now allow the larger-scale physiological processes in the body to be measured and probed with considerable precision in a whole-body clinical setting. For example, it is now possible to routinely obtain detailed measurements of mechanical deformation in the heart throughout the cardiac cycle, to measure gas movement in the airways of the breathing lung, and to obtain high-resolution digital images of the structure of many internal organs (Hunter et al. 2002).

The mathematical framework for interpreting and modeling the function of organs and organ systems is, moreover, well developed because it has been able to build on 150 years of developments in the physical and engineering sciences. Most of the manufactured devices we rely on in the modern world (cars, airplanes, cell phones, and so on) are based on a thorough understanding of the underlying physical laws of nature, such as conservation of mass, momentum, energy, and the corresponding equations of electromagnetism, mechanics, and so on. Sophisticated software tools, based on these physical principles, are available for engineers and physicists for designing new products and analyzing their behavior.

Applying these tools to biological organ systems is fairly straightforward, although it does pose some particular challenges. Nature almost never exploits geometric symmetry in the way engineers do, and thus the full 3D shape and structure of organs has to be modeled. Nature also uses anisotropic and inhomogeneous materials and seldom limits herself to a linear range of behavior, whereas engineering structures typically employ homogeneous isotropic components with linear material properties (Hunter and Borg 2003; Crampin et al. 2004).

If we are to link these two ends of the biological spectrum—from genes and proteins to whole-organ physiology via protein pathways to cell and tissue structure and function—we need to address the challenge of multi-scale modeling across spatial scales from nanometers to meters and temporal scales of microseconds to a human lifetime (see Figure 17.1). The benefits of achieving such a synthesis, especially one that incorporates patient-specific data, would be considerable (Noble 2002).

In this chapter we give a brief overview of a project attempting to achieve this synthesis, called the IUPS Physiome Project. The concept of a “Physiome Project” was presented in a report from the Commission on Bioengineering in Physiology to the International Union of Physiological Sciences (IUPS) Council at the 32nd World Congress in Glasgow in 1993. The term *physiome* comes from *physio* (life) + *ome* (as a whole), and is intended to provide a “quantitative description of physiological dynamics and functional behavior of the intact organism” (Bassingthwaight 1995, 2000).

We begin by illustrating some of the organ models that have been developed and the progress likely in the next few years. We then discuss the open-source software tools being developed to facilitate international collaboration on the project.

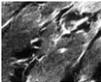
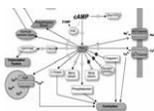
Physiological level		Types of model	3D Imaging devices
Organ systems cardiovascular, respiratory, musculo-skeletal, digestive, skin, urinary, nervous, endocrine, lymphatic, male reproductive, female reproductive, special sense organs		Systems theory	MRI, CT, PET 3D Ultrasound 1mm or 10^{-3} m
Organ 50 organs		Continuum theory Conservation of mass Conservation of momentum Conservation of charge	
Tissue Epithelial Connective Muscle Nerve		Conservation equations Passive flux equations Carrier mediated transport Electroneutrality constraints	MicroCT, Optical Coherence Tomography 10μ or 10^{-6} m Serial sections 1μ or 10^{-6} m
Cell 200 cell types		Conservation equations Passive flux equations Carrier mediated transport Electroneutrality constraints	Confocal microscopy 1μ or 10^{-6} m 2-photon microscopy 100nm or 10^{-7} m Electron tomography 1nm or 10^{-9} m
Organelle Cell membrane, mitochondria, nucleus endoplasmic reticulum, ribosomes, Golgi apparatus, centrioles Lysosomes, peroxisomes.		Continuum models Stochastic models	
Cell function Membrane receptors Membrane ion channels Signaling pathways Metabolic pathways Transport Motility Maintenance Cell cycle			
Proteins, carbohydrates & lipids		Molecular dynamics Quantum mechanics	Xray diffraction 1A or 10^{-10} m
Post-translational modifications		Markov models Boolean network models	
Protein folding			
Translation			
Post-transcriptional modifications			
Gene regulation			
Transcription			
Genes 19,000 genes			

Figure 17.1. The multi-scale modeling hierarchy.

II. ORGAN SYSTEMS: CURRENT PROGRESS AND FUTURE PLANS

The most highly developed “physiome” model of an organ is probably the heart model (Kohl et al. 2001), developed as a collaboration among research groups at the Universities of Auckland, Oxford, and California (San Diego). The geometry and fibrous-sheet structure of the heart has been measured and modeled with finite-element techniques (see Figures 17-2a- through c), which allow the equations of large deformation elasticity theory to be solved during the cardiac cycle under appropriate ventricular boundary conditions (Nash and Hunter 2000).

The coronary vessel tree (arteries and veins) is also modeled, and the Navier-Stokes equations governing blood flow are solved with coupling between the soft tissue mechanics of the ventricular wall and the deformable blood vessel wall (see Figure 17.1d) (Smith et al. 2000, 2002). The reaction-diffusion equations governing propagation of the wave of electrical excitation are also solved on this finite-element geometry and coupled to the mechanics through calcium release from RyR channels in the sarcoplasmic reticulum (stimulated by L-type calcium channels in the T-tubules) and through calcium binding to troponin-C proteins on the myofilaments (Hunter et al. 2003).

Currently, the heart model incorporates myofilament mechanics (Hunter et al. 1998; Stevens et al. 2003), the electrophysiology of ion channels (Noble and Rudy 2001; Smith and Crampin 2004), and the coupling between these (Nickerson et al. 2001), but does not yet include the regulation of the myofilament proteins and ion channels by signal transduction pathways—the next immediate target (Saucerman et al. 2004). Other developments currently underway include models of the heart valves and fluid mechanics of the ventricular cavities (coupled to wall mechanics), a structurally detailed model of the atria, and a model of the heart’s intrinsic nervous

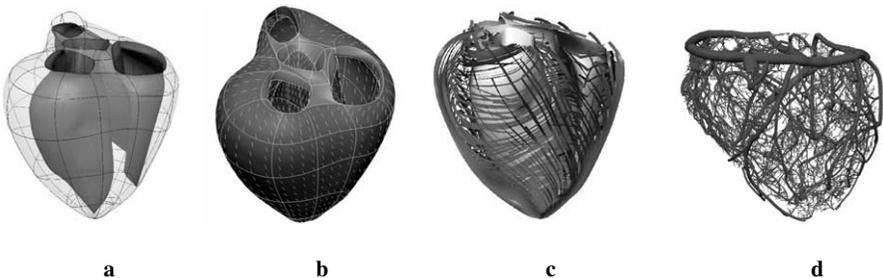


Figure 17.2. The geometry and microstructure of ventricular myocardium. (a) Finite-element surfaces fitted to measurements from the left and right ventricles of the pig heart. (b) 3D finite-element model of the heart. The elements use high-order basis functions (cubic Hermite) and therefore relatively few are required to provide an accurate description of ventricular anatomy. (c) Two layers of streamlines (one on the epicardial surface and one midway through the wall) are used to visualize the epicardial and midwall fiber directions. (d) The coronary arteries modeled from pig heart data. (Images a and b from Stevens & Hunter, copyright 2003, used by permission. Images c and d from Hunter et al., copyright 2005, used by permission.) (see color plate 14).

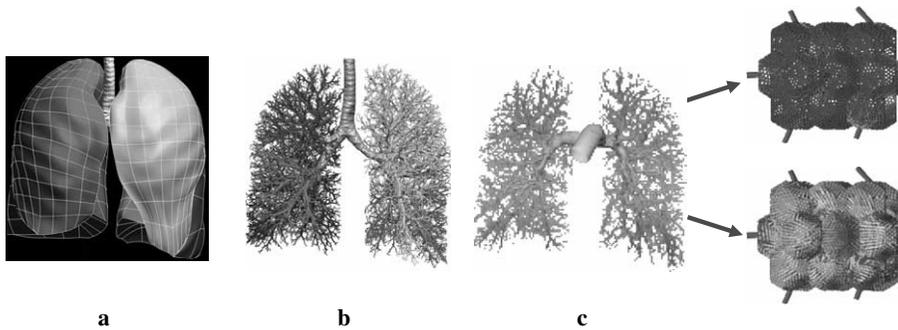


Figure 17.3. (a) Finite-element model of the five human lobes, derived by geometry fitting to high-resolution CT imaging of a normal human lung. (b) Airway model derived by geometry fitting the uppermost airways (down to generation 6–9) to CT imaging and “filling” the volumes in a using a branching algorithm. (c) Pressures predicted by solution of Navier–Stokes equations in the elastic pulmonary arterial tree, in an upright lung under normal gravity. The figures to the right show microcirculatory cell transit simulated in the apical and basal lung regions (see color plate 15).

system. The spatial distribution of ion channels in the heart is also being incorporated, as this is known to be important in the mechanisms of reentrant arrhythmia. An important application of the model is the ability to follow the consequences of an ion channel mutation, or drug-binding modification of a channel, on the pattern of activation in the intact heart (Noble 2002; Smith et al. 2004).

