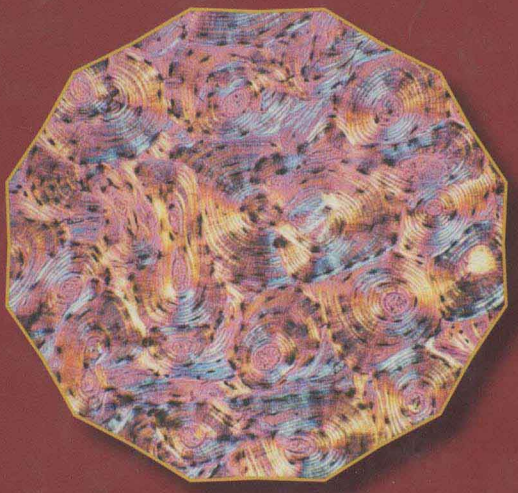
Wiley Series in Bioinformatics • Yi Pan and Albert Y. Zomaya, Series Editors

Elements of Computational Systems Biology



Edited by
Huma M. Lodhi
Stephen H. Muggleton

WILEY

ELEMENTS OF COMPUTATIONAL SYSTEMS BIOLOGY

Edited by

Huma M. Lodhi
Stephen H. Muggleton

Imperial College London, London, UK

WILEY

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Lodhi, Huma M.
Elements of computational systems biology / Huma M. Lodhi, Stephen H. Muggleton.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-18093-8 (cloth)
1. Systems biology. 2. Computational biology. I. Muggleton, Stephen. II. Title.
QH324.2.L64 2010
570.285–dc22 2009028768

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ELEMENTS OF COMPUTATIONAL SYSTEMS BIOLOGY

Wiley Series on
Bioinformatics: Computational Techniques and Engineering

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of *biology*, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

Series Editors: **Professor Yi Pan** and **Professor Albert Y. Zomaya**
pan@cs.gsu.edu zomaya@it.usyd.edu.au

PREFACE

Recently there has been a huge interest in the development of computational methodologies for modeling and simulating biological processes. The book facilitates the design of effective and efficient techniques by introducing key elements of the emerging field of computational systems biology. It gives an in-depth description of core subjects including biological network modeling, analysis, and inference. It presents a measured introduction to foundational topics such as genomics and describes state-of-the-art software tools.

The collaborations between experts from highly diverse areas ranging from biology to computer science are crucial for the progress in computational systems biology. The book is aimed at fostering close collaborations between biologists, chemists, physicists, mathematicians and computer scientists by providing ground-breaking research. It provides an inspiration and basis for the future development and applications of novel computational and mathematical methods to solving complex and unsolved problems in biology.

The book is intended for researchers and scientists from the fields of biology, chemistry, mathematics, physics, and computer science who are interested in computational systems biology or focused on developing, refining, and applying computational and mathematical approaches to solving biological problems. It is organized in a way so that the experts from the industry such as biotechnology and pharmaceutical companies will find it very useful and simulating. The book is accessible to students and provides knowledge that he/she requires.

We wish to thank Wiley for the support and help in the processing of the book. We would also like to thank Yanqing Zhang, Bart Bijnens, Antti Honkela, Zhongming Zhao, Nicos Angelopoulos, Roman Rosipal, Jae-Hyung Lee, Zhaolei Zhang, Ying Liu, Wenyuan Li, Dong Xu, Giovani Gomez Estrada, Li Liao, Leming Zhou, and Etienne Birmele for their help in the reviewing process.

H. M. LODHI AND S. H. MUGGLETON

Feburary, 2009

CONTRIBUTORS

Tatsuya Akutsu, Kyoto University, Kyoto, Japan

Panayiotis V. Benos, Departments of Computational Biology and Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Alessandra Carbone, Génomique Analytique, Université Pierre et Marie Curie–Paris 6, FRE3214 CNRS-UPMC, 15 rue de l'Ecole de Médecine, 75006 Paris, France

Carsten Carlberg, Life Sciences Research Unit, University of Luxembourg, Luxembourg, UK

Matteo Cavaliere, The Microsoft Research - University of Trento, CoSBi, Trento, Italy

