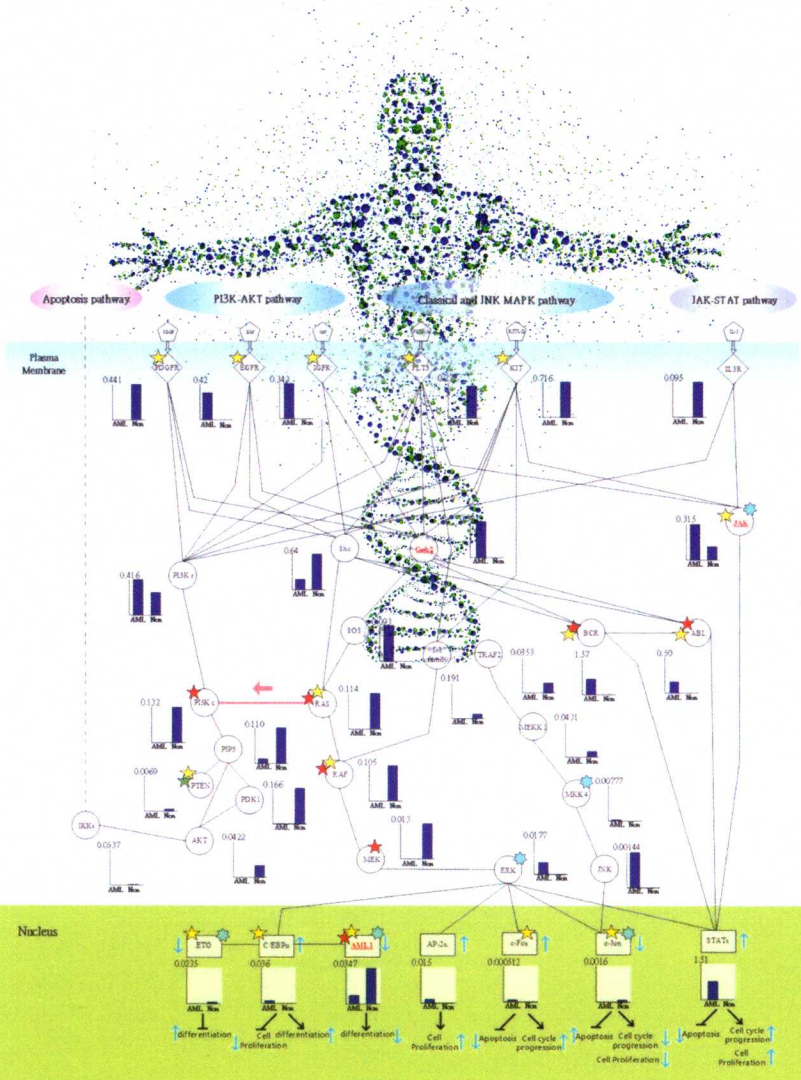
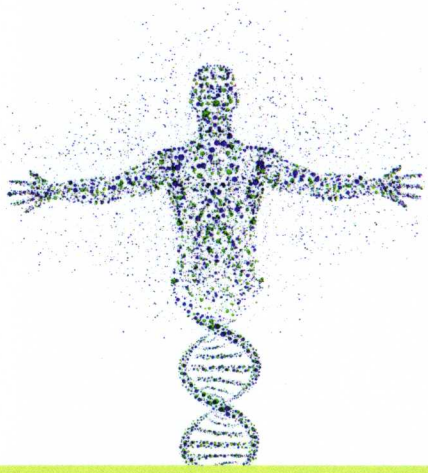
BIG MECHANISMS IN SYSTEMS BIOLOGY

Big Data Mining, Network Modeling, and Genome-Wide Data Identification





BOR-SEN CHEN & CHENG-WEI LI

BIG MECHANISMS IN SYSTEMS BIOLOGY

Big Data Mining, Network Modeling, and Genome-Wide Data Identification

BOR-SEN CHEN
National Tsing Hua University, Hsinchu, Taiwan

CHENG-WEI LI
National Tsing Hua University, Hsinchu, Taiwan

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1800, San Diego, CA 92101-4495, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-809479-2

Publisher: Mica Haley
Acquisition Editor: Rafael Teixeira
Editorial Project Manager: Mariana Kuhl
Production Project Manager: Karen East and Kirsty Halterman
Designer: Victoria Pearson

