

# Systems Biology

*Volume 1:*  
*Genomics*

Edited by  
Isidore Rigoutsos  
&  
Gregory Stephanopoulos

**OXFORD**  
UNIVERSITY PRESS  
2007

## Series in Systems Biology

Edited by Dennis Shasha, New York University

### EDITORIAL BOARD

Michael Ashburner, University of Cambridge  
Amos Bairoch, Swiss Institute of Bioinformatics  
Charles Cantor, Sequenom, Inc.  
Leroy Hood, Institute for Systems Biology  
Minoru Kanehisa, Kyoto University  
Raju Kucheralapati, Harvard Medical School

Systems Biology describes the discipline that seeks to understand biological phenomena on a large scale: the association of gene with function, the detailed modeling of the interaction among proteins and metabolites, and the function of cells. Systems Biology has wide-ranging application, as it is informed by several underlying disciplines, including biology, computer science, mathematics, physics, chemistry, and the social sciences. The goal of the series is to help practitioners and researchers understand the ideas and technologies underlying Systems Biology. The series volumes will combine biological insight with principles and methods of computational data analysis.

*Cellular Computing*, edited by Martyn Amos

*Systems Biology, Volume I: Genomics*, edited by Isidore Rigoutsos and Gregory Stephanopoulos

*Systems Biology, Volume II: Networks, Models, and Applications*, edited by Isidore Rigoutsos and Gregory Stephanopoulos

# Contents

*Contributors* xi

*Systems Biology: A Perspective* xiii

- 1** Prebiotic Chemistry on the Primitive Earth 3  
*Stanley L. Miller & H. James Cleaves*
  - 2** Prebiotic Evolution and the Origin of Life:  
Is a System-Level Understanding Feasible? 57  
*Antonio Lazcano*
  - 3** Shotgun Fragment Assembly 79  
*Granger Sutton & Ian Dew*
  - 4** Gene Finding 118  
*John Besemer & Mark Borodovsky*
  - 5** Local Sequence Similarities 154  
*Temple F. Smith*
  - 6** Complete Prokaryotic Genomes:  
Reading and Comprehension 166  
*Michael Y. Galperin & Eugene V. Koonin*
  - 7** Protein Structure Prediction 187  
*Jeffrey Skolnick & Yang Zhang*
  - 8** DNA–Protein Interactions 219  
*Gary D. Stormo*
  - 9** Some Computational Problems Associated  
with Horizontal Gene Transfer 248  
*Michael Syvanen*
  - 10** Noncoding RNA and RNA Regulatory Networks  
in the Systems Biology of Animals 269  
*John S. Mattick*
- Index 303

# Contributors

JOHN BESEMER

Department of Biology  
Georgia Institute of Technology  
Atlanta, Georgia  
john@amber.biology.gatech.edu

MARK BORODOVSKY

Department of Biology  
Georgia Institute of Technology  
Atlanta, Georgia  
mark.borodovsky@biology.gatech.edu

H. JAMES CLEAVES

The Scripps Institution of  
Oceanography  
University of California, San Diego  
La Jolla, California  
hcleaves@ucsd.edu

IAN DEW

Steck Consulting, LLC  
Washington, DC  
ian@catmandew.com

MICHAEL Y. GALPERIN

National Center for Biotechnology  
Information  
National Institutes of Health  
Bethesda, Maryland  
galperin@ncbi.nlm.nih.gov

EUGENE V. KOONIN

National Center for Biotechnology  
Information  
National Institutes of Health  
Bethesda, Maryland  
koonin@ncbi.nlm.nih.gov

ANTONIO LAZCANO

Faculty of Science  
Universidad Nacional  
Autónoma de México  
Mexico City, Mexico  
alar@correo.unam.mx

JOHN S. MATTICK

Institute for Molecular Bioscience  
University of Queensland  
Brisbane, Australia  
j.mattick@imb.uq.edu.au

STANLEY L. MILLER

Scripps Institution of  
Oceanography  
University of California,  
San Diego  
La Jolla, California  
smiller@ucsd.edu