Wilbur E. Channels, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

Wai-Ki Ching, University of Hong Kong, Hong Kong, China

Kwang-Hyun Cho, Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305–701 Korea

Sang-Mok Choo, School of Electrical Engineering, University of Ulsan, Ulsan, Korea

Vincent Danos, University of Edinburgh, Edinburgh, UK

Andrei Doncescu, LAAS-CNRS, Avenue du Colonel Roche, Toulouse, France

Jérôme Féret, ENS-INRIA-CNRS, Paris, France

Duncan Gillies, Department of Computing, Imperial College London, London, UK

Jeremy Gunawardena, Department of Systems Biology, Harvard Medical School, Boston, MA, USA

Jörg Hakenberg, Computer Science and Engineering, Arizona State University, Tempe, AZ, USA

Russell Harmer, CNRS-Paris-Diderot, Paris, France

Merja Heinäniemi, Life Sciences Research Unit, University of Luxembourg, Luxembourg

Pablo A. Iglesias, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

Katsumi Inoue, Department of Informatics, The Graduate University for Advanced Studies and National Institute of Informatics, Chiyoda-ku, Tokyo, Japan

Tae-Hwan Kim, Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305–701 Korea

Jean Krivine, Harvard University, Cambridge, MA, USA

Ulf Leser, Knowledge Management in Bioinformatics, Humboldt-Universität zu Berlin, Berlin, Germany

Huma M. Lodhi, Department of Computing, Imperial College London, London SW7 2AZ, UK

Anthony Mathelier, Génomique Analytique, Université Pierre et Marie Curie–Paris 6, FRE3214 CNRS-UPMC, 15 rue de l'Ecole de Médecine, 75006 Paris, France

Itay Mayrose, Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

Tommaso Mazza, The Microsoft Research - University of Trento, CoSBi, Trento, Italy

Alok Mishra, Department of Computing, Imperial College London, London, UK

Elaine Murphy, University of Edinburgh, Edinburgh, UK

Conrad Plake, Biotechnological Centre, Technische Universität Dresden, Dresden, Germany

Tal Pupko, Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

Sung-Young Shin, Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305–701 Korea

Alain B. Tchagang, Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Jean-Philippe Vert, Mines ParisTech – Institut Curie, Paris, France

Stephen T. C. Wong, The Center for Biotechnology and Informatics, The Methodist Hospital Research Institute and Department of Radiology, The Methodist Hospital Weill Cornell Medical College, Houston, TX, USA

Ling-Yun Wu, The Center for Biotechnology and Informatics, The Methodist Hospital Research Institute and Department of Radiology, The Methodist Hospital Weill Cornell Medical College, Houston, TX, USA

Yoshitaka Yamamoto, Department of Informatics, The Graduate University for Advanced Studies, Chiyoda-ku, Tokyo, Japan

Yoshihiro Yamanishi, Mines ParisTech - Institut Curie, Paris, France

Xiaobo Zhou, The Center for Biotechnology and Informatics, The Methodist Hospital Research Institute and Department of Radiology, The Methodist Hospital Weill Cornell Medical College, Houston, TX, USA

I

OVERVIEW

1

ADVANCES IN COMPUTATIONAL SYSTEMS BIOLOGY

Huma M. Lodhi

Department of Computing, Imperial College London, London, SW7 2AZ, UK

1.1 INTRODUCTION

Computational systems biology, a rapidly evolving field, is at the interface of computer science, mathematics, physics, and biology. It endeavors to study, analyze, and understand complex biological systems by taking a coordinated integrated systems view using computational methodologies. From the middle of the twentieth century till present, we have been witnessing breakthrough discoveries in biology that range from molecular structure of deoxyribonucleic acid (DNA) to the generation of the sequence of the euchromatic portion of the human genome. There have also been recent advances in sophisticated computational methodologies, high-throughput biotechnologies, and computational power. The stunning developments in diverse disciplines such as biology and computer science are playing a key role in the fast progression of the emerging field. Computational systems biology provides a point of convergence for genomics, proteomics, metabolomics, and computational modeling. It is characterized by its focus on experimental data, computational techniques, and hypotheses testing [1–3].

Open and unsolved problems in biology range from understanding structure and dynamics of biological systems to prediction and inference in the complex systems.

