METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For other titles published in this series, go to
www.springer.com/series/7651

METHODS IN MOLECULAR BIOLOGY™

Computational Systems Biology

Edited by

Jason McDermott

Pacific Northwest National Laboratory, Richland, WA, USA

Ram Samudrala

University of Washington, Seattle, WA, USA

Roger E. Bumgarner

University of Washington, Seattle, WA, USA

Kristina Montgomery

University of Washington, Seattle, WA, USA

Reneé Ireton

Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Humana Press

Editors
Jason McDermott
Pacific Northwest National Laboratory
Computational Biology & Bioinformatics Group
Richland, WA, USA
jason.mcdermott@pnl.gov

Ram Samudrala
Department of Microbiology
University of Washington
960 Republican St., WA, USA
ram@compbio.washington.edu

Roger E. Bumgarner
Department of Microbiology
University of Washington
960 Republican St., WA, USA
rogerb@u.washington.edu

Kristina Montgomery
Department of Microbiology
University of Washington
960 Republican St., WA, USA
kmontgom@u.washington.edu

Reneé Ireton
Fred Hutchinson Cancer Research Center
Public Health Sciences Div.
Molecular Diagnostics Program
WA, USA
rireton@fhcrc.org

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-58829-905-5 e-ISBN 978-1-59745-243-4
DOI 10.1007/978-1-59745-243-4

Library of Congress Control Number: 2009920380

Cover illustration: Figure 1 in Chapter1.

Printed on acid-free paper

springer.com

Preface

Computational systems biology is the term that we use to describe computational methods to identify, infer, model, and store relationships between the molecules, pathways, and cells ("systems") involved in a living organism. Based on this definition, the field of computational systems biology has been in existence for some time. However, the recent confluence of high-throughput methodology for biological data gathering, genome-scale sequencing, and computational processing power has driven a reinvention and expansion of this field. The expansions include not only modeling of small metabolic *(1–3)* and signaling systems *(2, 4)* but also modeling of the relationships between biological components in very large systems, including whole cells and organisms *(5–15)*. Generally, these models provide a general overview of one or more aspects of these systems and leave the determination of details to experimentalists focused on smaller subsystems. The promise of such approaches is that they will elucidate patterns, relationships, and general features, which are not evident from examining specific components or subsystems. These predictions are either interesting in and of themselves (e.g., the identification of an evolutionary pattern) or interesting and valuable to researchers working on a particular problem (e.g., highlight a previously unknown functional pathway).

Two events have occurred to bring the field of computational systems biology to the forefront. One is the advent of high-throughput methods that have generated large amounts of information about particular systems in the form of genetic studies, gene and protein expression analyses and metabolomics. With such tools, research to consider systems as a whole are being conceived, planned, and implemented experimentally on an ever more frequent and wider scale. The other event is the growth of computational processing power and tools. Methods to analyze large data sets of this kind are often computationally demanding and, as is the case in other areas, the field has benefited from continuing improvements in computational hardware and methods.

The field of computational biology is very much like a telescope with two sequential lenses: one lens represents the biological data and the other represents a computational and/or mathematical model of the data. Both lenses must be properly coordinated to yield an image that reflects biological reality. This means that the design parameters for both lenses must be designed in concert to create a system that yields a model of the organism, which provides both predictive and mechanistic information. The chapters in this book describe the construction of subcomponents of such a system. Computational systems biology is a rapidly evolving field and no single group of investigators has yet developed a complete system that integrates both data generation and data analysis in such a way so as to allow full and accurate modeling of any single biological organism. However, the field is rapidly moving in that direction. The chapters in this book represent a snapshot of the current methods being developed and used in the area of computational systems biology. Each method or database described within represents one or more steps on the path to a complete description of a biological system. How

these tools will evolve and ultimately be integrated is an area of intense research and interest. We hope that readers of this book will be motivated by the chapters within and become involved in this exciting area of research.

Organization of the Book

This volume is organized into five major parts: Network Components, Network Inference, Network Dynamics, Function and Evolutionary Systems Biology, and Computational Infrastructure for Systems Biology. Each section is described briefly below.