Typeset by MPS Limited, Chennai, India

CONTENTS

CHAPTER 1

Introduction to Big Mechanisms in Systems Biology

INTRODUCTION

Currently, the reductionist approach in science leads to a causal model that has small fragments from complicated systems. However, constructing causal models of entire systems is difficult because the required information is distributed across the exhaustive literature. In the United States, The Defense Advanced Research Projects Agency (DARPA) created the Big Mechanism program (BMP) to develop technology for constructing, understanding, and reasoning of large, complicated systems such as climatic, economic, ecological, and biological systems. At present, BMP focuses on cancer signaling pathways, but the technology is intended to be applied for general purposes [1,2]. In the future, the BMP aims to produce machines that can read literature and assemble causal fragments found in individual articles into larger causal models. For example, the computer can gather information from the literature on cancer biology, extract fragments of causal mechanisms from publications, assemble the mechanisms into executable models of unprecedented scale and fidelity, use these models to explain and predict aspects of cancer biology, and even test these predictions in vitro [1,2]. BMP aims to use existing online sources of biological knowledge as well as existing machine reading and information extraction methods, such as big data and big data mining methods, to develop representation and inference methods for mechanistic biology models and methods for the machine to understand the claims and evidence in papers, algorithms, or system modeling for inferring causal relationships through system identification via big data and big data mining [2].

In this chapter, we introduce Big Mechanisms, which are mechanistic models of complicated systems with too many elements and relationships or too many possible current or future states to be easily comprehended by humans. Next, we introduce Big Mechanisms in systems biology via system modeling and identification through genome-wide data and big data mining methods. Since the focus of the BMP is on systems biology, in this book, the scope of Big Mechanisms in systems biology is on the system mechanisms of the cell cycle, signal flow, immunity, regeneration, infection, aging, evolution, carcinogenesis, and medicine based on system models, big database mining, and genome-wide high-throughput experimental data. This book is

Big Mechanisms in Systems Biology
DOI: http://dx.doi.org/10.1016/B978-0-12-809479-2.00001-9

designed to develop systematic methods to help scientists understand the big mechanisms of very complicated biological systems.

1.1 INTRODUCTION TO BIG MECHANISMS

The DARPA has recently introduced a search for technologies based on a new kind of science in which research is integrated automatically or semi-automatically into causal, explanatory models [1]. It emphasizes the need for solutions to complex systems such as ecosystems, brains, economics, and social systems that have parts and processes that are currently studied piecemeal, with literature and data that are fragmented, distributed, and inconsistent. It addresses the problem of information overload that we are all currently experiencing in "Big Data" environments. In recent years, "Big Data" has transformed science, engineering, medicine, healthcare, finance, business, military, and ultimately society itself. For example, analyses of command information flows during military crises have been suggested as an approach to ontogenetic learning that avoids both the problem of describing the subject covered in a document, and the problem of integrating new subject matter into a predetermined classification code. We need to make computer-readable documents and data, and assemble the fragments into explanatory pattern. This would be part of the Big Mechanisms that explain causes and effects within systems.

Therefore, Big Mechanisms are made up of many small mechanisms, which are dispersed and need to be integrated into a knowledge base to understand the mechanisms of big and complex systems. However, our knowledge of these mechanisms is increasingly fragmented, voluminous, and inconsistent. It is of central importance to plan and follow the big database mining process to create a knowledge base, and investigate how to put this knowledge base into both a relational and graph format to perform analyses and visualizations for the development of knowledge and for application. DARPA's goal is to develop technologies for a new kind of science in which many areas of research are integrated into causal, explanatory, or mechanistic models, with deeper semantics to represent the causal and often kinetic models in response to the challenge of big data that science and industry are focus today.

At present, the domain of the BMP is systems biology [1,3,4]. The amount of data produced by the study of biological systems in different species and at different scales, from molecules to ecosystems, is growing exponentially. As this increase in information presents challenges to some areas of biology, systems biology researchers employ bioinformatics technologies to handle large amounts of data, from local production through to storage in publicly accessible and integrative depositories. Systems biology is a new area of study in biology that seeks to incorporate bioinformatics, statistics, mathematics, physics, chemistry, biology, and engineering, and that promises to advance and transform our ability to increase biological understanding, program biological functions, and render the engineering of biological circuits or pathways faster

and more predictable. A large-scale, integrative, and multidisciplinary approach is needed for systems biology to flourish.