Elements of Computational Systems Biology Edited by Huma M. Lodhi and Stephen H. Muggleton

In the postgenomic era, systems-based approaches may provide a solution to such unsolved problems. It is believed that some answer to the question "what is life" may be obtained by taking a broader, integrated view of biology [4]. However, applications of systems-based techniques to biology are not new. Such methods and frameworks have been applied to analyze biological processes since early twentieth century [5, 6]. Norbert Wiener's groundbreaking work [7] is a well-known example of these applications.

The purpose and objective of this chapter is to review cutting-edge and long-ranging research in the field of computational systems biology in the recent years. However, the review is not meant to be exhaustive. We briefly describe novel methodologies to build multiscale biological models in Section 1.2. In Section 1.3, we present an overview of the applications of proteomics techniques to study biological processes. We then summarize computational systems biology methods to examine and understand aging in Section 1.4. Section 1.5 describes systems-based techniques for drug design, where such methods are revolutionizing the process of drug discovery. Efficient software tools and infrastructure are crucial to solving complex biological problems. In Section 1.6, we review tools for systems biology.

1.2 MULTISCALE COMPUTATIONAL MODELING

In the postgenomic era, researchers seek to focus their attention to studying and analyzing biological networks and pathways by the use of multiscale computational modeling techniques. A model can be viewed as a representation of a biological system, where the representation can comprise a set of differential equations [8], a set of first-order logic clauses [9], and so on. Biological models that incorporate multiple scales such as time and space or multiple timescales may be viewed as multiscale models [10]. Chapter 2 gives an in-depth account of mathematical and computational models in systems biology.

Development of efficient and effective computational methodologies to perform modeling, simulation, and analysis of complex biological processes is a challenging task. Traditionally, mathematical and computational models have been developed by considering a single scale. However, it is now feasible to incorporate multiple scales in the process of model building due to recent advances in computational power and technology. Generally, multiscale models are constructed by using sophisticated techniques including numerical methods and integration approaches. Multiscale model of the heart [11, 12] is a well-known example of an application of these modeling techniques.

Multiscale computational modeling and simulation methods are showing promising results in the field of oncology. The development of three-dimensional multiscale brain tumor model by Zhang et al. [13] is an attempt in this direction. The dynamics of tumor growth were simulated by using an agent-based multiscale model where microscopic scale, macroscopic scale, and molecular scale were incorporated in the *in silico* model. In micro-macroscopic environment, a virtual brain tissue block was represented by points in three-dimensional lattice. The lattice was

divided into four cubes that illustrated the behavior of chemotactically acting tumor cells. The chemotaxis distribution of transforming growth factor alpha (TGFα), glucose, and oxygen tension were illustrated in a set of mathematical equations. It was observed that the amount of TGFα and glucose was chemoattractant, and diffusion of glucose occurred at a constant rate. In order to incorporate molecular scale, epidermal growth factor receptor (EGFR) gene–protein interaction network model [14] was used in conjunction with cell cycle module. The authors used a simplified EGFR network that comprised of EGFR and TGFα genes. The mathematical model of EGFR gene–protein network was represented as a set of differential equations. The authors utilized the cell cycle model presented in Tyson and Novak [15] and Alacron et al. [16]. The implementation of the software systems was carried out by combining in-house code with an agent-based software tool, namely, MASON (`http://cs.gmu.edu/ eclab/projects/mason/`). In order to study and analyze tumor growth and spread, 10 simulations were performed. The results demonstrated an increase in tumor volume with respect to time, where the relationship between tumor volume and time was not linear. There was a sharp increase in volume growth at later time intervals. The study found that migrating and proliferating cells exhibited a dynamic behavior with respect to time. Furthermore, the cells caused spatiotemporal tumor growth. The results showed that the number of migrating cells was greater than the number of proliferating cells over time, where the high concentration of phospholipase C gamma (PLCγ) might be the key factor behind the phenomenon. In summary, the study demonstrated a successful construction of multiscale computational model of the complex multifaceted biological process. However, the approach is not free from shortcomings as described below:

- A simple EGFR network was used.
- Clonal heterogeneity within tumor was not examined.