JEFFREY SKOLNICK

New York State Center of  
Excellence in Bioinformatics  
and Life Sciences  
University at Buffalo  
The State University of  
New York  
Buffalo, New York  
skolnick@buffalo.edu

TEMPLE F. SMITH

BioMolecular Engineering  
Resource Center  
Boston University  
Boston, Massachusetts  
tsmith@darwin.bu.edu

GARY D. STORMO

Department of Genetics  
Washington University in St. Louis  
St. Louis, Missouri  
stormo@genetics.wustl.edu

GRANGER SUTTON

J. Craig Venter Institute  
Rockville, Maryland  
GSutton@venterininstitute.org

MICHAEL SYVANEN

Department of Medical  
Microbiology and Immunology  
University of California Davis School  
of Medicine  
Sacramento, California  
msyvanen@ucdavis.edu

YANG ZHANG

Center for Bioinformatics  
University of Kansas  
Lawrence, Kansas  
yzhang@ku.edu

# Systems Biology: A Perspective

As recently as a decade ago, the core paradigm of biological research followed an established path: beginning with the generation of a specific hypothesis a concise experiment would be designed that typically focused on studying a small number of genes. Such experiments generally measured a few macromolecules, and, perhaps, small metabolites of the target system.

The advent of genome sequencing and associated technologies greatly improved scientists' ability to measure important classes of biological molecules and their interactions. This, in turn, expanded our view of cells with a bevy of previously unavailable data and made possible genome-wide and cell-wide analyses. These newly found lenses revealed that hundreds (sometimes thousands) of molecules and interactions, which were outside the focus of the original study, varied significantly in the course of the experiment.

The term *systems biology* was coined to describe the field of scientific inquiry which takes a global approach to the understanding of cells and the elucidation of biological processes and mechanisms. In many respects, this is also what physiology (from the Greek *physis* = nature and *logos* = word-knowledge) focused on for the most part of the twentieth century. Indeed, physiology's goal has been the study of function and characteristics of living organisms and their parts and of the underlying physiochemical phenomena. Unlike physiology, systems biology attempts to interpret and contextualize the large and diverse sets of biological measurements that have become visible through our genomic-scale window on cellular processes by taking a holistic approach and bringing to bear theoretical, computational, and experimental advances in several fields. Indeed, there is considerable excitement that, through this integrative perspective, systems biology will succeed in elucidating the mechanisms that underlie complex phenomena and which would have otherwise remained undiscovered.

For the purposes of our discussion, we will be making use of the following definition: "Systems biology is an integrated approach that brings together and leverages theoretical, experimental, and computational approaches in order to establish connections among important molecules or groups of molecules in order to aid the eventual mechanistic explanation of cellular processes and systems." More specifically, we view systems biology as a field that aims to uncover concrete molecular relationships for targeted analysis through the interpretation

of cellular phenotype in terms of integrated biomolecular networks. The fidelity and breadth of our network and state characterization are intimately related to the degree of our understanding of the system under study. As the readers will find, this view permeates the treatises that are found in these two books.

Cells have always been viewed as elegant systems of immense complexity that are, nevertheless, well coordinated and optimized for a particular purpose. This apparent complexity led scientists to take a reductionist approach to research which, in turn, contributed to a rigorous understanding of low-level processes in a piecemeal fashion. Nowadays, completed genomic sequences and systems-level probing hold the potential to accelerate the discovery of unknown molecular mechanisms and to organize the existing knowledge in a broader context of high-level cellular understanding. Arguably, this is a formidable task. In order to improve the chances of success, we believe that one must anchor systems biology analyses to specific questions and build upon the existing core infrastructure that the earlier, targeted research studies have allowed us to generate.