Another example of organ-scale modeling is shown in Figure 17.3. Here, the pulmonary circulation is modeled from the level of the whole organ (Figure 17.3a) down to the blood cells transiting through the capillary bed (Figure 17.3c). Fernandez et al. (2004) and Tawhai et al. (2004) show how subject-specific models of human (or animal) lobes can be routinely derived from high-resolution CT imaging of an individual subject, or for subjects imaged as part of the digital Lung Atlas (Li et al. 2003; Hoffman et al. 2004). These models are finite-element meshes in which equations for (for example) the large nonlinear deformation of the lung tissue are solved (Tawhai et al. 2005).

The geometry of the largest pulmonary arterial and venous blood vessels can also be defined from the same CT imaging (Burrowes et al. 2005a, 2005b), but for the modeling approach described here—in which *in vivo* geometric data is used in preference to more extensive *postmortem* data—an additional technique must be employed to model the non-imaged vessels such that the resulting trees are anatomically consistent. One approach is to use the maximum extent of the imaged blood vessels in combination with a volume-filling branching algorithm, such as has been used by Tawhai et al. (2004) to model both the human and ovine bronchial airway trees (Figure 17.3b).

This technique generates “accompanying” blood vessels within the geometry of the lobes, and additional steps can be included to model the extensive system of supernumerary vessels (Burrowes et al. 2005a, 2005b). At the microcirculatory level, the pulmonary capillary bed forms a dense mesh of short capillary segments,

wrapped over the alveoli in a continuous “sheet.” Burrowes et al. (2004) have modeled this physical relationship, representing the pulmonary capillary bed as a multi-segmented finite-element mesh generated over a 3D anatomically-consistent alveolar sac such that adjacent alveoli in the model share a single sheet of capillaries. The alveolar sac model used in the study was generated as a volume-filling structure. However, the method developed by Burrowes et al. would work equally well over an alveolar model derived directly from microstructural measurements.

The advantage of this approach is that it exploits the imaged geometry of the lobes and of the vascular or bronchial tree, producing models with realistic spatial relationships among airways, veins, arteries, and lung tissue. The models can therefore be readily exploited to couple multiple processes at the same physical scale (e.g., tissue mechanics and blood flow) or to couple over multiple scales (e.g., Newtonian flow in the large elastic blood vessels, and two-phase fluid transit in the microcirculation). For example, Burrowes et al. (2005b) solved the Navier-Stokes equations in elastic venous and arterial trees, subject to boundary conditions for pressures at the heart, pressures at the capillary bed, a gravitational acceleration vector, and transpulmonary pressure (Figure 17.3c).

Because the vascular models are “embedded” within the mesh of the lobes, simulation of soft tissue deformation of the lung provides (1) a change in geometry of the vascular models and (2) pressures acting on the vessels that result from expansion or recoil of the parenchymal tissue. The tissue pressures derived from the lung deformation and peripheral vessel flow calculations of the Navier-Stokes solution provide boundary conditions for simulation of red blood cell and neutrophil transit through the alveolo-capillary bed (Burrowes et al. 2004). Two capillary solutions are shown in Figure 17.3 for different tissue pressures. Conversely, the pressure drop predicted by the microcirculatory model provides updated pressure boundary conditions for re-resolution of flow in the arteries and veins.

Because the multi-scale models are separable, they can be coupled to different geometric or functional models than those that have been described here. For example, models that represent the average branching structure (Weibel 1963) and the average branching asymmetry (Horsfield et al. 1971)—or that have been measured directly from casts (Phalen et al. 1978; Schmidt et al. 2004)—can be used in place of the anatomically-based bronchial, venous, or arterial trees. The next stages of development will focus on interactions with other organs or muscle groups (the heart, diaphragm, and other respiratory muscles), and on incorporating the spatial distribution of airway and blood vessel smooth muscle.

A third example of organ-level modeling is shown in Figure 17.4 for the musculo-skeletal system. All bones and muscles of the human musculo-skeletal system have now been modeled, although not all yet have the detailed structure incorporated (muscle fiber directions and trabecular bone density). These models are being used for studying normal and abnormal gait, and for applications in surgical planning and virtual surgery training (Fernandez et al. 2001).

Now that all of the muscles and bones of the human body have been modeled, a database is being established to allow open Internet access to the models.

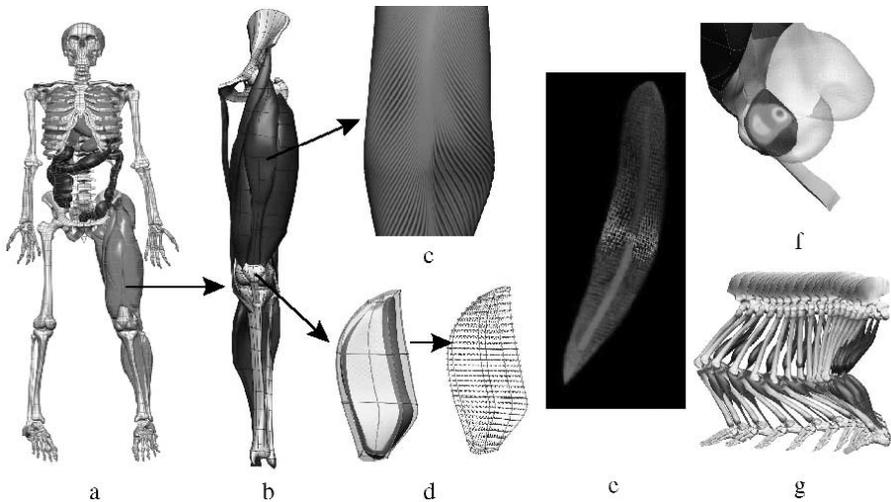


Figure 17.4. (a) A model of the human skeleton and (b) lower limb subset. (c) The fiber orientation in the rectus femoris muscle and (d) spatially varying material properties in the patella (red, cortical bone; blue, cancellous bone). (e) The muscle surface potential arising from a single nerve activation. (f) Investigation of loads on the patella cartilage during flexion and (g) model used to study the gait of a cerebral palsy subject. (From Hunter, Smith, Fernandez, and Tawhai, copyright 2005, by permission.) (see color plate 16).

The anatomies of the blood circulation and nerve pathways are also being incorporated.

III. OPEN STANDARDS AND OPEN-SOURCE TOOLS FOR THE PHYSIOME PROJECT

In this section we discuss the XML-based modeling standards and associated software tools currently being developed for the Physiome Project (see note in Acknowledgments). Much of the effort over the last five years has been in developing the CellML cell modeling standard and associated ontologies for describing biological systems (Cuellar et al. 2003; Lloyd et al. 2004) (see also www.cellml.org).

The web-accessible database of CellML models based on peer-reviewed journal publications currently contains about 300 models in the following categories: signal transduction pathway models, metabolic pathway models, cardiac electrophysiological models, calcium dynamics models, immunology models, cell cycle models, other cell-type electrophysiological models, smooth and skeletal muscle models, mechanical models, and constitutive law models. This database is now widely used by the biological modeling community, and is also being translated into SBML (Systems Biology Markup Language for biochemical reaction networks; see www.sbml.org) for use by the systems biology community.

An application programming interface (API) has been created for reading and writing the CellML files (see cellml.sourceforge.net). Open-source software tools are being developed for authoring CellML models, for rendering the models graphically, and for running model simulations (see www.bioeng.auckland.ac.nz/physiome/physiome_project.php). CellML uses Content MathML to represent the underlying mathematical relationships between/among model variables. A web-accessible ontology for accessing anatomical information and models is being set up using *Terminologia Anatomica* identifiers.

The graphical user interface (based on Mozilla/XUL), shown in Figure 17.5, is being developed for interacting with multi-scale Physiome models across all organ systems via the ontology database. It is intended both as a means of navigating the model databases and as a means of running model simulations and viewing simulation results.

IV. DISCUSSION

In this chapter we have illustrated anatomically and biophysically based models of three organ systems (heart and circulation, lungs, musculo-skeletal) of the twelve organ systems. Other organ systems either currently well developed or well underway are the digestive system, the skin (integument), the kidney and urinary system, and the lymphatic (immune) system. Others just starting are the endocrine system, the nervous system, the special-sense organs, and the male and female reproductive systems.

As these organ systems are modeled at the tissue/organ level and coupled to models at the cellular level (such as the cardiac ion channels in the heart model), it is becoming feasible to consider mapping these cell-, tissue-, and organ-level physiological processes into the proteome—the entire set of proteins. There are only about 200 cell types, each distinguished by the relative expression levels of various proteins that make up the cellular processes of metabolism, signal transduction, transport, motility, organization of the cytoskeletal structure, and operation of the cell cycle.

V. CONCLUSIONS

The greatest challenge of all will be modeling gene expression and how the 19,000 genes of the human genome are transcribed and, via splice variants, translated (and subsequently modified by the addition of carbohydrates and so on) into the 100,000 or so proteins that define biological function at a molecular scale. To achieve this goal requires a high degree of international collaboration based on open-source software and freely available user-friendly web-browser interfaces such as those we have described. The bioinformatics community has shown how this can be done for genomic and proteomic databases. It needs now to be done by the

Organ level view:

Controls for scene viewers. Access to tools such as data fitting etc. Each window can be toggled between reduced/full screen/icon.

