Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari
Chern Han Yong
Limsoon Wong

Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari, Chern Han Yong, Limsoon Wong

Complexes of physically interacting proteins constitute fundamental functional units that drive almost all biological processes within cells. A faithful reconstruction of the entire set of protein complexes (the "complexosome") is therefore important not only to understand the composition of complexes but also the higher level functional organization within cells. Advances over the last several years, particularly through the use of high-throughput proteomics techniques, have made it possible to map substantial fractions of protein interactions (the "interactomes") from model organisms including Arabidopsis thaliana (a flowering plant), Caenorhabditis elegans (a nematode), Drosophila melanogaster (fruit fly), and Saccharomyces cerevisiae (budding yeast). These interaction datasets have enabled systematic inquiry into the identification and study of protein complexes from organisms. Computational methods have played a significant role in this context, by contributing accurate, efficient, and exhaustive ways to analyze the enormous amounts of data. These methods have helped to compensate for some of the limitations in experimental datasets including the presence of biological and technical noise and the relative paucity of credible interactions.

In this book, we systematically walk through computational methods devised to date (approximately between 2000 and 2016) for identifying protein complexes from the network of protein interactions (the protein-protein interaction (PPI) network). We present a detailed taxonomy of these methods, and comprehensively evaluate them for protein complex identification across a variety of scenarios including the absence of many true interactions and the presence of false-positive interactions (noise) in PPI networks. Based on this evaluation, we highlight challenges faced by the methods, for instance in identifying sparse, sub-, or small complexes and in discerning overlapping complexes, and reveal how a combination of strategies is necessary to accurately reconstruct the entire complexosome.

ABOUT ACM BOOKS

ACM Books is a new series of high quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

BOOKS.ACM.ORG · WWW.MORGANCLAYPOOLPUBLISHERS.COM

ISBN 978-1-970001-55-6

90000

9 781970 001556

SRIHARI
YONG
WONG

COMPUTATIONAL PREDICTION OF PROTEIN COMPLEXES FROM PROTEIN INTERACTION NETWORKS

ACM | MORGAN & CLAYPOOL

Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari

The University of Queensland Institute for Molecular Bioscience

Chern Han Yong

Duke-National University of Singapore Medical School

Limsoon Wong

National University of Singapore

ACM Books #16

Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari, Chern Han Yong, Limsoon Wong

books.acm.org
www.morganclaypoolpublishers.com

ISBN: 978-1-97000-155-6 hardcover
ISBN: 978-1-97000-152-5 paperback
ISBN: 978-1-97000-153-2 ebook
ISBN: 978-1-97000-154-9 ePub

Series ISSN: 2374-6769 print 2374-6777 electronic

DOIs:

10.1145/3064650 Book
10.1145/3064650.3064651 Preface
10.1145/3064650.3064652 Chapter 1
10.1145/3064650.3064653 Chapter 2
10.1145/3064650.3064654 Chapter 3
10.1145/3064650.3064655 Chapter 4
10.1145/3064650.3064656 Chapter 5
10.1145/3064650.3064657 Chapter 6
10.1145/3064650.3064658 Chapter 7
10.1145/3064650.3064659 Chapter 8
10.1145/3064650.3064660 Chapter 9
10.1145/3064650.3064661 References, Bios

A publication in the ACM Books series, #16
Editor in Chief: M. Tamer Özsu, *University of Waterloo*

First Edition

10 9 8 7 6 5 4 3 2 1

Computational Prediction of Protein Complexes from Protein Interaction Networks

ACM Books

Editor in Chief

M. Tamer Özsu, *University of Waterloo*

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari, *The University of Queensland Institute for Molecular Bioscience*
Chern Han Yong, *Duke-National University of Singapore Medical School*
Limsoon Wong, *National University of Singapore*
2017

Shared-Memory Parallelism Can Be Simple, Fast, and Scalable

Julian Shun, *University of California, Berkeley*
2017

The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations

Editors: Sharon Oviatt, *Incaa Designs*
Björn Schuller, *University of Passau and Imperial College London*
Philip R. Cohen, *Voicebox Technologies*
Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*
Gerasimos Potamianos, *University of Thessaly*
Antonio Krüger, *German Research Center for Artificial Intelligence (DFKI)*
2017

Communities of Computing: Computer Science and Society in the ACM

Thomas J. Misa, Editor, *University of Minnesota*
2017

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining

ChengXiang Zhai, *University of Illinois at Urbana–Champaign*
Sean Massung, *University of Illinois at Urbana–Champaign*
2016

An Architecture for Fast and General Data Processing on Large Clusters

Matei Zaharia, *Massachusetts Institute of Technology*
2016

Reactive Internet Programming: State Chart XML in Action
Franck Barbier, *University of Pau, France*
2016

Verified Functional Programming in Agda
Aaron Stump, *The University of Iowa*
2016

The VR Book: Human-Centered Design for Virtual Reality
Jason Jerald, *NextGen Interactions*
2016

Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age
Robin Hammerman, *Stevens Institute of Technology*
Andrew L. Russell, *Stevens Institute of Technology*
2016

Edmund Berkeley and the Social Responsibility of Computer Professionals
Bernadette Longo, *New Jersey Institute of Technology*
2015

Candidate Multilinear Maps
Sanjam Garg, *University of California, Berkeley*
2015

Smarter Than Their Machines: Oral Histories of Pioneers in Interactive Computing
John Cullinane, *Northeastern University; Mossavar-Rahmani Center for Business and Government, John F. Kennedy School of Government, Harvard University*
2015

A Framework for Scientific Discovery through Video Games
Seth Cooper, *University of Washington*
2014

Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers
Bryan Jeffrey Parno, *Microsoft Research*
2014

Embracing Interference in Wireless Systems
Shyamnath Gollakota, *University of Washington*
2014

Dedicated to the Honors, Masters, and Ph.D. students who worked over the years on the different aspects of PPI networks by being part of the computational biology group at the Department of Computer Science, National University of Singapore.