The diversity of molecules and reactions participating in the various cellular functions can be viewed as an impediment to the pursuit of a more complete understanding of cellular function. However, it actually represents a great opportunity as it provides countless possibilities for modifying the cellular machinery and commandeering it toward a specific goal. In this context, we distinguish two broad categories of questions that can guide the direction of systems biology research. The first category encompasses topics of medical importance and is typically characterized by forward-engineering approaches that focus on preventing or combating disease. The second category includes problems of industrial interest, such as the genetic engineering of microbes so as to maximize product formation, the creation of robust-production strains, and so on. The applications of the second category comprise an important reverse-engineering component whereby microbes with attractive properties are scrutinized for the purpose of transferring any insights learned from their functions to the further improvement and optimization of production strains.

## PRIOR WORK

As already mentioned, and although the term *systems biology* did not enter the popular lexicon until recently, some of the activities it encompasses have been practiced for several decades. As we cannot possibly be exhaustive, we present a few illustrative examples of approaches that have been developed in recent years and successfully applied to relatively small systems. These examples can serve as useful guides in our attempt to tackle increasingly larger challenges.

## Metabolic Control Analysis (MCA)

Metabolic pathways and, in general, networks of reactions are characterized by substantial stoichiometric and (mostly) kinetic complexity in their own right. The commonly applied assumption of a single rate-limiting step leads to great simplification of the reaction network and often yields analytical expressions for the conversion rates. However, this assumption is not justified for most biological systems where kinetic control is not concentrated in a single step but rather is distributed among several enzymatic steps. Consequently, kinetics and flux control of a bioreaction network represent properties of the entire system and can be determined from the characteristics of individual reactions in a bottom-up approach or from the response of the overall system in a top-down approach. The concepts of MCA and distribution of kinetic control in a reaction pathway have had a profound impact on the identification of target enzymes whose genetic modification permitted the amplification of the product flux through a pathway.

## Signaling Pathways

Signal transduction is the process by which cells communicate with each other and their environment and involves a multitude of proteins that can be in active or inactive states. In their active (phosphorylated) state they act as catalysts for the activation of subsequent steps in the signaling cascade. The end result is the activation of a transcription factor which, in turn, initiates a gene transcription event. Until recently, and even though several of the known proteins participate in more than one signaling cascade, such systems were being studied in isolation from one another. A natural outcome of this approach was of course the ability to link a single gene with a single ligand in a causal relationship whereby the ligand activates the gene. However, such findings are not representative in light of the fact that signaling pathways branch and interact with one another creating a rather intricate and complex signaling network. Consequently, more tools, computational as well as experimental, are required if we are to improve our understanding of signal transduction. Developing such tools is among the goals of the recently formed Alliance for Cellular Signaling, an NIH-funded project involving several laboratories and research centers ([www.signaling-gateway.org](http://www.signaling-gateway.org)).

## Reconstruction of Flux Maps

Metabolic pathway fluxes are defined as the actual rates of metabolite interconversion in a metabolic network and represent most informative measures of the actual physiological state of cells and organisms. Their dependence on enzymatic activities and metabolite concentrations makes them an accurate representation of carbon and energy flows through the various pathway branches. Additionally, they are very

important in identifying critical reaction steps that impact flux control for the entire pathway. Thus, flux determination is an essential component of strain evaluation and metabolic engineering. Intracellular flux determination requires the enumeration and satisfaction of all intracellular metabolite balances along with the use of sufficient measurements typically derived from the introduction of isotopic tracers and metabolite and mass isotopomer measurement by gas chromatography–mass spectrometry. It is essentially a problem of constrained parameter estimation in overdetermined systems with overdetermination providing the requisite redundancy for reliable flux estimation. These approaches are basically methods of network reconstruction whereas the obtained fluxes represent properties of the entire system. As such, the fluxes accurately reflect changes introduced through genetic or environmental modifications and, thus, can be used to assess the impact of such modifications on cell physiology and product formation, and to guide the next round of cell modifications.