1.2 BIG MECHANISMS IN SYSTEMS BIOLOGY

This book addresses Big Mechanisms of systems biology by system identification and big data mining methods using models of biological systems [2,5]. Faced with a large volume of available literature, complicated mechanisms, small prior knowledge, few classes on the topic, and causal and mechanistic language, biological research is currently undergoing revolutionary changes in response to the integration of powerful technologies. Immunity, regeneration, infection, aging, evolution, and carcinogenesis are complicated biological systems, and related biological knowledge is fragmented, voluminous, and even inconsistent. The authors have observed remarkably poor agreement among different databases regarding components of these biological systems. These inconsistencies may reflect the underlying biology with time-varying systems and signal transduction events are often context-dependent. This raises a significant problem for mechanistic modeling since it is not clear which genes/proteins to include in models or experimental measurements.

In this book, we first construct candidate biological networks from omics data and big data mining that contain many false positives, interaction inconsistencies, and contradictions (see Fig. 1.1). System models are then employed to describe the relationship between components in the biological network. System identification (reverse engineering) methods are used to identify the interaction parameters of candidate biological network by experimental measurement via microarray data, NGS (next-generation sequencing) data, or real-time PCR data in a specific biological condition, such as lung cancer. Finally, the system order detection method is employed to determine the system order (the number of interactions) of the biological network so we can delete the insignificant interactions from the system order and prune false positives, inconsistencies, or contradictory interactions in the candidate biological network to obtain the real biological network of a specific biological condition.

After detection and removing inconsistencies between the experimental data and candidate biological network, biological networks in different biological conditions are obtained. We can compare two biological networks in different conditions to find significant changes in interactions or evolution of biological conditions (e.g., from a normal to cancer condition) as the network markers of these conditions change. From some related biological websites, like GO Enrichment Analysis, we can find biological functions or pathways to interpret system mechanisms of these network markers and identify adequate drug targets by methods to analyze sensitivity and robustness for the therapeutic treatment.

In this book, the Big Mechanisms of biological systems, like the immune system, regeneration process, photosynthetic system, and cancer are constructed by systems biology and big database mining methods. Systems biology sits at the heart of a new integrative 21st century paradigm of biology. It has introduced a search for a new

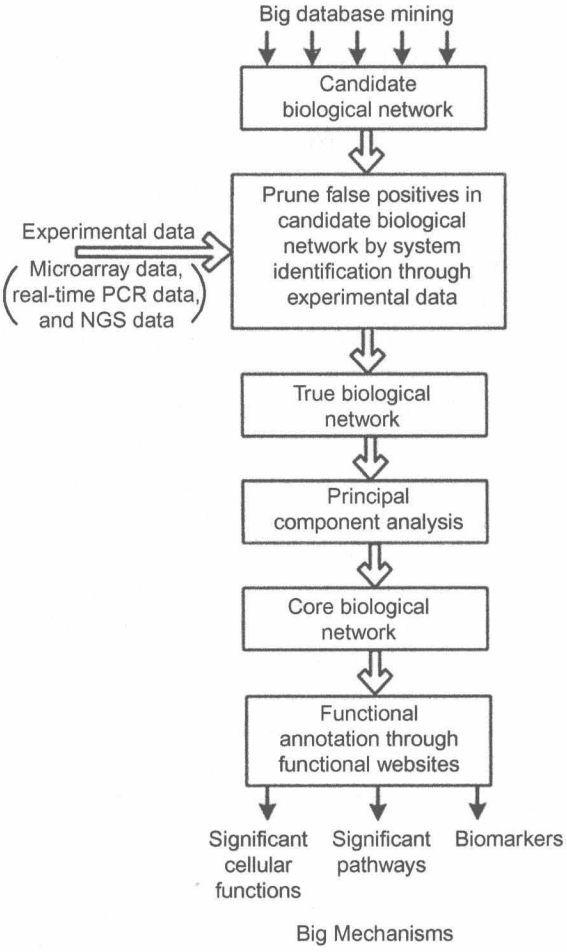


Figure 1.1 A block diagram of Big Mechanisms of systems biology via system identification and big data mining.

kind of scientific research based on big data mining and causal and explanatory models of experimental data. If we could resolve Big Mechanisms of systems biology by system modeling, system identification, and big data, the results could provide a paradigm for solutions to Big Mechanisms of complex systems, such as ecosystems, brains, economics, and social systems in the future.