Part I – Network Components

This section focuses on methods to identify subcomponents of the complete networks. Ultimately, such subcomponents will need to be integrated with each other or used to inform other methods to arrive at a complete description of a biological system. This section begins with two methods for the prediction of transcription factor binding sites. In the first, Chapter 1, Mariño-Ramirez et al. describe a method for the prediction of transcription factor binding sites using a Gibbs sampling approach. In Chapter 2, Liu and Bader show how DNA-binding sites and specificity can be predicted using sophisticated structural analysis. Chapters 3–5 discuss methods to predict protein–protein interaction (PPI) networks, and Chapter 6 builds on predicted PPIs to identify potential regulatory interactions. Finally, Chapter 7 discusses the inherent modularity that is observed in biological networks with a focus on networks of PPIs.

Part II – Network Inference

This section focuses on methodologies to infer transcriptional networks on a genome-wide scale. In general, the methods described within focus on using either mRNA expression data or mRNA expression data coupled with expression quantitative trait locus (eQTL) data. To a large extent, method development in this area is driven primarily by the ubiquitous mRNA expression data that are available in the public domain or that are relatively easily generated within a single laboratory. These methods have been tremendously enabled by the development of array technology and hence predominately model mRNA levels (as that is the most ubiquitous data type). Chapters 8 and 9 present two methods for identifying and modeling transcriptional regulatory networks, while Chapter 10 focuses on inferring mRNA expression networks from eQTL data. Chapter 11 is a review of different methods for inferring and modeling large scale networks from expression and eQTL data.

Part III – Network Dynamics

Systems are not static entities. They change over time and in response to a variety of perturbations. Ultimately, computational systems biology will have to develop methods and corresponding data sets that allow one to infer and model the kinetics and

dynamics of reactions between all the chemical moieties in a cell. The chapters in this section focus on such methods. Chapter 12 discusses methods to infer both static co-expression networks and a finite-state Markov chain model for mimicking the dynamic behavior of a transcriptional network. Chapter 13 focuses on quantitative models of system behavior based on differential equations using biochemical control theory, whereas Chapter 14 focuses on the use of stochastic kinetic simulations. Both approaches have applications where one is superior to the other. At this point in time, it is not clear which methods will turn out to be most useful in dynamically modeling the largest number of biological systems. In general, this is likely the case for most of the technologies described in this book, so it is useful for readers to familiarize themselves with several concepts. Specifically, both Chapters 13 and 14 provide an excellent discussion of a variety of historical approaches to the dynamical modeling of biological systems and the relative merits and downsides to each. Chapter 15 provides an excellent introduction to considerations for the interplay between experimental design and dynamic modeling using lambda phage as an example system. The methods and considerations described within are generally applicable to other biological systems and highlight the importance of integrating the direction of wet bench work and computational modeling to more rapidly refine the models.

Part IV – Function and Evolutionary Systems Biology

The ultimate representation of the function of a given biological moiety is a complete description of all the reactions in which it participates and the relative rates of said reactions. At present, we are quite distant from this goal for most biological molecules or systems. However, we are able to use computational methods to predict the most likely functions of a given protein and even predict which portions and specific sequences of the protein contribute most to that function. This section is focused on methods used to infer protein function and on the relationships between function and evolution.

Ultimately, the reason to study and research "systems" biology is to understand biological function at a given hierarchical level (be it a single catalytic site or entire pathways). The interplay between the detailed atomic study of function and the large-scale study of systems will enable us to achieve this goal. This section contains chapters that address the interdependence of these two aspects: individual algorithms or techniques to understand the functional role of atoms or residues in single molecules (e.g., proteins), which in turn are extrapolated to understand their greater role in terms of biological or organismal function. Conversely and complementarily, the role of larger systems and their influence on single molecules is also explored. Together, all these chapters illustrate the strong dependence between single molecules and entire pathways or systems.