### **Metabolic Engineering**

Metabolic engineering is the field of study whose goal is the improvement of microbial strains with the help of modern genetic tools. The strains are modified by introducing specific transport, conversion, or deregulation changes that lead to flux redistribution and the improvement of product yield. Such modifications rely to a significant extent on modern methods from molecular biology. Consequently, the following central question arises: “What is the real difference between genetic engineering and metabolic engineering?” We submit that the main difference is that metabolic engineering is concerned with the entire metabolic system whereas genetic engineering specifically focuses on a particular gene or a small collection of genes. It should be noted that over- or underexpression of a single gene or a few genes may have little or no impact on the attempt to alter cell physiology. On the other hand, by examining the properties of the metabolic network as a whole, metabolic engineering attempts to identify targets for amplification as well as rationally assess the effect that such changes will incur on the properties of the overall network. As such, metabolic engineering can be viewed as a precursor to functional genomics and systems biology in the sense that it represents the first organized effort to reconstruct and modify pathways using genomic tools while being guided by the information of postgenomic developments.

### **WORDS OF CAUTION**

In light of the many exciting possibilities, there are high expectations for the field of systems biology. However, as we move forward, we should not lose sight of the fact that the field is trying to tackle a

problem of considerable magnitude. Consequently, any expectations of immediate returns on the scientific investment should be appropriately tempered. As we set out to forecast future developments in this field, it is important to keep in mind several points.

*Despite the wealth of available genomic data, there are still a lot of regions in the genomes of interest that are functional and which have not been identified as such.* In order to practice systems biology, lists of “parts” and “relationships” that are as complete as possible are needed. In the absence of such complete lists, one generally hopes to derive at best an approximate description of the actual system’s behavior. A prevalent misconception among scientists states that nearly complete lists of *parts* are already in place. Unfortunately, this is not the case—the currently available parts lists are incomplete as evidenced by the fact that genomic maps are continuously updated through the addition or removal of (occasionally substantial amounts of) genes, by the discovery of more regions that code for RNA genes, and so on.

*Despite the wealth of available genomic data, knowledge about existing optimal solutions to important problems continues to elude us.* The current efforts in systems biology are largely shaped by the available knowledge. Consequently, optimal solutions that are implemented by metabolic pathways that are unknown or not yet understood are beyond our reach. A characteristic case in point is the recent discovery, in sludge microbial communities, of a *Rhodocyclus*-like polyphosphate-accumulating organism that exhibits enhanced biological phosphorus removal abilities. Clearly, this microbe is a great candidate to be part of a biological treatment solution to the problem of phosphorus removal from wastewater. Alas, this is not yet an option as virtually nothing is known about the metabolic pathways that confer phosphorus removal ability to this organism.

*Despite the wealth of available genomic data, there are still a lot of important molecular interactions of whose existence we are unaware.* Continuing on our parts and relationships comment from above, it is worth noting another prevalent misconception among scientists: it states that nearly complete lists of *relationships* are already in place. For many years, pathway analysis and modeling has been characterized by protein-centric views that comprised concrete collections of proteins participating in well-understood interactions. Even for well-studied pathways, new important protein interactions are continuously discovered. Moreover, accumulating experimental evidence shows that numerous important interactions are in fact effected by the action of RNA molecules on DNA molecules and by extension on proteins. Arguably, posttranscriptional gene silencing and RNA interference represent one area of research activity with the potential to substantially revise our current

understanding of cellular processes. In fact, the already accumulated knowledge suggests that the traditional protein-centric views of the systems of interest are likely incomplete and need to be augmented appropriately. This in turn has direct consequences on the modeling and simulation efforts and on our understanding of the cell from an integrated perspective.

*Constructing biomolecular networks for new systems will require significant resources and expertise.* Biomolecular networks incorporate a multitude of relationships that involve numerous components. For example, constructing gene interaction maps requires large experimental investments and computational analysis. As for global protein–protein interaction maps, these exist for only a handful of model species. But even reconstructing well-studied and well-documented networks such as metabolic pathways in a genomic context can prove a daunting task. The magnitude of such activities has already grown beyond the capabilities of a single investigator or a single laboratory.