1.3 THE SCOPE OF BIG MECHANISMS OF SYSTEMS BIOLOGY IN THIS BOOK

In order to describe the system mechanisms and behaviors, system modeling and system identification methods are introduced with big data mining in Chapter 2, System

Modeling and System Identification Methods for Big Mechanisms in Biological Systems. Linear and nonlinear models and related systematic and mathematic analyses for biological networks are introduced to investigate the Big Mechanisms in Systems Biology. Then, least square parameter estimation and maximum likelihood parameter estimation methods are introduced to estimate the system parameters of biological networks using experimental data or high-throughput data [5]. Some stochastic testing, systems theory, and system order detection methods are provided for statistical inference and hypothesis testing of biological mechanisms in bioinformatics and systems biology [6,7].

In Chapter 3, Procedure for Exploring Big Mechanisms of Systems Biology Through System Identification and Big Database Mining, how to construct Big Mechanisms in systems biology by integrating many smaller mechanisms from the dispersed heterogeneous omics data via system models and big database mining is introduced. Since the main domain of Big Mechanisms is systems biology, several examples, i.e., how to construct a gene regulatory network (GRN), protein–protein interaction network (PPIN), and integrated genetic and epigenetic networks (IGEN) from omics data, are given to illustrate the procedure for following big data mining processes and system identification procedures to elucidate system-based Big Mechanisms.

The most remarkable features of cells and entire organisms are their ability to reproduce, and the cell cycle entails an ordered series of macromolecular events that lead to cell division and the production of two daughter cells. As a consequence, we choose Chapter 4, Big Cellular Mechanisms in the Cell Cycle by System Identification and Big Data Mining, to look at the regulation of transcription in the progression of the yeast cell cycle as an example for analyzing Big Cellular Mechanisms using big data mining processes and system identification procedures in yeast [8–10].

Since the control of transcription factors (TFs) on their target genes is one of the most important cellular mechanisms, it is more appealing to identify the regulatory trans/cis mechanisms of TFs via dynamic regulatory models and genome-wide microarray data. Therefore, in Chapter 5, Big Regulatory Mechanisms in the Transcriptional Regulation Control of Gene Expression Using a Stochastic System Model and Genome-Wide Experimental Data, stochastic system models and experimental data are used to analyze Big Regulatory Mechanisms in transcription control [11–13].

The cells sense extracellular signals through intercellular communication or stress responses, cellular responses to sudden environmental stresses or physiological changes that provide living organisms with the opportunity for survival and further development. Therefore, defending against environmental stresses with protective mechanisms is an important topic. In Chapter 6, Big Mechanisms of Information Flow in Cellular Systems in Response to Environmental Stress Signals via System Identification and Data Mining, cellular systems under environmental stresses are given as examples to identify Big Mechanisms of information flow and protective mechanisms in response to environmental stresses using system identification and database mining [14–17].

The major task of the immune system is to defend the host against infections. Recently, a large variety of experimental techniques and high-throughput NGS data has been created and provides a sound scientific basis for integrative approaches to advanced, quantitative, and qualitative systems biology, using computation and modeling that will allow us to understand Big Mechanisms of complex immune responses. In Chapter 7, Big Offensive and Defensive Mechanisms in Systems Immunity From System Modeling and Big Data Mining, system modeling and big data mining are used to identify the big offensive and defensive mechanisms in the immune system [18–20].

Regeneration is one of the most intriguing and fascinating biological phenomena, but the molecular and cellular bases of regeneration are still not fully understood. Unlike the vertebrates with high regenerative capacity such as amphibians and fish, mammals have a restricted ability to regenerate lost cells, tissues, and organs at the adult stage. Therefore, understanding the molecular mechanism of regeneration in organisms is not only valuable in shedding light on the longstanding question of regeneration mechanisms, but it is also useful for stem cell-based therapies in the future. In Chapter 8, Big Regeneration Mechanisms via Systems Biology and Big Database Mining Methods, Big Regeneration Mechanisms are given as an example of using systems biology and big database mining methods [21,22].

Cancer is a genetic disease. Many screening arrays for finding new cancer genes have been constructed. Cancer is also a systemic disease. Therefore, tumorigenesis mechanisms should be investigated from a systems biology perspective. In Chapter 9, Big Tumorigenesis Mechanisms in Systems Cancer Biology via Big Database Mining and Network Modeling, cancer systems biology is used to identify Big Tumorigenesis Mechanisms through big database mining and network modeling [20,23–25].