Part V – Computational Infrastructure for Systems Biology

To represent and organize the large amounts of experimental data and software tools, database frameworks must be created and made available to the larger biological

community. This chapter focuses on computational methods and databases as well as data representations necessary to both integrate and export systems biology information to an end user. The user may be the biologist searching for their gene of interest or they may be the bioinformatician looking for trends in protein function among higher eukaryotes. Several groups are working on this extremely difficult task of providing semantic meaning to the large amounts of underlying biological data collected from single and high-throughput experiments, as well as computational predictions. (As a parenthetical comment, this is a significantly much harder problem than one faced by Internet search engines such as a Google, which at this point do not provide any semantic meaning to a query.) We present only a few such examples in this section (and in this book). One primary focus is on the Bioverse framework, database, and web application, which was developed by the editors of this book. However, we also describe the Biozon as well as the SEBIN and CABIN frameworks. The abstract representations required to model biological systems are still in fruition, and a complement of many tools, technologies, databases, and algorithms will have to be integrated in the future as our knowledge expands.

Acknowledgments

It goes without saying that this volume would not have been possible without the efforts of countless biologists who provided the raw data that enable modeling of biological systems. We first and foremost thank all these researchers who provide the raw data for all the computational modeling that enables the field of "computational" systems biology. We also thank all the researchers who investigate these problems using sophisticated modeling techniques, particularly those described in this volume. We thank two particular editors of this volume, Reneé Ireton and Kristina Montgomery, who have dealt with the thankless job of undertaking the proofreading the chapters in this book and associated bureaucratic requirements. We finally thank the McDermott, Bumgarner, and Samudrala groups for their critical comments as this volume was being prepared, as well as our respective families for their patience. The administrative aspects of this work were in part supported by the NSF CAREER award to Dr. Ram Samudrala.

References

1. Ishii N, Robert, M, Nakayama, Y, Kanai, A and Tomita, M. Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 2004;113(1-3):281–94.
2. Ekins S, Nikolsky, Y, Bugrim, A, Kirillov, E and Nikolskaya, T. Pathway mapping tools for analysis of high content data. Methods Mol Biol 2006;356:319–50.
3. Lafaye A, Junot, C, Pereira, Y, et al. Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. J Biol Chem 2005;280(26):24723–30.
4. Stevenson-Paulik J, Chiou, ST, Frederick, JP, et al. Inositol phosphate metabolomics: merging genetic perturbation with modernized radiolabeling methods. Methods 2006;39(2):112–21.
5. Ideker T, Thorsson, V, Ranish, JA, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001;292(5518):929–34.
6. Pe'er D, Regev, A, Elidan, G and Friedman, N. Inferring subnetworks from perturbed expression profiles. Bioinformatics 2001;17 Suppl 1:S215–24.

7. Pilpel Y, Sudarsanam, P and Church, GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001;29(2):153–9.
8. Ideker T, Ozier, O, Schwikowski, B and Siegel, AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002;18 Suppl 1:S233–40.
9. Kelley BP, Sharan, R, Karp, RM, et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 2003;100(20):11394–9.
10. Shannon P, Markiel, A, Ozier, O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.
11. Ideker T. A systems approach to discovering signaling and regulatory pathways–or, how to digest large interaction networks into relevant pieces. Adv Exp Med Biol 2004;547:21–30.
12. Schadt EE, Monks, SA, Drake, TA, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 2003;422(6929):297–302.
13. Schadt EE and Lum, PY. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. J Lipid Res 2006.
14. McDermott J and Samudrala R. Bioverse: Functional, structural and contextual annotation of proteins and proteomes. Nucleic Acids Res 2003;31(13):3736–7.
15. McDermott J, Bumgarner, R and Samudrala, R. Functional annotation from predicted protein interaction networks. Bioinformatics 2005;21(15):3217–26.