It has been found that the distribution of tumor cells is not homogeneous, and the cells exhibit heterogeneous patterns. Techniques that account for clonal heterogeneity of tumor cell populations can be vital to analyze and study the development of cancerous diseases. Furthermore, clonal heterogeneity can strongly impact the design of effective therapeutic strategies. Therefore, many studies examined heterogeneity in tumors [17, 18]. Zhang et al. [19] extended their multiscale computational modeling technique [13] to investigate the clonal heterogeneity by incorporating genetic instability. The extended model included doubling time of cell and cell cycle. Other parameters such as cell–cell adhesion were also considered so that the strength of the chemoattractants' (TGFα, oxygen tension, and glucose) impact on cancer cells adhesion and rate of cell migration could be investigated. The authors used Shannon's entropy for the quantification of tumor heterogeneity. Shannon entropy in this context can be calculated as follows: Let c_i denote the occurrence of clone i in the tumor, the entropy is given by $\sum_i c_i \ln(c_i)$, where the higher values of Shannon's entropy represent more clonal heterogeneity.

The results of the study showed an increase in tumor total volume over time, where the tumor was categorized into three regions on the basis of the distance between it

and the nutrient source. It was observed that there was a general increase in the values of Shannon's entropy for all the three regions. However, there was highest clonal heterogeneity in the region closest to the nutrient source at early time stages where the region exhibited a homogeneous pattern at later stages. The study inferred that cancer could spread faster due to clonal heterogeneity as compared to homogeneous cell populations in tumor.

The complexity of the mechanisms of development and morphogenesis establishes a need to design effective and efficient computational techniques to investigate and analyze the biological process. In a recent study, Robertson et al. [20] presented a multiscale computational framework to investigate morphogenesis mechanisms in *Xenopus laevis*. Mammalian cells share similarities with *X. laevis* in terms of signaling network and cell behavior. A multiscale model was constructed by integrating an intercellular signaling pathway model with the multicellular model of mesendoderm migration. The authors implemented Wnt/β-catenin signaling pathway model that was presented by Lee et al. [21], whereas an agent-based approach was applied to build mesendoderm migration model. In order to simulate mesendoderm cells' migration, it was viewed that each cell comprised of nine sections, where each section was modeled as an agent. Mesendoderm migration was facilitated by the use of fibronectin extracellular matrix substrate. The study found that fibronectin gradient was a key factor behind the cellular movement. It was also observed that polarity signals [22] might be important for mesendoderm migration and morphogenesis. The simulations also demonstrated the importance to keep the cadherin binding strength in balance with the integrin binding strength. Although the study establishes the efficacy of multiscale computational methodologies to studying morphogenesis, the proposed approach may not be computationally attractive for large-scale simulations.

Physiome project [12] is well known for the development of multiscale modeling infrastructures. Given that standard modeling languages are useful for sharing biological data and models, three markup languages, namely, CellML (`http://www.cellml.org/`), FieldML, and ModelML, have been developed in the project. CellML [23] is characterized by its ability to capture three-dimensional information regarding cellular structures. It can also incorporate mathematical knowledge and metadata. FieldML, a related language, is known for its incorporation of spatial information. The third systems biology modeling language, namely, ModelML, is characterized by its ability to encode physical equations that illustrate complex biological processes. The efficacy of the languages was established by building multiscale heart models [12].

It has been found that same input, to constituent parts of a system, can produce different outputs. Such variations may be produced by factors including alterations in the concentration of system's components. It is desirable to design techniques and methods that can provide robustness to variations. Shinar et al. [24] presented a robust method by exploiting molecular details. The authors coined the term "input–output relation" for the association between input signal strength and output. The study investigated the input–output relation in bacterial signaling systems.

1.3 PROTEOMICS

Proteomics, the study of proteins, is viewed crucial to analyze and understand biological systems, as protein is the building block of life. Mass spectrometry (for details see Chapter 17) is a well-known proteomics technology that is showing a huge impact on the development of the field of computational systems biology. Several recent studies have identified the significant role of proteomics techniques in solving complex biological problems [25–27].