Living organisms are complex systems characterized by emergent properties. A ubiquitous property in complex systems is robustness. Robustness is not a trivial biological mechanism to be studied using a reductionism approach. Robust mechanisms are frequently and widely distributed, guaranteeing organization of the system and cooperation between various parts. Signal transductivity, a system property, is complementary to robustness, i.e., a biological system with greater network robustness will show greater loss of signal transductivity and vice versa. Since aging and cancer are both systems diseases of somatic cell evolution, it is appealing to investigate the network robustness and signal transductivity of somatic cells in the aging and carcinogenic processes. In Chapter 10, Big Evolutionary Mechanisms of Network Robustness and Signaling Transductivity in Aging and Carcinogenic Process by System Modeling and Database Mining, the Big Mechanisms in aging processes and carcinogenesis are given as examples of examining the mechanisms of network robustness and signaling transductivity through system modeling and database mining [26,27].

Aging, an extremely complex and system-level process, has attracted much attention in medical research, especially as chronic diseases are quite prevalent in the elderly

population. These may be the result of both accumulated genetic and epigenetic variations that lead to intrinsic perturbations and environmental changes that may stimulate signaling in the body. Decades of research have demonstrated that higher incidence of many pathophysiological phenotypes, such as cancer, inflammation, diabetes, cardiovascular diseases, and neurodegenerative disorders, are associated with the process of aging. In Chapter 11, Big Mechanisms of Aging via System Identification and Big Database Mining, we first investigate the systematic mechanism of GRN in the aging process by systems biology method via microarray data. Then we will investigate specific core genetic and epigenetic network for cellular mechanisms of human aging [28].

In the past few decades, medicine has sought a "magic bullet" that targets a simple disease-causing molecule and that could cure cancer, diabetes, and other complex diseases. Although some drugs have proven successful, many others have been found to be ineffective or the cause of significant side effects. This disappointing outcome highlights the limitation of single target drug paradigms and is considered to be the underlying cause of stagnation in the productivity of pharmaceutical industry. However, compound efficacy and safety in humans, including toxicity and pharmacokinetic profiles, rather than target selection, are the criteria that determine which drug candidates enter the clinic. Therefore, drug discovery using database mining and systems biology will become the most important approach to drug discovery. Finally, in Chapter 12, Big Drug Design Mechanisms via Systems Biology and Big Database Mining, Big Drug Design Mechanisms are introduced as therapeutic treatments using systems biology and big database mining. At first, we will give an overview of drug discovery using systems biology, and then introduce how to investigate core and specific network markers for multiple drug targets from the perspective of systems biology. Finally, a cocktail drug design is also introduced for multiple drug targets [29,30].

REFERENCES

[1] A data science Big Mechanism for DARPA, <http://semanticommunity.info/>.
[2] Cohen PR. DARPA's Big Mechanism program. Phys Biol 2015;12(4):1–9.
[3] Chen B-S, Wu C-C. Systems biology: an integrated platform for bioinformatics, systems synthetic biology and systems metabolic engineering. New York, NY: Nova Science Pub Inc.; 2014.
[4] Klipp E, Liebermeister W, Wierling C, Kowald A, Lehrach H, Herwig R. Systems biology. Weinheim: Wiley; 2011.
[5] Johansson R. System modeling and identification. Englewood Cliffs, NJ: Prentice Hall; 1993.
[6] Chen BS, Li CW. Constructing an integrated genetic and epigenetic cellular network for whole cellular mechanism using high-throughput next-generation sequencing data. BMC Syst Biol 2016;10:18.
[7] Wang YC, Chen BS. Integrated cellular network of transcription regulations and protein-protein interactions. BMC Syst Biol 2010;4:20.
[8] Chen HC, Lee HC, Lin TY, Li WH, Chen BS. Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. Bioinformatics 2004;20(12):1914–27.