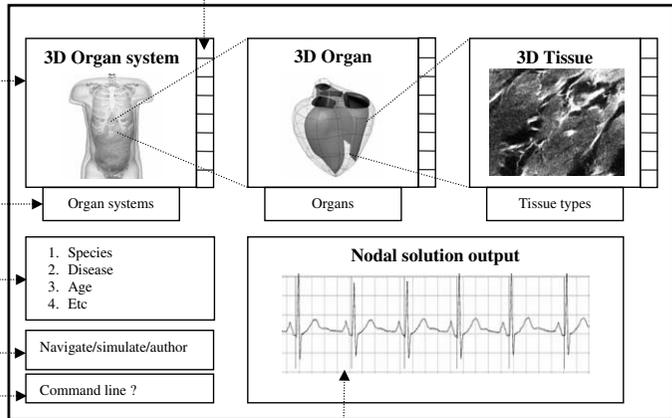
3D Cmgui window.
Note: ROI for scene on window to the the right is indicated.

Drop down menus where the options displayed for organs reflect the choice made for organ systems.

Selection of species, etc.

Switch between modes.

Read script files that run simulation (eqtns, b.c.s etc)



Simulation results shown for any node selected in above 3D windows

Cell level view:

The 3D cell & organelle windows show the spatial distributions of proteins, whereas the RH window shows the topological layout of pathways.

3D protein structures from PDB etc.

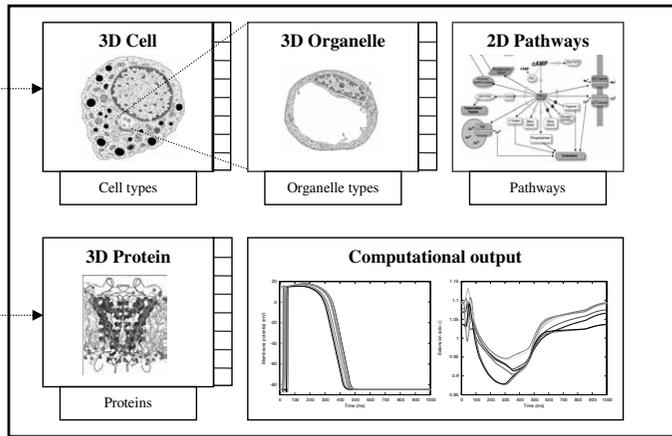


Figure 17.5. The graphical user interface for multi-scale modeling in the Physiome Project. Each window is linked to the one on its left to display information at a finer spatial scale. The models are retrieved from the ZOPE CellML/ontology database. The layout within the Mozilla window is defined by XUL. The 2D pathway graphs use SVG.

community of computational physiologists to extend these databases up to the level of physiological models that can be applied in a clinical setting.

ACKNOWLEDGMENTS

The authors are grateful to their colleagues in the Bioengineering Institute at the University of Auckland for help with many of the studies referred to. We would also

like to thank the Marsden Fund; the Foundation for Research, Science and Technology; and the Woolf-Fisher Trust for help with funding. IUPS is the International Union of Physiological Sciences (www.iups.org) and the IUPS Physiome Project is run under the auspices of the IUPS Physiome and Bioengineering Committee chaired by Peter Hunter (University of Auckland) and Aleksander Popel (Johns Hopkins University).

Note that the Physiome Project needs to support both public good science and industry involvement. All software tools (including APIs, authoring tools, visualization tools, graphical user interfaces, and simulation tools) developed for the Physiome Project are open source under a Mozilla or LGPL licence. That is, modifications must be put back into the public domain, but apart from this there is no restriction on commercial use (a commercial package, for example, can link the open-source library).

REFERENCES

- Bassingthwaighte, J. B. (1995). Towards modelling the human physiome. *Adv. Expt. Med. Biol.* **382**:331–339.
- Bassingthwaighte, J. B. (2000). Strategies for the Physiome Project. *Annals of Biomedical Engineering* **28**:1043–1058.
- Burrowes, K. S., Tawhai, M. H., and Hunter, P. J. (2004). Modeling RBC and neutrophil distribution through an anatomically based pulmonary capillary network. *Ann. Biomed. Eng.* **32**(4):585–595.
- Crampin, E. J., Halstead, M., Hunter, P. J., Nielsen, P. M. F., Noble, D., Smith, N. P., and Tawhai, M. (2004). Computational physiology and the Physiome Project. *Experimental Physiology* **89**:1–26.
- Burrowes, K., Hunter, P., and Tawhai, M. (2005a). Anatomically-based finite element models of the human pulmonary arterial and venous trees including supernumerary vessels. *J. Appl. Physiol.* **99**(2):731–738.
- Burrowes, K. S., Tawhai, M. H., and Hunter, P. J. (2005b). Physiology, Function, and Structure from medical images: Proceedings of SPIE, vol. 5746, 257–266.
- Cuellar, A., Lloyd, C., Nielsen, P., Halstead, M., Bullivant, D., Nickerson, D., and Hunter, P. (2003). An overview of CellML 1.1, a biological model description language. *Trans. Soc. Model. Sim. Int.* **79**(12):740–747.
- Fernandez, J. W., Mithraratne, P., Thrupp, S. F., Tawhai, M. T., and Hunter, P. J. (2004). Anatomically based geometric modeling of the musculo-skeletal system and other organs. *Biomechanics and Modeling in Mechanobiology* **2**(3):139–155.
- Hoffman, E., Clough, A., Christensens, G., Lin, C., McLennan, G., Reinhardt, J., Simon, B., Sonka, M., Tawhai, M., van Beek, E., and Wang, G. (2004). The comprehensive imaging-based analysis of the lung a forum for team science. *Acad. Radiol.* **11**(12): 1370–1380.
- Horsfield, K., Dart, G., Olsen, D. E., Filley, G. F., and Cumming, G. (1971). Models of the human bronchial tree. *J. Appl. Physiol.* **31**:207–217.
- Hunter, P. J., and Borg, T. K. (2003). Integration from proteins to organs: The Physiome Project. *Nature Reviews Molecular and Cell Biology* **4**:237–243.

- Hunter, P. J., McCulloch, A. D., and ter Keurs, H. E. (1998). Modeling the mechanical properties of cardiac muscle. *Prog. Biophys. Mol. Biol.* **69**:289–331.
- Hunter, P. J., Robbins, P., and Noble, D. (2002). The IUPS Human Physiome Project. *European J. Physiol.* **445**(1):1–9.
- Hunter, P. J., Pullan, A. J., and Smaill, B. H. (2003). Modelling total heart function. *Ann. Review of Biomedical Engineering* **5**:147–177.
- Hunter, P., Smith, N., Fernandez, J., and Tawhai, M. (2005). Integration from proteins to organs: The IUPS Physiome Project. *Mech. Ageing Dev.* **126**(1):187–192.
- Kohl, P., Noble, D., and Hunter, P. J. (eds.) (2001). The integrated heart: Modelling cardiac structure and function. *Philosophical Transactions of the Royal Society* **A359**:1047–1337.
- Li, B., Christensen, G. E., Hoffman, E. A., McLennan, G., and Reinhardt, J. M. (2003). Establishing a normative atlas of the human lung: Intersubject warping and registration of volumetric ct images. *Acad. Radiol* **10**:255–265.
- Lloyd, C., Halstead, M., and Nielsen, P. (2004). CellML: its future, present and past. *Prog. Biophys. Molec. Biol.* **85**(2/-3):433–450.
- Nash, M. P., and Hunter, P. J. (2000). Computational mechanics of the heart: From tissue structure to ventricular function. *J. Elasticity* **61**:113–141.
- Nickerson, D. P., Smith, N. P., and Hunter, P. J. (2001). A model of cardiac cellular electro-mechanics. *Phil. Trans. R. Soc. Lond.* **A359**:1159–1172.
- Noble, D. (2002). Modeling the heart: From genes to cells to the whole organ. *Science* **295**(5560):1678–1682.
- Noble, D., and Rudy, Y. (2001). Models of cardiac ventricular action potentials: Iterative interaction between experiment and simulation. *Phil. Trans. R. Soc. Lond.* **A359**(1783):1127–1142.
- Phalen, R. F., Yeh, H. C., Schum, G. M., and Raabe, O. G. (1978). Application of an idealized model to morphometry of the mammalian tracheobronchial tree. *Anat. Rec.* **190**:167–176.
- Saucerman, J. J., Healy, S. N., Belik, M. E., Puglisi, J. L., and McCulloch, A. D. (2004). Proarrhythmic consequences of a *kcnq1* akap-binding domain mutation: Computational models of whole cells and heterogeneous tissue. *Circ. Res.* **95**(12):1216–1224.
- Schmidt, A., Zidowitz, S., Kriete, A., Denhard, T., Krass, S., and Peitgen, H. O. (2004). A digital reference model of the human bronchial tree. *Comput. Med. Imaging Graph.* **28**(4):203–211.
- Smith, N. P., and Crampin, E. J. (2004). Development of models of active ion transport for whole-cell modeling: Cardiac sodium-potassium pump as a case study. *Progress in Biophysics & Molecular Biology* **85**(2/-3):387–405.
- Smith, N. P., Pullan, A. J., and Hunter, P. J. (2000). The generation of an anatomically accurate geometric coronary model. *Annals Biomedical Engineering* **28**:14–25.
- Smith, N. P., Pullan, A. J., and Hunter, P. J. (2002). An anatomically based model of transient coronary blood flow in the heart. *SIAM J. Applied Mathematics* **62**:990–1018.
- Smith, N. P., Nickerson, D. P., Crampin, E. J., and Hunter, P. J. (2004). Multi-scale computational modelling of the heart. *Acta Numerica* **13**:371–431.
- Stevens, C., and Hunter, P. J. (2003). Sarcomere length changes in a 3D mathematical model of the pig ventricles. *Progress in Biophysics & Molecular Biology* **82**(1/2/-3):229–241.
- Tawhai, M. H., Hunter, P. J., Tschirren, J., Reinhardt, J. M., McLennan, G., and Hoffman, E. A. (2004). CT-based geometry analysis and finite element models of the human and ovine bronchial tree. *J. Appl. Physiol.* **97**(6):2310–2321.
- Tawhai, M. H., Nash, M. P., Tschirren, J., Hoffman, E. A., and Hunter, P. J. (2005). Physiology, function, and structure from medical imaging: Proceedings of SPIE, vol. 5746, 84–91.
- Weibel, E. R. (1963). *Morphometry of the Human Lung*. Berlin: Springer-Verlag.