*Even when one works with a biomolecular network database, the system picture may be incomplete or only partially accurate.* In the postgenomic era, the effort to uncover the structure and function of genetic regulatory networks has led to the creation of many databases of biological knowledge. Each of these databases attempts to distill the most salient features from incomplete, and at times flawed, knowledge. As an example, several databases exist that enumerate protein interactions for the yeast genome and have been compiled using the yeast two-hybrid screen. These databases currently document in excess of 80,000 putative protein–protein interactions; however, the knowledge content of these databases has only a small overlap, suggesting a strong dependence of the results on the procedures used and great variability in the criteria that were applied before an interaction could be entered in the corresponding knowledge repository. As one might have expected, the situation is less encouraging for those organisms with lower levels of direct interaction experimentation and scrutiny (e.g., *Escherichia coli*) or which possess larger protein interaction spaces (e.g., mouse and human); in such cases, the available databases capture only a minuscule fraction of the knowledge spectrum.

*Carrying out the necessary measurements requires significant resources and expertise.* Presently, the only broadly available tool for measuring gene expression is the DNA chip (in its various incarnations). Conducting a large-scale transcriptional experiment will incur unavoidable significant costs and require that the involved scientists be trained appropriately. Going a step further, in order to measure protein levels, protein states, regulatory elements, and metabolites, one needs access to complex and specialized equipment. Practicing systems biology will

necessitate the creation of partnerships and the collaboration of faculty members across disciplines. Biologists, engineers, chemists, physicists, mathematicians, and computer scientists will need to learn to speak one another's language and to work together.

*It is unlikely that a single/complex microarray experiment will shed light on the interactions that a practitioner seeks to understand.* Even leaving aside the large amounts of available data and the presence of noise, many of the relevant interactions will simply not incur any large or direct transcriptional changes. And, of course, one should remain mindful of the fact that transcript levels do not necessarily correlate with protein levels, and that protein levels do not correlate well with activity level. The situation is accentuated further if one considers that both transcript and protein levels are under the control of agents such as microRNAs that were discovered only recently—the action of such agents may also vary temporally contributing to variations across repetitions of the same experiment.

*Patience, patience, and patience: the hypotheses that are derived from systems-based approaches are more complex than before and disproportionately harder to validate.* For a small system, it is possible to design experiments that will test a particular hypothesis. However, it is not obvious how this can be done when the system under consideration encompasses numerous molecular players. Typically, the experiments that have been designed to date strove to keep most parameters constant while studying the effect of a small number of changes introduced to the system in a controlled manner. This conventional approach will need to be reevaluated since now the number of involved parameters is dramatically higher and the demands on system controls may exceed the limits of present experimentation.

## **ABOUT THIS BOOK**

From the above, it should be clear that the systems biology field comprises multifaceted research work across several disciplines. It is also hierarchical in nature with single molecules at one end of the hierarchy and complete, interacting organisms at the other. At each level of the hierarchy, one can distinguish “parts” or active agents with concrete static characteristics and dynamic behavior. The active agents form “relationships” by interacting among themselves within each level, but can also be involved in inter-level interactions (e.g., a transcription factor, which is an agent associated with the proteomic level, interacts at specific sites with the DNA molecule, an agent associated with the genomic level of the hierarchy).

Clearly, intimate knowledge and understanding of the specifics at each level will greatly facilitate the undertaking of systems

biology activities. Experts are needed at all levels of the hierarchy who will continue to generate results with an eye toward the longer-term goal of the eventual mechanistic explanation of cellular processes and systems.

The two books that we have edited try to reflect the hierarchical nature of the problem as well as this need for experts. Each chapter is contributed by authors who have been active in the respective domains for many years and who have gone to great lengths to ensure that their presentations serve the following two purposes: first, they provide a very extensive overview of the domain's main activities by describing their own and their colleagues' research efforts; and second, they enumerate currently open questions that interested scientists should consider tackling. The chapters are organized into a "Genomics" and a "Networks, Models, and Applications" volume, and are presented in an order that corresponds roughly to a "bottom-up" traversal of the systems biology hierarchy.