[9] Wu WS, Li WH, Chen BS. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. BMC Bioinformatics 2006;7:421.
[10] Wu WS, Li WH, Chen BS. Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. BMC Bioinformatics 2007;8:188.
[11] Chang YH, Wang YC, Chen BS. Identification of transcription factor cooperativity via stochastic system model. Bioinformatics 2006;22(18):2276–82.
[12] Chang YH, Wang YC, Chen BS. Nonlinear dynamic trans/cis regulatory circuit for gene transcription via microarray data. Gene Regul Syst Bio 2007;1:151–66.
[13] Lin LH, Lee HC, Li WH, Chen BS. Dynamic modeling of cis-regulatory circuits and gene expression prediction via cross-gene identification. BMC Bioinformatics 2005;6:258.
[14] Li CW, Chen BS. Identifying functional mechanisms of gene and protein regulatory networks in response to a broader range of environmental stresses. Comp Funct Genomics 2010;Article ID:408705, 20 pages.
[15] Wu WS, Li WH, Chen BS. Reconstructing a network of stress-response regulators via dynamic system modeling of gene regulation. Gene Regul Syst Bio 2008;2:53–62.
[16] Yao CW, Hsu BD, Chen BS. Constructing gene regulatory networks for long term photosynthetic light acclimation in *Arabidopsis thaliana*. BMC Bioinformatics 2011;12:335.
[17] Przytycka TM, Kim YA. Network integration meets network dynamics. BMC Biol 2010;8:48.
[18] Chen BS, Yang SK, Lan CY, Chuang YJ. A systems biology approach to construct the gene regulatory network of systemic inflammation via microarray and databases mining. BMC Med Genomics 2008;1:46.
[19] Kuo ZY, Chuang YJ, Chao CC, Liu FC, Lan CY, Chen BS. Identification of infection- and defense-related genes via a dynamic host-pathogen interaction network Using a *Candida albicans*-zebrafish infection model. Journal of Innate Immunity 2013;5(2):137–52.
[20] Chen BS, Li CW. Constructing an integrated genetic and epigenetic cellular network for whole cellular mechanism using high-throughput next-generation sequencing data. BMC Syst Biol 2016;10(1):18.
[21] Wu CC, Lin C, Chen BS. Dynamic network-based relevance score reveals essential proteins and functional modules in directed differentiation. Stem Cells Int 2015;2015:792843.
[22] Wu CC, Tsai TH, Chang C, Lee TT, Lin C, Cheng IHJ, et al. On the crucial cerebellar wound healing-related pathways and their cross-talks after traumatic brain injury in *Danio rerio*. PLoS One 2014;9(6).
[23] Chu LH, Chen BS. Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. BMC Syst Biol 2008;2:56.
[24] Li CW, Chu YH, Chen BS. Construction and clarification of dynamic gene regulatory network of cancer cell cycle via microarray data. Cancer Inform 2006;2:223–41.
[25] Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Med Genomics 2011;4:2.
[26] Chen BS, Tsai KW, Li CW. Using nonlinear stochastic evolutionary game strategy to model an evolutionary biological network of organ carcinogenesis under a natural selection scheme. Evol Bioinform 2015;11:155–78.
[27] Tu CT, Chen BS. On the increase in network robustness and decrease in network response ability during the aging process: a systems biology approach via microarray data. IEEE-ACM Trans Comput Biol Bioinform 2013;10(2):468–80.
[28] Li CW, Wang WH, Chen BS. Investigating the specific core genetic-and-epigenetic networks of cellular mechanisms involved in human aging in peripheral blood mononuclear cells. Oncotarget 2016;7(8):8556–79.
[29] Chen BS, Li CW. Analysing microarray data in drug discovery using systems biology. Expert Opin Drug Discov 2007;2(5):755–68.
[30] Wong YH, Lin CL, Chen TS, Chen CA, Jiang PS, Lai YH, et al. Multiple target drug cocktail design for attacking the core network markers of four cancers using ligand-based and structure-based virtual screening methods. BMC Med Genomics 2015;8(Suppl. 4):S4.

CHAPTER 2

System Modeling and System Identification Methods for Big Mechanisms in Biological Systems

INTRODUCTION

Consider the cellular system as shown in Fig. 2.1. Through intercellular communication and cellular stress responses, the cells sense extracellular signals. Different external changes or events may induce signaling in cells. Typical signals are hormones, pheromones, pathogens, heat, cold, light, osmotic pressure, and changes in the appearance or concentration of substances such as glucose, potassium ions, calcium ions, or cyclic adenosine monophosphate (cAMP) [1–7]. In the signal transduction pathway, extracellular signals are perceived by transmembrane receptors. The receptor changes its own state from susceptible to active and then it triggers subsequent cellular processes.