Subject Index

Page numbers followed by *t* indicate tables; page numbers followed by *f* indicate figures.

- A**
- Actin filaments, 151f, 152
- Active transport, 333
- Adaptive algorithms, error analysis and, 338
- AFM. See Atomic force microscopy
- Algorithms and Methods for the
 - Development of Biochemical Ontology-based Database Systems (Ambos), 19, 21, 27
- Ambos. See Algorithms and Methods for the Development of Biochemical Ontology-based Database Systems
- Analysis biological discovery, computational systems biology and, 4–6
- AnatML. See Anatomical Markup Language
- Anatomical Markup Language (AnatML), 27
- Annotated genome sequencing, 172
- APL. See Average path length
- Aptamers, 366
- Architectural features, of large network dynamics, 312–314
- ArrayExpress, 18
- Artificial entities, 30
- ASAP database, 199
- Atomic force microscopy (AFM), 366
- Attractors, cell fates as, 304–306
- Automated classification, of subcellular location patterns, 367–368, 368f
- Automated microscopy, pattern classification and, 368–369
- Automatic text summarization, 41
- Average path length (APL), 176, 179, 180f, 182f
 - degree distribution and, 177–179
- Axin response, Wnt pathway and, 100f
- B**
- B*-splines, example of, 209f
- BASE. See BioArray Software Environment
- Bayesian network
 - for gene networks modeling, 210–211
 - nonparametric regression and, 207–208
 - searching optimal, 214–217
- Bayesian network and nonparametric regression criterion (BNRC), 210
- Bayesian network literature, 211
- Bayesian network model, 205, 206
- Bayesian network, optimal, algorithm for, 216
- Bayesian network searching, greedy heuristics for, 215–217
- Bead arrays, 366
- Behavior analysis, 4, 5
- Behavior, emergent, 7
- Beta-catenin degradation, Wnt pathway and, 99f
- Beta-catenin response, Wnt pathway and, 99f
- Bifurcation analysis, 121t
- Binding site detection, algorithm for, 212
- Bio ThermoKinetics (BTK), 109
- BioArray Software Environment (BASE), 18
- Biochemical networks
 - dynamic, 7
 - top-down modeling of, 229–244
- Biochemical networks modeling, of
 - computational systems biology, 4–6
- Biochemical reaction networks
 - computational models of, 5, 127–142
 - kinetic models of, 130f
 - regulatory models of, 135
 - stoichiometric model of, 130f, 131f
 - structural models of, 129–135, 130f, 131f
- Biochemical systems theory (BST), 139
- Biocyc, 207
- BioD, 113
- Bioinformatics, statistical, 9
- Bioinformatics databases, public, 3

- Biological computing, advanced topics in, 9–10
 - Biological data, 1
 - databases for, 17–19
 - Biological data production, genome-wide, 205
 - Biological materials, experiment modules and, 22–25
 - Biological Pathway Exchange (BioPAX), 115
 - Biological system identification, mathematical theory of, 243
 - Biomaterials, 32
 - Biomedical literature, text
 - clustering/summarization in, 40–41
 - Biomedical literature data mining, experiment results of, 50–52
 - Biomedical ontologies, 41
 - Biomolecular computing, 9
 - BioPAX. *See* Biological Pathway Exchange
 - BioSig data model, 73
 - coarse representation of, 62f
 - Biosimulation tool, GON as, 224–225
 - BioSPICE, 4, 116, 117–118, 123, 207
 - BIOSSIM, 113, 115, 116
 - BioUML, 4, 118–119, 123
 - Bipartite graph, 49
 - Blood vessels, geometry of, 387, 388
 - BMAL1, 278, 281, 282
 - BNRC. *See* Bayesian network and nonparametric regression criterion
 - Boolean functions, 105, 106f, 312, 315
 - gene regulatory mechanism and, 307f
 - Boolean networks
 - discrete modeling methods and, 232
 - formalism in, 306–308
 - Bottom-up modeling, 5, 231
 - v. top-down modeling, 231–232
 - Bow-tie structure, 179–182, 181f
 - GSC and, 182–184
 - Braunschweig Enzyme Database (BRENDA), 17, 174–175
 - BRENDA. *See* Braunschweig Enzyme Database
 - BST. *See* Biochemical systems theory
 - BTK. *See* Bio ThermoKinetics
 - Building models, 104–105
- C**
- Canalizing function, 313
 - Cartesian grid, 343
 - Cell(s)
 - control mechanisms of, 128
 - multiscale representations of, 7–9
 - structural organization of, 150–152, 151f
 - Cell arrays, 1
 - Cell compartmentalization, spatiotemporal systems biology and, 327–329
 - Cell cycle
 - gene network estimation of, 213f
 - regulation of, 155
 - Cell differentiation, 293–294
 - Cell division cycle, 293–294
 - Cell fate dynamics, gene regulatory network and, 299–300
 - Cell fates, 293, 294, 294f, 296
 - as attractors, 304–306
 - cell types and, 297–298
 - molecular pathways and, 298–299
 - transitions, 294f
 - Cell heterogeneity, spatiotemporal systems biology and, 327–329
 - Cell Illustrator, 225
 - Cell marker, 295
 - Cell Markup Language (CellML), 4, 27, 103, 109, 110, 111, 112, 115, 120t, 122
 - Cell states, 293, 363, 364, 367
 - Cell surface receptors, 150
 - Cell transport routes, 359f
 - Cell types, cell fates and, 297–298
 - CellDesigner, 113
 - CellML. *See* Cell Markup Language
 - CellSim, 336, 337, 338, 339, 343–357
 - 3D kinase phosphatase model and, 349–352, 350f
 - code base of, 344
 - downloading/compiling with, 344–345
 - examples of, 345–355
 - Mpi parallelization in, 344
 - parameter optimization of, 354–355
 - sensitivity analysis for, 352–354
 - visualization of, 356–357, 356f
 - CellSimVis, 356–357, 356f
 - Cellular biology, 127

- Cellular communication, 149
- Cellular control analysis, 121t
- Cellular dynamics, systems level analysis of, 2
- Cellular networks modeling tools, mature/accessible, 120t
- Cellular simulator, 343–355
- Central services, 20, 20f
- Chemical kinetics, mathematics of, 331–332
- Chromatin protein-to-protein interaction network, data mining/extraction of, 50–52
- Circadian clock, mammalian
 computational model for, 278–280, 279f
 model of, 277–284
 syndromes and, 283f
- Circadian gene regulatory mechanism, HFPN model, 221, 222f
- Circadian oscillations
 bifurcation diagram of, 268, 269f, 273, 273f
 core molecular model for, 262
 deterministic v. stochastic simulations for, 265f
 model schemes for, 252, 253f
 in PER-TIM model, 256, 256f
 robustness of, 263–274, 266f
 ten-variable deterministic model for, 255–262, 257f, 258f, 259f
- Circadian regulatory network, multiple oscillations sources in, 280–281
- Circadian rhythm computational model
 molecular mechanisms of, 282–284
 physiological disorders and, 282–284
 sensitivity analysis of, 281–282
- Circadian rhythms, 6, 7
 computational models for, 249–287, 264t
 deterministic models for, 262
 dynamical bases of temperature compensation in, 261–262
 gene network modeling for, 220–224
 introduction to, 249–251
 irregular time series/trajectory of, 270, 272f
 long-term suppression of, 260, 261f
 mathematical modeling of, 250
 molecular mechanism of, 250
 non-developed stochastic models for, 274–277, 275t, 276t
 oscillatory behavior in, 281f
 stochastic models for, 262–277
- CLOCK–BMAL1, 277, 279, 282
- CLOCK proteins, 255, 277
- CMP. See Common multipotent precursor cell
- Colony organization, radiation effects and, 76f, 77f
- Comma-separated values (CSVs), 25
- Commission on Bioengineering in Physiology, 384
- Common Lisp, 92
- Common multipotent precursor cell (CMP), 305
- Compartmental models, 162
- Complex formation, signal transduction and, 156
- Complexity, 7
- Component integration, reactions and, 22
- Component/reaction module, 15, 20, 20f, 21
- Comprehensive Microbial Resource, 195
- Computational imaging, 8
- Computational models
 of biochemical reaction networks, 5, 127–142
 for circadian rhythms, 249–287, 264t
 for mammalian circadian clock, 278–280, 279f
- Computational systems biology, 6. See also Systems biology
 analysis biological discovery and, 4–6
 areas of, 3–9
 biochemical networks modeling of, 4–6
 challenges in, 9–11
 enabling technologies of, 3–4
 evolutionary diagram of, 2f
 multiscale representations, 7
 outlook for, 12–13
 scientific community and, 11
 software tools release for, 119t
- Computer-assisted modeling, 19
- Computing
 advanced topics in, 9–10
 with DNA, 10