Contents

Contributors

L. Aravind • *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

M. Madan Babu • *MRC Laboratory of Molecular Biology, Cambridge, UK*

Joel S. Bader • *Department of Biomedical Engineering and High-Throughput Biology Center, John Hopkins University, Baltimore, MD, USA*

Olivier Bodenreider • *National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

Richard Bonneau • *Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*

Roger E. Bumgarner • *Department of Microbiology, University of Washington, Seattle, WA, USA*

Liran Carmel • *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

Danny Z. Chen • *Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA*

Jake Y. Chen • *Informatics and Technology Complex (IT), Indiana University School of Informatics, IUPUI, Indianapolis, IN, USA*

Jason V. Davis • *Department of Computer Science, University of Texas at Austin, Austin, TX, USA*

William Dirks • *Center for Integrative Genomics, University of California, Berkeley, Berkeley, CA, USA*

Eric Franosa • *Bioinformatics Program, Boston University, Boston, MA, USA*

Zach Frazier • *Department of Microbiology, University of Washington, Seattle, WA, USA*

Michal Guerquin • *Department of Microbiology, University of Washington, Seattle, WA, USA*

Chengbang Huang • *Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA*

Jesus Izaguirre • *Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA*

Simon P. Kanaan • *Department of Computer Science, University of Notre Dame, Notre Dame, IN, USA*

Grigory Kolesov • *Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA*

Eugene V. Koonin • *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

David Landsman • *Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA*

Benjamin Lang • *MRC Laboratory of Molecular Biology, Cambridge, UK*

Insuk Lee • *Center of Systems and Synthetic Biology, Institute for Molecular Biology, University of Texas at Austin, Austin, TX, USA*
Huai Li • *Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA*
Bolan Linghu • *Bioinformatics Program, Boston University, Boston, MA, USA*
L. Angela Liu • *Department of Biomedical Engineering and Institute for Multiscale Modeling of Biological Interactions, John Hopkins University, Baltimore, MD, USA*
Aviv Madar • *Center for Comparative Functional Genomics, New York University, New York, NY, USA*
Edward M. Marcotte • *Center for Systems and Synthetic Biology and Department of Chemistry and Biochemistry, Institute for Molecular Biology, University of Texas at Austin, Austin, TX, USA*
Leonardo Mariño-Ramirez • *Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA*
Jason McDermott • *Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA, USA*
Leonid A. Mirny • *Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA*
Sean D. Mooney • *Department of Medical and Molecular Genetics, Center for Computational Biology and Bioinformatics, IUPUI, Indianapolis, IN, USA*
Adam Christopher Palmer • *School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia*
Michel F. Pettigrew • *ScienceOps, Bothell, WA, USA*
Linda Petzold • *Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA, USA*
Shafquat Rahman • *Mathworks Inc., Natick, MA, USA*
Imran Rashid • *Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA*
Erzsébet Ravasz Regan • *Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA*
Haluk Resat • *Pacific Northwest National Laboratory, Richland, WA, USA*
Igor B. Rogozin • *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
Ram Samudrala • *Department of Microbiology, University of Washington, Seattle, WA, USA*
Herbert M. Sauro • *Department of Bioengineering, University of Washington, Seattle, WA, USA*
Itai Sharon • *Department of Computer Science, Technion - Israel Institute of Technology, Haifa, Israel*
Keith Edward Shearwin • *School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia*
Mudita Singhal • *Computational Biology and Bioinformatics Group, Computational and Informational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA*
John Spouge • *Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA*

Yu Sun • *Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA*
Umar Syed • *Department of Computer Science, Princeton University, Princeton, NJ, USA*
Ronald Taylor • *Computational Biology and Bioinformatics Group, Computational and Informational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA*
Kannan Tharakaraman • *Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA*
Duangdao Wichadakul • *BIOTEC, Pathumthani, Thailand*
Yuri I. Wolf • *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
Stefan Wuchty • *Northwestern Institute of Complexity, Northwestern University, Evanston, IL, USA*
Yu Xia • *Bioinformatics Program and Department of Chemistry, Boston University, Boston, MA, USA*
Ka Yee Yeung • *Department of Microbiology, University of Washington, Seattle, WA, USA*
Golan Yona • *Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA; Department of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel*
Eunseog Youn • *Department of Computer Science, Texas Tech University, Lubbock, TX, USA*
Ming Zhan • *Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA*

