Computational Biology

Volume 24

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at http://www.springer.com/series/5769

Sourav S. Bhowmick · Boon-Siew Seah

Summarizing Biological Networks

Springer

Sourav S. Bhowmick
School of Computer Science
and Engineering
Nanyang Technological University
Singapore
Singapore

Boon-Siew Seah
School of Computer Science
and Engineering
Nanyang Technological University
Singapore
Singapore

ISSN 1568-2684
Computational Biology
ISBN 978-3-319-54620-9 ISBN 978-3-319-54621-6 (eBook)
DOI 10.1007/978-3-319-54621-6

Library of Congress Control Number: 2017935562

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To our parents and wives.

Preface

Data do not give up their secrets easily. They must be tortured to confess.

Jeff Hopper, Bell Labs

The desire to study biology from a systems perspective has led to an emergence of new science—biological network analysis. Biological network models biological entities (e.g., proteins and genes) and their relationships (e.g., physical and genetic interactions) to characterize their cooperative activity within a system. With the rapid growth of such network data, the information overload problem has become a major stumbling block to analyze these networks, making human interpretation of such data increasingly difficult. Hence, there is a growing need to construct methods for large-scale topological and functional summaries of biological networks to understand the underlying mechanics of biological systems.

This book presents frameworks, as they stand today, that allow biologists to rapidly visualize and comprehend high-level topological and functional summary of the processes that govern biological systems via topological or functional organization *within* a biological network (intra-system processes) and relationships *between* biological networks (inter-system processes). Drawing on well-founded principles in data mining, systems biology, and bioinformatics, we present a multi-resolution and multi-perspective analysis paradigm to address this broad goal. Note that it is reasonable to expect this picture to change with time.

As a representative example of biological networks, we utilize protein–protein interaction (PPI) networks in majority part of this book. Our discussion is divided into five parts. First, we have attempted to review, as accurately as possible, a wide spectrum of approaches proposed by the bioinformatics community to cluster PPI networks and highlight their strengths and limitations. The results of such clustering can be considered as a summary of topological or functional modules in the underlying PPI network. In particular, a pervasive desire of this review is to

emphasize the uniqueness of the network clustering problem in the context of PPI networks and highlight why a panoply of generic network clustering algorithms proposed by the data mining community cannot be leveraged to address this problem effectively.

Second, we review a closely related problem to PPI network clustering, functional summarization, which can enable us to make sense out of the information contained in large PPI networks by generating multi-level functional summaries. We discuss a data-driven and generic PPI network summarization framework that constructs higher level functional summary to summarize the underlying PPI network to obtain a concise, interpretable representation of the network. It generates the "best" summary from both interaction and annotation data by maximizing information gain for a specific resolution. We evaluate the performance of this framework on several real-world PPI networks, its superiority over network clustering, and showcase its applicability in comprehending Alzheimer's disease network.

Third, we discuss a technique that summarizes a PPI network in a multi-perspective manner. This is based on the fact that a biological system can be seen from different functional perspectives (e.g., components in a PPI network can be organized by localization, process, disease, etc.). Each discovered perspective represents a distinct interpretation of how the network can be functionally summarized. The performance of this framework is extensively discussed with several real-world PPI networks highlighting the limitations of network clustering paradigm to generate such multi-perspective summary. We also performed a case study using human autophagy system to illustrate the utility of this framework.

Fourth, we discuss a data-driven effort to construct summaries of differential functional responses of gene interaction networks that undergo "rewiring" after environmental change. Experimental evaluation with real-world dataset demonstrates the superiority of this technique to address the differential network summarization problem.

The last topic consists of several open problems of this young field. The list presented should by no means be considered exhaustive and is centered around challenges and issues currently in vogue. Nevertheless, readers can benefit by exploring the research directions given in this part.

The book is suitable for use in advanced undergraduate- and graduate-level courses on biological networks. It has sufficient material that can be covered as part of a semester-long course, thereby leaving plenty of room for an instructor to choose topics. An undergraduate course in algorithms, graph theory, and basic cell biology should suffice as a prerequisite for most of the chapters. A good knowledge of C++/Java programming language is sufficient to code the algorithms described herein. For completeness, we have provided background information on several topics in Chap. 2: the central dogma of biology, protein–protein interactions, high-throughput experimental techniques to analyze protein–protein interactions,

and annotations of these interactions with Gene Ontology. The knowledgeable reader may omit this chapter and perhaps refer back to comparisons while reading later chapters of this book.

We hope that this book will serve as a catalyst in helping this burgeoning area of biological network summarization grow and have practical impact.

Singapore
December 2016

Sourav S. Bhowmick
Boon-Siew Seah

Acknowledgements

It is a great pleasure for us to acknowledge the assistance and contributions of a large number of individuals to this effort. First, we would like to thank our publisher Springer-Verlag for their support. In particular, we would like to acknowledge the efforts, help, and patience of Melissa Fearon and Jennifer Malat, our primary contacts for this edition.

The work reported in this book grew out of the PANORAMA project at the Nanyang Technological University (NTU), Singapore, and Massachusetts Institute of Technology (MIT), USA, under the auspices of Singapore-MIT Alliance graduate program. In this project, we explored issues on building frameworks that allow biologists to rapidly visualize the processes that govern biological systems. Specifically, the chapters in this book are part of Boon Siew's thesis work under the guidance of Sourav. Some of these chapters are published in reputable journals and conferences in the area of bioinformatics and systems biology.

Dr. C. Forbes Dewey, Jr. of MIT, who was a key collaborator for this project, deserves the first thank you. Not only did he introduce us to interesting and exciting field of network biology and network medicine, but he was also always willing to discuss ideas with us, no matter how strange they were.

In addition, we would also like to express our gratitude to all the group members and collaborators, past and present, in our **Co**mputational **S**ystems **B**iolog**y** research group (COSBY). In particular, Dr. Huey Eng Chua (NTU), Dr. Jie Zheng (NTU), Dr. Lisa Tucker-Kellogg (Duke-NUS Medical School), Dr. Hanry Yu (NUS), and Mengxuan Chen made substantial contributions to the broader aspect of our research in network biology.

Quite a few people have helped us with the initial vetting of the text for this book. It is our pleasure to acknowledge them all here. We would like to thank Scientific Publishing Services (SPS) for carefully proofreading the complete book in a short span of time and suggesting the changes which have been incorporated.

Sourav and Boon Siew would like to acknowledge their parents and family members who gave them incredible support throughout the years. A special thanks goes to Dr. Paul Matsudaira (MIT, NUS) and Dr. Hew Choy Leong (NUS), who

were a great motivator during our early days when we were grappling with the new field of systems biology. They were the major force behind our continuous strive for breaking out from the comfort zone of computer science to explore problems that are at the intersection of two or more disparate fields. It has been a great learning experience for us.

Finally, we would like to thank the Singapore-MIT Alliance for the generous resources and financial support provided for the PANORAMA project. We would also like to thank the School of Computer Science and Engineering at the Nanyang Technological University for allowing the use of their resources to help complete this book.

Singapore
December 2016

Sourav S. Bhowmick
Boon-Siew Seah

Contents