Proteomics methods and data can be useful for the reconstruction of biological networks. Recently, Rho et al. [28] presented a computational framework to reconstruct biological networks. The framework is based on the use of proteomics data and technologies to build and analyze computational models of biological networks. It is termed as integrative proteomic data analysis pipeline (IPDAP). IPDAP incorporates a number of network modeling and analysis tools. The component tools of IPDAP can be applied to reconstruct biological networks by fusing different types of proteomics data. The successful application of IPDAP to different cellular and tissue systems demonstrated the efficacy and functionality of the framework.

In another study, Zhao et al. [29] investigated signal transduction by applying techniques from optimization theory and exploiting proteomics and genomics data. They formulated the network identification problem as an integer linear programming problem. The proteomics (protein–protein interaction) data were represented as weighted undirected graph, where the nodes and the edges represented proteins and interaction between pair of proteins, respectively. The results of the study confirmed the efficacy of the approach in searching optimal signal transduction networks from the data.

Cell cycle comprises a series of ordered events by which cell replication and division take place. Studying cell cycle regulation provides useful insights in cancer growth and spread. The relationship between cell cycle and cancer has been a focus of many studies [30, 31]. In Sigal et al. [32], a proteomics approach was applied to investigate cell cycle mechanisms. The approach is based on the use of time-lapse microscopy to study protein dynamics. The study identified cell cycle-dependent changes in protein localization, where 40 percent of the investigated nuclear proteins demonstrated cell cycle dependence. Another challenging problem is to find patterns of polarized growth in cells where such growth is viewed as an important process in organisms. In order to investigate the biological problem, Narayanaswamy et al. [33] conducted a study by using budding yeast as the model system. The proposed computational method is based on the use of microarray image analysis and a machine learning technique, namely, naive Bayes algorithm. The study found 74 localized proteins including previously uncharacterized proteins and observed novel patterns of cell polarization in budding yeast.

In a recent study [34], a computational technique is presented for predicting peptide retention times. The method is at the intersection of two machine learning approaches, namely, neural networks and genetic algorithms. In order to predict the retention times, an artificial neural network is trained and the predicted values are further optimized

by using a genetic algorithm. The method was successfully applied to *Arabidopsis* proteomics data.

1.4 COMPUTATIONAL SYSTEMS BIOLOGY AND AGING

Aging is a complex phenomenon that has not been well understood. In aging, we witness gradual diminishing/decreasing functions at different levels, including organs and tissues. Cell division has been viewed as a key process in aging since long [35, 36]. Recently, de Magalhaes and Faragher [37] have elucidated that aging might be affected by variations in cell division. Hazard rates and nutrition may be the key factors that influence the longevity of cellular organisms [38]. There are a number of theories that describe how aging occurs. Kirkwood [38] listed five different theories that are as follows:

- Somatic mutation theory
- Telomere loss theory
- Mitochondrial theory
- Altered proteins and waste accumulation theory
- Network theory

Aging has been extensively studied in *Caenorhabditis elegans* (nematode), mice, humans, and fruit flies. A number of genes that extend organisms' life span have been discovered. Several studies on aging found that genetic mutations could increase longevity [39–41]. Furthermore, aging genes with their associated pathways may influence the variations in aging between different species but may not have any affect on the differences in aging within a particular specie [42]. Gene expression and pathway analysis can provide useful means to identify aging-related similarities and differences between various species [43], where the efficacy of DNA microarray technology, in studying aging, is significant [44]. In a recent study on aging, DNA microarray experiments were utilized to show that aging in *C. elegans* is influenced by GATA transcriptional circuit [45].

Advances in computational systems biology have led to the development of tools and methods for solving highly complex problem of aging. For example, Xue et al. [46] addressed the key issue regarding aging by applying an analytic method to human/fruit fly protein–protein interaction network, namely, NP analysis [47]. The method is based on the identification of active modules in network, where the chosen module comprised of protein–protein interaction subnetwork between genes that show (positive or negative) correlation during aging. The application of the method to human brain aging identified four modules. Among these modules, the two showed transcriptionally anticorrelation with each other. The other two modules comprised of immunity genes and translational genes, respectively. In order to study correlation between genes in other species during aging, the method was applied to fruit fly interactome. The results of the study showed that in addition to two transcriptionally anticorrelated