- Connectionist theory, 121t
 Consensus subcellular location tree, 370f
 Constraint-based modeling, 195
 Contact information, 24f
 Continuous modeling, discrete modeling
 and, 240–242
 Control analysis, 140–141
 Control coefficients, 141
 CORBA, 116
 Core molecular model
 for circadian oscillations, 262
 molecular noise in, 262–263
 Coupling, 334
 Cry genes. *See* Cryptochrome genes
 Cryptochrome (Cry) genes, 221, 277
 Cryptochrome (Cry) proteins, 279, 284
 CSVs. *See* Comma-separated values
 CYC proteins, 255, 277
 Cycling behavior, 6
 Cytometry
 flow, 8, 365
 slide-based, 365
 Cytomics, 8, 357, 363–376
 abstract on, 363
 computational imaging in, 364–371, 368f,
 370f
 conclusions on, 375–376
 data analysis and, 372–373
 definition of, 364
 discussion of, 373–375
 high-content image analysis and, 371
 information density and, 365
 innovative preparation in, 366–367
 introduction to, 363–364
 labeling techniques in, 366–367
 location proteomics and, 367–371
 single-cell image analysis, 364–367
 system-wide data analysis and, 373
 Cytomics analysis
 high-throughput imaging and, 371
 in tissues, 371
 Cytoskeleton, 152
 Cytosolic proteins, 157
- D**
- DAEs. *See* Differential algebraic equations
 DAG. *See* Diacylglycerol
 DARPA, 117
- Data
 generation of, 4
 integration, 16, 21, 22. *See also*
 Databases
 standards for, 26
 Data discretization, 239–240
 different levels of, 233f
 Data mining/extraction, of chromatin
 protein-to-protein interaction network,
 50–52
 Database content, guided workflow
 annotation/exploration of, 63f
 Databases
 for biological data, 17–19
 components of, 20, 20f
 for elements, 17
 for experimental data, 18–19
 for images, 59, 60f, 371
 for information resources, 17–18
 integrated, 3, 21, 22
 for location proteomics, 370
 for modeling, 19
 for systems biology, 15–34
 DBcat. *See* Public Catalog of Databases
 DDEs. *See* Delay differential equations
 De novo experimental designs, systems
 biology and, 10–11
 Death-inducing signaling complex (DISC),
 162
 Decomposition, of metabolic networks,
 185, 185f, 196f
 Degree distribution, APL and, 177–179
 Delay differential equations (DDEs), 163
 Dephosphorylation, 158
 Deterministic hybrid, 105, 106f
 Deterministic models, 250
 for circadian rhythms, 262
 Diacylglycerol (DAG), 157
 Differential algebraic equations (DAEs),
 106, 106f, 163
 quantitative models and, 107–108
 Diffusion, 8
 spatiotemporal systems biology and,
 329–330
 Diffusion equation, mathematics of,
 330–343
 Diffusion operator, 335–336
 second order, 338

- Dimensionality reduction, 47
- DISC. See Death-inducing signaling complex
- Discrete modeling methods, 232–239
- Boolean networks and, 232
 - continuous modeling and, 240–242
 - finite-state polynomial models and, 234–239
 - mathematical theory for, 242
 - multi-state discrete models and, 233–234
- DNA, computing with, 10
- DOT code, 89
- Downloading/compiling, with CellSim, 344–345
- Drosophila circadian clock
- experimental observations of, 251–252
 - modeling of, 251–262, 253f, 254f, 257f, 258f, 259f
- Drugs
- discovery companies for, 365
 - new availability of, 2
- DVD interface, of polynomial models, 239f
- Dynamic Bayesian networks, gene network estimation and, 213–214
- Dynamic biochemical networks, 7
- E**
- E1. See Ubiquitin-activating enzyme
- E2. See Ubiquitin-conjugating enzyme
- E3. See Ubiquitin ligase
- EAV. See Entity attribute model
- ED pathways. See Entner-Doudoroff pathways
- EFMA. See Elementary flux mode analysis
- EGF-R. See Epidermal growth factor receptor
- Electrophoretic mobility shift assay (EMSA), 165
- Elementary flux mode analysis (EFMA), 131, 132, 134
- Elements, databases for, 17
- Eliza-like patterns, 43
- Elliptic regions, of segmentation process, 65–66
- Emergent behavior, 7, 104
- Emerging phenotypes, multiscale representations of, 7–9
- EMSA. See Electrophoretic mobility shift assay
- Enabling Grid for E-science (EGEE), 10
- Endoplasmic reticulum, 152
- Ensemble approach, large network dynamics of, 311–312
- Entity attribute model (EAV), 23
- Entner-Doudoroff (ED) pathways, 183
- ENZYME databank, 194
- Enzyme information, 1
- Enzyme-reaction relationships, 6
- EPA. See Extreme pathway analysis
- Epidermal growth factor receptor (EGF-R), 154, 159f
- ERK. See Extracellular signal regulated kinase
- Error analysis, adaptive algorithms and, 338
- Euler integrators, 336
- Evolution, progressive, 7
- Exchange standards, 4
- ExpASy. See Expert Protein Analysis System
- Experiment modules, 15, 20, 20f, 22–26
- biological materials and, 22–25
 - experimental data and, 22–25
- Experimental data, databases for, 18–19
- Experimental setup, 24f
- Experimental values, 24f
- Expert Protein Analysis System (ExpASy), 19
- Extracellular matrix (ECM), 59, 74
- Extracellular signal regulated kinase (ERK), 159f, 160
- Extreme pathway analysis (EPA), 131, 132, 134
- F**
- Familial advanced sleep/wake cycle syndrome (FASPS), 282, 283
- FASPS. See Familial advanced sleep/wake cycle syndrome
- FBA models, 132
- FCS. See Fluorescence correlation spectroscopy
- Field Markup Language (FieldML), 27
- FieldML. See Field Markup Language
- Finite-state polynomial models, discrete modeling methods and, 234–239
- Finite state set, 7
- FLIM. See Fluorescence lifetime imaging

- Flow cytometry, 8, 365
- Fluorescence correlation spectroscopy (FCS), 358
- Fluorescence lifetime imaging (FLIM), 365
- Fluorescence microscopy, 365
- Fluorescence recovery after photo-bleaching (FRAP), 358, 370
- Fluorescence resonance energy transfer (FRET), 358, 365, 370
- Flux analysis, 132, 133
- FlyBase, 18
- Forward time-centered space algorithm (FTCS), 335
- FRAP. *See* Fluorescence recovery after photo-bleaching
- Frequency analysis, 121t
- Frequent term set-based concept clustering, 47–48
- FRET. *See* Fluorescence resonance energy transfer
- FTCS. *See* Forward time-centered space algorithm
- Functional analysis, metabolic networks and, 184–186, 185f, 186t
- Functional modules, in metabolic network, 185f
- Functional network reconstructions, 194, 195f
- G**
- GAP. *See* GTPase-activating protein
- GATA-1, 305
- GE-Miner. *See* Gene Expression Miner
- GEF. *See* Guanine nucleotide-exchange factor
- GEISHA, 40
- GenBank, 17
- Gene(s)
- regulatory relationships between, 4
 - variants of, 1
- Gene-enzyme relationships, 6
- Gene expression, discrete behavior of, 304f
- Gene Expression Miner (GE-Miner), 46
- Gene Expression Omnibus (GEO), 18
- Gene expression profiles, in gene expression state space, 295–297
- Gene expression state space
- biological implications of, 317–319
 - gene expression profiles in, 295–297
- Gene network(s), 6, 205–225
- petri-net-based modeling of, 217–225
- Gene network estimation
- advanced methods for, 211–217
 - algorithm for, 212
 - cell cycle, 213f
 - dynamic Bayesian networks and, 213–214
 - general framework for, 211–212
 - from microarray gene expression data, 207–211
 - multi-source biological information for, 211–213
 - promoter regions of, 212–213
 - protein-to-protein interactions in, 213
- Gene network information, creation of, 205
- Gene network modeling
- Bayesian networks for, 210–211
 - for circadian rhythms, 220–224
- Gene Network Sciences, 113, 120
- Gene Networks International (GNI), 225
- Gene Ontology (GO), 41
- Gene products, biomolecular complexity of, 2
- Gene regulatory mechanism, boolean function and, 307f
- Gene regulatory networks, 135
- cell fate dynamics and, 299–300
- Generative models, location proteomics and, 370–371
- Genome expression, 57
- Genome information, metabolic network reconstruction from, 169–187, 173f
- Genome International Sequencing Consortium, 1
- Genome-scale metabolic networks, 170
- reconstruction of, 170–174
- Genome-scale model, 66
- Genome-scale network topology, 314–316
- Genome-wide data/computational issues, 206f
- Genome-wide networks, v. small circuits, 309–310
- Genomic Object Net (GON), 207, 217, 224, 225
- as biosimulation tool, 224–225