The active receptor stimulates an internal signaling cascade. This cascade frequently includes a series of changes in protein phosphorylation states. The sequence of state changes crosses the nuclear membrane. Eventually, some transcription factors (TFs) are activated or deactivated, which change their ability to bind a set of genes that produce the corresponding proteins in response to extracellular signals or stresses. In addition to genetic regulation by TFs, epigenetic regulation due to DNA methylation and microRNAs (miRNA) can also alter gene expression. However, owing to the complex nature of dynamic biological systems, knowledge of their components and interactions is not sufficient to interpret behaviors in the cellular system. In this book, as shown in Fig. 1.1, we employ big database mining to construct candidate-integrated genetic and epigenetic networks (GENs) to interpret some cellular mechanisms for some biological conditions. However, there exists a large amount of false positives in a big database mining process. Therefore, it is very important to prune these false positive regulations and interactions from the candidate GEN by real high-throughput experimental data. We need to model interactions and regulations in the candidate GEN and then prune these false positive interactions and regulations to obtain the real GEN through system modeling and identification methods by microarray data or next generation sequencing (NGS) data.

Therefore, in this chapter, we will describe the models for these cellular systems and use experimental data, such as microarray data, NGS data, or real-time polymerase

Big Mechanisms in Systems Biology
DOI: http://dx.doi.org/10.1016/B978-0-12-809479-2.00002-0

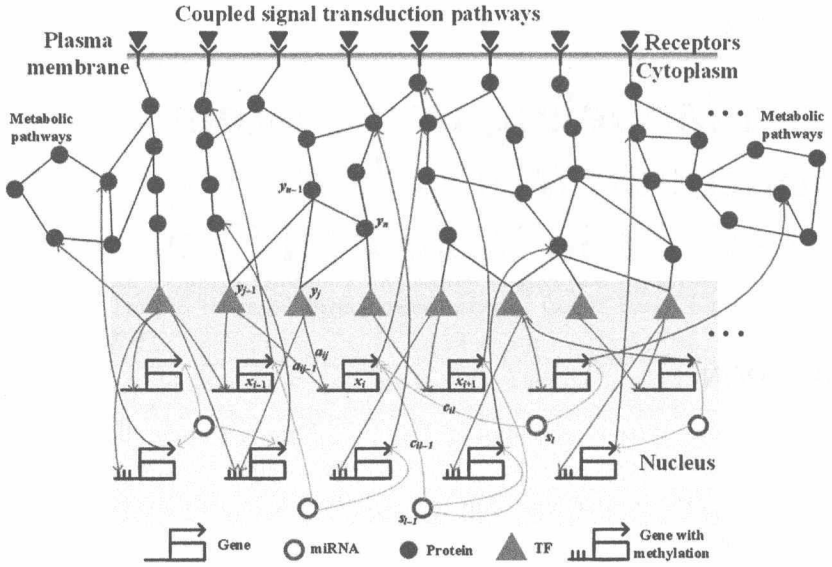


Figure 2.1 An integrated signal transduction pathway (i.e., protein–protein interaction network (PPIN), gene regulatory network (GRN), and miRNA regulatory network).

chain reaction (PCR) data, to identify the parameters of the model that will allow us to efficiently and accurately integrate this data into the Big Mechanism, and to interpret the behaviors of cellular systems by system modeling and identification methods. To begin, we will model and identify the protein–protein interaction network (PPIN) and gene regulatory network (GRN) by protein expression and microarray data, respectively. Then, these genetic networks will be integrated with epigenetic regulation by miRNA expression data to construct an integrated genetic and epigenetic cellular network. Further, core network extracted by applying principal network projection (PNP) to the integrated genetic and epigenetic cellular network is also introduced in this chapter to investigate significant network mechanism by principal component analysis (PCA) via singular value decomposition (SVA) method.

For the convenience of illustration, only linear system models are introduced for system identification of biological networks to investigate the Big Mechanisms through genome-wide high-throughput data. When necessary, the system identification methods based on nonlinear system models will be introduced in the following chapters.

2.1 DYNAMIC SYSTEM MODELS AND THEIR PARAMETER ESTIMATION BY TIME-PROFILE EXPERIMENTAL DATA

For the simplicity of analysis, we first discuss only the coupled signal transduction pathways shown in Fig. 2.2, which is extracted from Fig. 2.1. In Fig. 2.2, $y_i(t)$ denotes