The "Genomics" volume begins with a chapter on prebiotic chemistry on the primitive Earth. Written by Stanley Miller and James Cleaves, it explores and discusses several geochemically reasonable mechanisms that may have led to chemical self-organization and the origin of life. The second chapter is contributed by Antonio Lazcano and examines possible events that may have led to the appearance of encapsulated replicative systems, the evolution of the genetic code, and protein synthesis. In the third chapter, Granger Sutton and Ian Dew present and discuss algorithmic techniques for the problem of fragment assembly which, combined with the shotgun approach to DNA sequencing, allowed for significant advances in the field of genomics. John Besemer and Mark Borodovsky review, in chapter 4, all of the major approaches in the development of gene-finding algorithms. In the fifth chapter, Temple Smith, through a personal account, covers approximately twenty years of work in biological sequence alignment algorithms that culminated in the development of the Smith–Waterman algorithm. In chapter 6, Michael Galperin and Eugene Koonin discuss the state of the art in the field of functional annotation of complete genomes and review the challenges that proteins of unknown function pose for systems biology. The state of the art of protein structure prediction is discussed by Jeffrey Skolnick and Yang Zhang in chapter 7, with an emphasis on knowledge-based comparative modeling and threading approaches. In chapter 8, Gary Stormo presents and discusses experimental and computational approaches that allow the determination of the specificity of a transcription factor and the discovery of regulatory sites in DNA. Michael Syvanen presents and discusses the phenomenon of horizontal gene transfer in chapter 9 and also presents computational questions that relate to the phenomenon. The first volume concludes with a chapter

by John Mattick on non-protein-coding RNA and its involvement in regulatory networks that are responsible for the various developmental stages of multicellular organisms.

The “Networks, Models, and Applications” volume continues our ascent of the systems biology hierarchy. The first chapter, which is written by Cristian Ruse and John Yates III, introduces mass spectrometry and discusses its numerous uses as an analytical tool for the analysis of biological molecules. In chapter 2, Chris Floudas and Ho Ki Fung review mathematical modeling and optimization methods for the de novo design of peptides and proteins. Chapter 3, written by William Swope, Jed Pitera, and Robert Germain, describes molecular modeling and simulation techniques and their use in modeling and studying biological systems. In chapter 4, Glen Held, Gustavo Stolovitzky, and Yuhai Tu discuss methods that can be used to estimate the statistical significance of changes in the expression levels that are measured with the help of global expression assays. The state of the art in high-throughput technologies for interrogating cellular signaling networks is discussed in chapter 5 by Jason Papin, Erwin Gianchandani, and Shankar Subramaniam, who also examine schemes by which one can generate genotype–phenotype relationships given the available data. In chapter 6, Dimitrios Mastellos and John Lambris use the complement system as a platform to describe systems approaches that can help elucidate gene regulatory networks and innate immune pathway associations, and eventually develop effective therapeutics. Chapter 7, written by Sang Yup Lee, Dong-Yup Lee, Tae Yong Kim, Byung Hun Kim, and Sang Jun Lee, discusses how computational and “-omics” approaches can be combined in order to appropriately engineer “improved” versions of microbes for industrial applications. In chapter 8, Markus Herrgård and Bernhard Palsson discuss the design of metabolic and regulatory network models for complete genomes and their use in exploring the operational principles of biochemical networks. Raimond Winslow, Joseph Greenstein, and Patrick Helm review and discuss the current state of the art in the integrative modeling of the cardiovascular system in chapter 9. The volume concludes with a chapter on embryonic stem cells and their uses in testing and validating systems biology approaches, written by Andrew Thomson, Paul Robson, Huck Hui Ng, Hasan Otu, and Bing Lim.

The companion website for *Systems Biology* Volumes I and II provides color versions of several figures reproduced in black and white in print. Please refer to <http://www.oup.com/us/sysbio> to view these figures in color:

Volume I: Figures 7.5 and 7.6

Volume II: Figures 3.10, 5.1, 7.4 and 9.8