- Genomics, 1
 sequencing of, 4
- GEO. *See* Gene Expression Omnibus
- Gepasi, 115, 116, 120t, 207
- GFP-tagged proteins. *See* Green fluorescent protein-tagged proteins
- Giant strong component (GSC), 180, 181f, 183, 185
 in bow-tie structure and, 182–184
 in metabolic network, 183f
- Glycine/serine (GS), 158
- Glycogen synthase kinases 3 (GSK-3), 155
- Glycolysis, 108f
- GNI. *See* Gene Networks International
- Gnu Public License (GPL), 343
- GO. *See* Gene Ontology
- Golgi apparatus, 152
- GON. *See* Genomic Object Net
- GPL. *See* Gnu Public License
- Gray-Scott model, 347f, 348
- Greedy heuristics, for Bayesian network searching, 215–217
- Greedy hill climbing, 215–217
- Green fluorescent protein (GFP)-tagged proteins, 165, 166, 358
- Growth factor stimulation, 156
- GS. *See* Glycine/serine
- GSK-3. *See* Glycogen synthase kinases 3
- GTP. *See* Guanosine triphosphate
- GTPase-activating protein (GAP), 156
- Guanine nucleotide-exchange factor (GEF), 156
- Guanosine triphosphate (GTP), 156
- H**
- Haplotype Mapping project, 1
- Harmonic cuts, of segmentation process, 66–67
- HDN. *See* Hybrid dynamic net
- Heart model, 386, 386f
- Hematopoietic stem cell (HSC), 305
- Hepatocyte regeneration, signaling pathways in, 159f
- Hessian terms, 341
- HFPN. *See* Hybrid Functional Petri Net
- High-content image analysis, cytomics and, 371
- High-dimensional attractors, experimental evidence for, 316–317
- High-throughput imaging, cytomics analysis and, 371
- High-throughput reconstruction, 172
- HPN. *See* Hybrid Petri Net
- HSC. *See* Hematopoietic stem cell
- HTML. *See* HyperText Markup Language
- Human genome, 296
- Human Genome Project, 1, 364
- Human Proteome Organization (HUPO), 26
- HUPO. *See* Human Proteome Organization
- Hybrid dynamic net (HDN), 219
- Hybrid Functional Petri Net (HFPN), 207, 217, 218, 219, 221
 elements of, 218f
 modeling with, 218
 operon model with, 219–225, 220f
- Hybrid Functional Petri Net (HFPN) model, 221–223
 of circadian gene regulatory mechanism, 221, 222f
 new interaction in, 224–225
 simulation inconsistency in, 224
 simulation result of, 224, 224f
- Hybrid Petri Net (HPN), 218
- Hypergraph partitioning algorithm, 48
- HyperText Markup Language (HTML), 27
- I**
- Image collage, query results for, 64f
- Image restoration techniques, 366
- Imaging protein kinetics, 369–370
- Informatics, of integrated imaging informatics, 60–63
- Information resources, databases for, 17–18
- Inhibitory arc, 219
- Integrated imaging informatics, 57–77
 applications of, 73–76
 architecture of, 59–60
 informatics of, 60–63
 introduction to, 57–59
 iterative voting and, 72
 layers of, 59–60
 quantitative analysis of, 63–73
 voting-based techniques of, 69–73

- Integrated regulatory models, 6
 metabolic models and, 191–202
- Integrative database solution, project
 workflow overview of, 33f
- Interactive database architecture, 20, 20f
- Interfering RNA (RNAi), 23
- International Union of Biochemistry and
 Molecular Biology (IUBM), 170
- International Union of Physiological
 Sciences (IUPS), 384
- International Union of Physiological
 Sciences (IUPS) Council, 384
- International Union of Physiological
 Sciences (IUPS) Physiome Project, 9.
 See also Physiome Project
 abstract of, 383
 conclusions on, 390–391
 discussion of, 390
 GUI for, 390, 391f
 introduction to, 383–384
 open-source tools for, 389–390
 open standards for, 389–390
 progress/plans of, 383–391
- Irreversible reactions, 172
- Iterative voting, integrated imaging
 informatics and, 72
- IUBM. See International Union of
 Biochemistry and Molecular Biology
- IUPS. See International Union of
 Physiological Sciences
- J**
- Jacobian, 338
- Jacobian, extended, calculation of, 342–343
- Jacobian, original, evaluation of, 341–342
- JAK. See Janus kinase
- JAK-STAT pathway. See Janus kinase-Signal
 transducer and activator of
 transcription pathway
- Janus kinase (JAK), 158
- Janus kinase-Signal transducer and
 activator of transcription (JAK-STAT)
 pathway, 158
- Jarnac/Designer, 113, 120t
 visual format of, 114f
- Java Web Simulation (JWS), 27
- JavaScript Query-Handler, 92
- JWS. See Java Web Simulation
- K**
- KEGG. See Kyoto Encyclopedia of Genes
 and Genomes
- Kernel matrix, 131, 134
- Kernel topography, 71f
- KineCyte, 113
- Kinetic measurements, challenges of,
 165–166
- Kinetic models
 analysis of, 136–137
 of biochemical reaction networks,
 130f
 simulation methods for, 137–140, 138f
- Kohn format, 113
- Kyoto Encyclopedia of Genes and
 Genomes (KEGG), 17, 21, 170, 171,
 176, 184, 185, 207
- L**
- Language-independent architecture, valis
 and, 82
- Large network dynamics
 architectural features of, 312–314
 of ensemble approach, 311–312
- Layered expression imaging, 366
- Ligand-binding, 157
- LIGAND database, 171
- Light pulse, phase shifting by, 259f
- Limit cycle, 312
- Linear-noise approximation (LNA), 139
- Linear pathways, 132
- Link matrix, 134
- LNA. See Linear-noise approximation
- Location pattern, protein clustering by, 369,
 370f
- Location proteomics, 364
 cytomics and, 367–371
 generative models and, 370–371
- M**
- MAGE-ML. See Micro Array Gene
 Expression Markup Language
- MAGE-OM. See Micro Array Gene
 Expression object model
- MAGE. See Micro Array Gene Expression
- MAGEC-ML. See Micro Array Gene
 Expression Communication markup
 language

- MAGEC. See Micro Array Gene Expression Communication
- Magnetic nanobeads, 366
- MALDI. See Matrix-Assisted Laser Desorption/Ionization
- Mammalian circadian genetic control mechanism, 221
- MAP kinases. See Mitogen-activated protein kinases
- MAPPER database, 17
- Mathematical analysis
 - of chemical kinetics, 331–332
 - of diffusion equation, 330–333
 - of reaction-diffusion equation, 333–343
- Mathematical modeling
 - of metabolic networks, 174–175, 175f
 - of signaling pathways, 161–165
- Matlab, 120
- Matrix-Assisted Laser Desorption/Ionization (MALDI), 23
- MCA. See Metabolic control analysis
- MedLine, 39, 42
- MedLine abstracts, keyword searches of, 51t
- MedMiner, 40
- Message broker, 118f
- Message-passing interface (MPI), 343
- Message-passing interface parallelization, in CellSim, 344
- Meta-heuristics, 9–10
- Metabolic control analysis (MCA), 140, 141
- Metabolic models, 6
 - experimental/computational data interpretation in, 198–201
 - integrated regulatory models and, 191–202
- Metabolic networks
 - analysis/simulation of, 195–196
 - decomposition of, 185, 185f, 196f
 - functional analysis and, 184–186, 185f, 186t
 - from genome information, 169–187, 173f
 - mathematical representation of, 174–175, 175f
 - metabolite graph representation of, 177, 177f
 - to modules, 184–186, 185f, 186t
 - network reconstruction and, 193–195
 - predicted/measured outputs of, 196
 - reaction content of, 178
 - reconstruction/representation of, 6, 170–177
 - structural analysis of, 177–186
- Metabolite graph representation, 178
 - of metabolic networks, 177, 177f
- Metacyc, 184
- Metadata, 30
 - model and, 29f
- MGED. See Microarray Gene Expression Data group
- MGI. See Mouse Genome Informatics
- MIAME. See Minimum Information About a Microarray Experiment
- Michaelis-Menten's formulation, 85, 162, 329
- Micro Array Gene Expression (MAGE), 61
 - ontology terms of, 63
- Micro Array Gene Expression Communication (MAGEC), 83, 84
- Micro Array Gene Expression Communication markup language (MAGEC-ML), 84
- Micro Array Gene Expression Markup Language (MAGE-ML), 18, 26, 83, 84
- Micro Array Gene Expression object model (MAGE-OM), 26, 83
- Microarray data analysis, 25, 205, 206
 - text mining enrichment for, 52
- Microarray data interpretation, ontology-enhanced text
 - clustering/summarization for, 46–50
- Microarray gene expression data, gene network estimation from, 207–211
- Microarray Gene Expression Data group (MGED), 18
- Micrographia* (Hooke), 79, 80
- MicroRNA, 157
- Microscopy, multichannel, 63–64
- Microscopy, automated, pattern classification and, 368–369
- Microtubules, 152
- Minimal cells, 10
- Minimum Information About a Microarray Experiment (MIAME), 18, 26
- Mitogen-activated protein (MAP) kinases, 155, 159f, 160