Color Plates

Color Plate 1 Uncovering the underlying modularity of a complex network. **(a)** Topological overlap illustrated on a small hypothetical network. On each link, we indicate the topological overlap for the connected nodes; and in parentheses next to each node, we indicate the node's clustering coefficient. **(b)** The topological overlap matrix corresponding to the small network shown in (**a**). The *rows and columns* of the matrix were reordered by the application of an average linkage clustering method to its elements, allowing us to identify and place close to each other those nodes that have high topological overlap. The *color code* denotes the degree of topological overlap between the nodes. The associated tree reflects the three distinct modules built into the model, as well as the fact that the EFG and HIJK modules are closer to each other in the topological sense than to the ABC module (Chapter 7, Fig. 3; *see* discussion on p. 151).

Color Plate 2 Topological modules in the *Escherichia coli* metabolism: the topologic overlap matrix, together with the corresponding hierarchical tree (*top and right*) that quantifies the relation between the different modules. The branches of the tree are *color-coded* to reflect the predominant biochemical classification of their substrates. The *color code* of the matrix denotes the degree of topological overlap shown in the matrix. The large-scale functional map of the metabolism, as suggested by the hierarchical tree, is also shown (*bottom*) (Chapter 7, Fig. 5; *see* discussion on p. 153).

Color Plate 3 Enlarged view of the substrate module of pyrimidine metabolism, along with a detailed diagram of the metabolic reactions that surround and incorporate it. The *colored boxes* in the background denote the first two levels of the three levels of nested modularity suggested by the hierarchical tree. *Red-outlined boxes* denote the substrates directly appearing in the reduced metabolism and thus on the tree (Chapter 7, Fig. 6; *see* discussion on p. 154 and full caption on p. 155).

Color Plate 4 Structural localization of putative SDRs and CERs in two-component system domains. (**a**) RR Spo0F (*red-brown ribbon*) bound to structural analog of the DD in Spo0B protein. The conserved His is shown in *purple*, the conserved Asp in RR in *magenta*. SDRs and CERs are shown in *yellow* or, when located on the α4 helix, in *white* (PDB entry *1F51*). (b) The non-catalytic conformation of HK homodimer. ADP is shown as a *purple* wireframe, the phosphate-accepting conserved His residue in *magenta spacefill*. SDRs and CERs on the ATPase are shown in *yellow*, or in *white* if located on the unresolved ATP-lid loop that was

superimposed from PhoQ kinase (PDB entry *1ID0*), or in *green* in the RR-specific CERs side patch. SDRs and CERs on the DD are shown in *red* on one homodimer and *orange* on another (PDB entry *2C2A*) (Chapter 18, Fig. 6; *see* discussion on p. 435).

Color Plate 5 Localization of putative SDRs and CERs on computationally obtained models (models provided by Marina et al *(27)*). (a) HK in the active conformation, the ATPase is docked on the DD so that transfer of the phosphoryl group is possible. SDRs and CERs on the ATPase domain are shown in *yellow* or *green* when located in the RR-specific CERs side patch. SDRs and CERs on the DD are shown in *red* on one homodimer and orange on another. (b) Spo0F computationally docked on HK and subsequently superimposed with RR from OmpR. RR (*brown-red ribbon*) (PDB entry *1KGS*) with its 4 helix swung ~90°: the phosphorylated Asp in the RR is shown in *magenta*, SDR and CERs are shown in *light red* or, when located the 4 helix in *white*. DD (*dark blue* and *dark green ribbon*): SDRs and CERs are shown in *light blue* on one dimer and in *light green* on another. ATPase (*yellow-green ribbon* on the *left* and *light-blue* on the *right*): the colors are the same as in (a) (Chapter 18, Fig. 7; *see* discussion on p. 439).

Color Plate 6 Valine aminoacyl-tRNA synthetase (PDB entry *1GAX*). The tRNA is shown as a *purple* wireframe structure, SDRs and CERs are *red* balls, and amino acid (valyl-adenylate analog) is in *yellow* wireframe (Chapter 18, Fig. 8; *see* discussion on p. 443).