- Model(s) *See also* biochemical, gene, metabolic, and signaling networks
 Boolean, 233, 238, 293
 components of, 20, 21
 continuous, 301
 discrete vs. continuous, 240
 human-readable formats of, 113
 integration, 130, 191, 386. *See also* multi-scale sinetic, 127
 metadata and, 29f
 pathways, 5, 88, 103, 107, 149, 169
 polynomial, 234
 state space of, 241f
 stochastic, *see* stochastic models
 visualization of, 113, 345
 wiring diagram of, 241f
- Model analysis, 121
 methods of, 121t
- Model building, 161
- Model cell and cultures, 11, 59, 73
- Model databases, 15, 114
- Model-driven discovery, 198f
- Model fitting, validation and, 121–122
- Model integration, 3, 130, 191, 383
- Model interchange standards, 103
- Model modules, 15, 20, 20f, 26–31
 database model storage for, 27–30
 models/standards of, 26–27
 SBW, 116
 simulation for, 30–31
- Model organisms, 49
- Model predictions, 199
- Modeling, databases for, 19
- Modeling approaches, 5
- Modeling biological systems, mathematical techniques for, 106f
- Modeling cycle, 5
- Modular interaction domains, 156
- Modules *See also* model modules
 metabolic networks to, 184–186, 185f, 186t
 pathways, 2f
 reaction, 85
- Molecular biology, 127, 128
- Molecular mechanisms, of circadian rhythm
 computational model, 282–284
- Molecular noise, in core molecular model, 262–263
- Molecular pathways, cell fates and, 298–299
- Molecules, regulatory relationships
 between, 4
- Mouse Genome Informatics (MGI), 18
- MPI. *See* Message-passing interface
- Multi-scale
 data correlation, 373
 representation of cells, 7–9
 representation of emerging phenotypes, 7–9
 representation of organs, 383f
 modeling hierarchy, 385f
- Multi-source biological information, for gene network estimation, 211–213
- Multi-state discrete models, discrete modeling methods and, 233–234
- Multicellularity, 8
 analysis of, 364
 multistability and, 293–320
- Multiparametric single-cell analysis, 372
- Multiple oscillations sources, in circadian regulatory network, 280–281
- Multiple shooting technique, 163
- Multistability, 8
 multicellularity and, 293–320
 in small gene circuit, 300–304, 302f
- N**
- N matrix, 132
- NAD, 133, 134
- NADH, 133
- National Center for Biotechnology Information (NCBI), 17
- Natural language processing, 4
 introduction to, 39–40
 ontology-enhanced biomedical literature mining and, 39–54
- Near infrared Raman spectroscopy, 365
- Network
 time series plot of, 241f
 wiring diagram of, 241f
- Network connectivity, examples of, 180f
- Network global connectivity, 179–182
- Network reconstruction; *See also* models
 metabolic networks and, 193–195
 regulatory networks and, 196–197
 biochemical, 127
- NFκB, 164

- NLP
for automatic pattern
generation/evaluation, 42–46
ontology-enhanced biomedical literature
mining and, 42–50
- Non-developed stochastic models, for
circadian rhythms, 274–277, 275t, 276t
- Nonenzyme-catalyzed reactions, 172
- Nonlinear relationships, 128
- Nonparametric regression
Bayesian networks and, 207–208
introduction to, 208–209
- Normal arc, 218
- NP-complete, 10
- NP-hard, 9–10
- Numerical analysis, of reaction-diffusion
equation, 333–343
- NYU microarray database (NYUMAD), 82,
83–84, 84
- NYUISM, 84
- NYUMAD. See NYU microarray database
- O**
- ODEs. See Ordinary differential equations
- OME. See Open Microscopy Environment
- OMIM. See Online Mendelian Inheritance
in Man
- Online Mendelian Inheritance in Man
(OMIM), 17
- Ontology-enhanced biomedical literature
mining, 4
natural language processing and,
39–54
NLP and, 42–50
significant terms of, 53t
- Ontology-enhanced text
clustering/summarization, for
microarray data interpretation, 46–50
- Open Microscopy Environment (OME), 19,
26, 61
- Open-source tools, for IUPS Physiome
Project, 389–390
- Open standards, for IUPS Physiome Project,
389–390
- Operator splitting, 334–335
- Operon model, with HFPN, 219–225,
220f
- Optical coherence tomography, 365
- Ordinary differential equations (ODEs), 85,
87, 105, 106f, 116, 139, 162, 164, 230,
334
- Organ models, 385f
of human skeleton, 388, 389f
- Organ scale modeling, 387, 387f
- Organ systems
current progress on, 386–389
future plans for, 386–389
- Oscillatory behavior, in circadian rhythms,
281f
- P**
- Partial differential equations (PDEs), 106f,
107, 139, 163, 230
- Passive transport, 333
- Pathways, 5, 103, 132, 149, 159, 183, 298;
See also models
- Pattern classification, automated
microscopy and, 368–369
- Pattern generation, 43–46
- PDEs. See Partial differential equations
- PDGF-R. See Platelet-derived growth factor
receptor
- PDK1. See Phosphoinositide-dependent
kinase
- PEDRo. See Proteomics Experimental Data
Repository
- PER-CRY complex, 277, 281, 282
- PER genes, 221, 277
- PER model, 255, 256
sustained oscillations of, 254, 254f
- PER protein, overexpression of, 252
- PER protein circadian oscillations, core
deterministic model for, 252–255, 253f,
254f
- PER-TIM complex, 7, 255, 256, 257f, 276,
277
- PER-TIM model, 259, 260, 276
circadian oscillations in, 256, 256f
phase locking in, 258f
- Petri-net-based modeling, of gene
networks, 217–225
- Petri nets, 7, 218
- Phase locking, in PER-TIM model, 258f
- Phase shifting, by light pulse, 259f
- Phase transitions, 314
- Phosphoinositide (PI), 155, 159f

- Phosphoinositide-3,4-bisphosphate (PtdIns-3,4-P₂), 156
- Phosphoinositide-3,4,5-trisphosphate (PtdIns-3,4,5-P₃), 156
- Phosphoinositide-dependent kinase (PDK1), 159f, 161
- 5-Phosphoribosyl diphosphate (PRPP), 183
- Phosphorylation, 153–154
- Physiological disorders, circadian rhythm computational model and, 282–284
- Physiome, 384
- Physiome Project, 384. *See also* International Union of Physiological Sciences (IUPS) Physiome Project
- PI. *See* Phosphoinositide
- Platelet-derived growth factor receptor (PDGF-R), 154
- Platforms, for systems biology, 115–119
- Polynomial models, DVD interface of, 239f
- Priming stage, of signaling pathways, 159f
- Principal component analysis (PCA), 47
- Progressive evolution, 7
- Project management, UML for, 30, 31f
- Project workflow overview, of integrative database solution, 33f
- Proliferation assay, 74f
- Proliferation stage, of signaling pathways, 159f
- Promoter regions, of gene network estimation, 212–213
- Protege programming model, 62
- Protein cell arrays, 1
- Protein clustering, by location pattern, 369, 370f
- Protein kinases, 154
- Protein phosphatases, 155
- Protein-to-protein interactions, 4
in gene network estimation, 213
- Proteolytic cleavage/degradation, signal transduction and, 156–157
- Proteomics Experimental Data Repository (PEDRo), 18
- PRPP. *See* 5-Phosphoribosyl diphosphate
- PtdIns-3,4-P₂. *See* Phosphoinositide-3,4-bisphosphate
- PtdIns-3,4,5-P₃. *See* Phosphoinositide-3,4,5-trisphosphate
- Public Catalog of Databases (DBcat), 19
- PubMed, 3, 17
- Pulmonary circulation, modeling of, 387, 387f
- Pure-reaction description, 139
- Pysces, 120t
- Python code, 89, 90, 91
- Q**
- Q-dot nanoparticles. *See* Quantum-dot nanoparticles
- Quantum (Q-) dot nanoparticles, 358, 366
- Quasi-equilibrium reactions, 141
- R**
- Radial symmetry, 69–70
- RDF. *See* Resource description framework
- Reaction(s)
component integration and, 22
definition of, 21
- Reaction content, of metabolic networks, 178
- Reaction-diffusion equation
mathematical analysis of, 333–343
numerical analysis of, 333–343
- Reaction information, 29f
- Reaction operator, 336–337
- Reactions models, 100
- Regularized centroid transform,
segmentation process and, 67–68
- Regulatory models, of biochemical reaction networks, 135
- Regulatory networks, 196–198
analysis/simulation in, 197
measured/predicted outputs of, 198
network reconstruction and, 196–197
- Regulatory rules, perturbation approach to, 201f
- Relevant novelty, 48
- Resource description framework (RDF), 27, 110
- Response coefficients, 141
- RNAi knockdowns, 10
- RNAi. *See* Interfering RNA
- Robustness, of circadian oscillations, 263–274, 266f
- Rosenbrock method, 339
- Runge-Kutta methods, 337

- S**
- S/MARt. *See* Scaffold/Matrix Attached Regions database
- SAGE. *See* Serial Analysis of Gene Expression
- SBML. *See* Systems Biology Markup Language
- SBW. *See* Systems Biology Workbench
- Scaffold/Matrix Attached Regions database (S/MARt), 17
- Scalable and Portable Information
Extraction (SPIE), 42–46
architecture, 42f
experimental results of, 51t
- Scalable vector graphics (SVG), 63
- Scale-free network, 178
- Scanning near-field optical microscopy (SNOM), 366
- Scientific community, computational systems biology and, 11
- Screening technologies, 358
- Second harmonic imaging, 365
- Second messenger, 157
- Segmentation process, 65f
elliptic regions of, 65–66
harmonic cuts of, 66–67
regularized centroid transform and, 67–68
representation/classification in, 69
variational approach to, 65–69
vector field partitioning in, 68–69
- SELDI. *See* Surface-Enhanced Laser Desorption/Ionization
- Sensitivity analysis
for CellSim, 352–354
of circadian rhythm computational model, 281–282
- Sentence extraction, 48
- Serial Analysis of Gene Expression (SAGE), 18
- SH. *See* Src-homology
- Signal transducer and activator of transcription (STAT), 158, 164
- Signal transduction, 6
biological foundations of, 149–166
complex formation and, 156
concepts/principles of, 150–158
pathways of, 150
proteolytic cleavage/degradation and, 156–157
- Signal transmission, from cell surface to nucleus, 152–156, 153f
- Signaling pathways
in hepatocyte regeneration, 159f
mathematical modeling of, 161–165
modeling approaches to, 161–163
priming stage of, 159f
proliferation stage of, 159f
termination stage of, 159f
- Simpatica, 4, 79–101
GUI design of, 89f
introduction to, 80–81
theoretical basis for, 84–87
valis and, 81–87
- Simple gene regulatory circuit, 302f
- Simulated gene regulatory network, 240, 241f
- Simulation
by computational models, 16, 349; *See* also models
for kinetic models, 137–140, 138f
model modules and, 30–31
setup for, 29f
- Single-cell PCR, 366
- SMAD signaling pathway, 158, 159f
- Small circuits, v. genome-wide networks, 309–310
- Small gene circuit, multistability in, 300–304, 302f
- SMD. *See* Stanford Microarray Database
- SNOM. *See* Scanning near-field optical microscopy
- SOCS. *See* Suppressor of cytokine signaling
- SOP. *See* Standard operating procedures
- Spatiotemporal imaging, 357–359, 359f
- Spatiotemporal sensitivity analysis, 338–341
- Spatiotemporal systems biology, 327–360
abstract on, 327
cell compartmentalization and, 327–329
cell heterogeneity and, 327–329
diffusion and, 329–330
introduction to, 327–330
theory of, 330–343
- Species information, 29f
- Spectroscopic optical coherence tomography, 365

- SPIE. See Scalable and Portable Information Extraction
- Src-homology (SH), 156
- Standard clustering algorithms, 47
- Standard operating procedures (SOP), 23, 24f, 25, 32
- Standards for Reporting Enzymology Data (STREND A), 26
- Stanford Microarray Database (SMD), 18
- STAT. See Signal transducer and activator of transcription
- State-explosion problem, 100–101
- State space, 100–101
- Steady state flux pattern, 132
- STED microscopy. See Stimulated emission depletion microscopy
- Stimulated emission depletion (STED) microscopy, 366
- Stochastic models, 251
of biological processes, 332–333
- Stochastic simulations, parameters for, 267t
- Stoichiometric model, of biochemical reaction networks, 130f, 131f
- STREND A. See Standards for Reporting Enzymology Data
- Strongly connected components, distribution of, 181f
- Structural analysis, 121t
- Structural models, of biochemical reaction networks, 129–135, 130f, 131f
- Structural proteomics, 4
- Subcellular location patterns, automated classification of, 367–368, 368f
- Suppression, 260
- Suppressor of cytokine signaling (SOCS), 158, 159f
- Surface-Enhanced Laser Desorption/Ionization (SELDI), 23
- Sustained oscillations, of PER model, 254, 254f
- SVG. See Scalable vector graphics
- SWISS-2DPAGE, 18
- Synthetic biology, 11
- SysBio-OM. See Systems Biology Object Model
- System behavior, 4
- System identification, 4, 5
- System-wide data analysis, cytomics and, 373
- Systems biology, 8, 127. See also Computational systems biology
aim of, 16
applications for, 119–122
database solution for, 19–31
databases for, 15–34
de novo experimental designs and, 10–11
definition of, 104
discovery approach of, 191–193, 192f
goal of, 57–58
levels of, 58f
platforms for, 115–119
research workflow, 16, 16f
supports of, 16, 16f
v. traditional biology, 191–193, 192f
- Systems Biology Markup Language (SBML), 4, 15, 19, 20, 27, 28, 30, 84, 103, 109, 110, 115, 122, 123
development tools of, 110–111
extensibility of, 111
practical considerations for, 111–112
usage of, 112
- Systems Biology Object Model (SysBio-OM), 26
- Systems biology standards
alternative, 112–113
future considerations for, 112
- Systems Biology Workbench (SBW), 4, 113, 116–117, 118f, 123
messaging protocols, 116, 117–118
- Systems of Life-Systems Biology, 30–31
- T**
- Ten-variable deterministic model, for circadian oscillations, 255–262, 257f, 258f, 259f
- Teranode, 120
- Termination stage, of signaling pathways, 159f
- Text clustering data flow, text summarization and, 46f
- Text clustering/summarization, in biomedical literature, 40–41
- Text mining enrichment, for microarray data analysis, 52

Text summarization, 48–49
text clustering data flow and, 46f
TextQuest, 40
TFs. *See* Transcription factors
TGF receptors. *See* Transforming growth factor beta receptors
3D cell culture models, 74–76, 76f, 77f
3D kinase phosphatase model, 349–352, 350f
Threshold response, 6
TIM degradation, 260
TIM phase delay, 260
Time-discrete dynamic system, 7
Tissue homeostasis, 294
Tissomics, 2, 371
TNF receptor-associated factor (TRAF), 160
TNF. *See* Tumor necrosis factor
Top-down modeling, 5, 7
abstract of, 229
of biochemical networks, 229–244
v. bottom-up modeling, 231–232
definition of, 230–232
introduction to, 229–230
Topology data, 315
TRAF. *See* TNF receptor-associated factor
TRANSCompel database, 17
Transcription factors (TFs), 212
TRANSFAC database, 17, 21
Transforming growth factor beta (TGF β) receptors, 155, 158, 159f
TRANSPATH database, 17, 21
Traveling salesman problem, 10
Tumor necrosis factor (TNF), 159f, 162
Turing machines, 10
2D cell culture models, 73–74, 75f
Tyramide signal amplification, 366
Tyrosine kinase activity, 154

U
UAB Proteomics Database, 18
Ubiquitin, 157
Ubiquitin-activating enzyme (E1), 157
Ubiquitin-conjugating enzyme (E2), 157
Ubiquitin ligase (E3), 157
UMLS. *See* Unified Medical Language System

Unified Medical Language System (UMLS), 41
Universal modeling language (UML), 24f, 25
object integration schema of, 29f
for project management, 30, 31f
Universal Protein Resource, 17
Unpredictability, 7
US DOE science grid, 10

V

Validation, model fitting and, 121–122
Valis, 82–83
bioinformatics environment of, 4
language-independent architecture and, 82
Simpathica and, 81–87
whole genome analysis and, 82
VCell, 115, 116, 120t
Vector field partitioning, in segmentation process, 68–69
Version control, 30–31
Voting -based techniques, of integrated imaging informatics, 69–73
Voting algorithm, 71f

W

Whole genome analysis, valis and, 82
WinSCAMP, 120t
Wnt pathway, 95f
axin response and, 100f
beta-catenin degradation and, 99f
beta-catenin response and, 99f
reactants of, 96f, 97f
steady-state analysis of, 98f
Wnt signaling example, 94–98, 95
Wnt subset, simulation of, 93f
WormBase, 18

X

Xenbase, 18
XSSYS, 85, 86, 87, 88, 92
XSSys query pane, 94f

Y

YADS, 115
Yeast gene functional families, 53t