Preface

The suggestion and motivation to write this book came from Limsoon, who thought that it would be a great idea to compile our (Sriganesh's and Chern Han's) Ph.D. research conducted at National University of Singapore on protein complex prediction from protein-protein interaction (PPI) networks into a comprehensive book for the research community. Since we (Sriganesh and Chern Han) completed our Ph.D.s not long ago, the timing could not have been better for writing this book while the topic is still fresh in our minds and the empirical set up (datasets and software pipelines) for evaluating the methods is still in a "quick-to-run" form. However, although we had our Ph.D. theses to our convenient disposal and reference, it is only after we started writing this book that we realized the real scale of the task that we had embarked upon.

The problem of protein complex prediction may be just one of the plethora of computational problems that have opened up since the deluge of proteomics (protein-protein interaction; PPI) data over the last several years. However, in reality this problem encompasses or directly relates to several important and open problems in the area—in particular, the fundamental problems of modeling, visualizing, and denoising of PPI networks, prediction of PPIs (novel as well as evolutionarily conserved), and protein function prediction from PPI data. Therefore, to write a comprehensive self-contained book, we had to cover even these closely related problems to some extent or at least allude to or reference them in the book. We had to do so without missing the connection between these problems and our central problem of protein complex prediction in the book.

The early tone to write the book in this manner was set by our review article in a 2015 special issue of *FEBS Letters*, where we covered a number of protein complex prediction methods which are based on a diverse range of topological, functional, temporal, structural, and evolutionary information. However, being only a single-volume article, the description of the methods was brief, and to compile

this description in the form of a book we had to delve a lot deeper into the algorithmic underpinnings of each of the methods, highlight how each method utilized the information (topological, functional, temporal, structural, and evolutionary) on which it was based in its own unique way, and evaluate and study the applications of the methods across a diverse range of datasets and scenarios. To do this well, we had to: (i) cover in substantial detail the preliminaries such as the experimental techniques available to infer PPIs, the limitations of each of these techniques, PPI network topology, modeling, and denoising, PPI databases that are available, and how functional, temporal, structural, and evolutionary information of proteins can be integrated with PPI networks; and (ii) we had to categorize protein complex prediction methods into logical groups based on some criteria, and dedicate a separate chapter for each group to make our description comprehensive. In the book, we cover (i) in Chapter 2 and in the form of independent sections within each of the other Chapters 3, 5, 6, 7, and 8. We cover (ii) by allocating Chapters 3 and 4 for "classical" methods and their comprehensive evaluation, Chapter 5 for methods that predict certain kinds of "challenging complexes" which the classical methods do not predict well, Chapter 6 for methods that utilize temporal and structural information, Chapter 7 for methods that utilize information on evolutionary conservation, and Chapter 8 for methods that integrate other kinds of omics datasets to predict "specialized" complexes—e.g., protein complexes in diseases.

The requirement for a book exclusively dedicated to the problem of protein complex prediction from PPI networks at this point in time cannot be understated. Over the last two decades, a major focus of high-throughput experimental technologies and of computational methods to analyze the generated data has been in genomics—e.g., in the analysis of genome sequencing data. It is relatively recently that this focus has started to shift toward proteomics and computational methods to analyze proteomics data. For example, while the complete sequence of the human genome was assembled more than a decade ago, it is only over the last three years that there have been similar large-scale efforts to map the human proteome. The ProteomicsDB (http://www.proteomicsdb.org/), Human Proteome Map (http://humanproteomemap.org/), and the Human Protein Atlas projects (http://www.proteinatlas.org/) have all appeared only over the last three years. Similarly is The Cancer Proteome Atlas (TCPA) project (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html) which complements The Cancer Genome Atlas project (TCGA). This means that developing more effective solutions for fundamental problems such as protein complex prediction has become all the more important today, as we try to apply these solutions to larger and more complex datasets arising from these newer technologies and projects. In this respect,

we had to write the book not just by considering protein complex prediction methods (i.e., their algorithmic details) as important in their own right but also by giving significant importance to the applications of these methods in the light of today's complex datasets and research questions. There are several sections within each of the Chapters 6, 7, and 8 that play this dual role, e.g., a section in Chapter 7 discusses the evolutionary conservation of core cellular processes based on conservation patterns of protein complexes, and a section in Chapter 8 discusses the dysregulation of these processes in diseases based on rewiring of protein complexes between normal and disease conditions.

In the end, we hope that we have done justice to what we intend this book to be. We hope that this book provides valuable insights into protein complex prediction and inspires further research in the area especially for tackling the open challenges, as well as inspires new applications in diverse areas of biomedicine.

Acknowledgments

Although this book is primarily concerned with the problem of protein complex prediction, the book also covers several other aspects of PPI networks. We would like to therefore dedicate this book to the students—Honors, Masters, and Ph.D. students—who worked on these different aspects of PPI networks by being part of the computational biology group at the Department of Computer Science, National University of Singapore, over the years. Several of the methods covered in this book are a result of the extensive research conducted by these students. Sriganesh would like to thank Hon Wai Leong (Professor of Computer Science, National University of Singapore) under whom he conducted his Ph.D. research on protein complex prediction; Mark Ragan (Head of Division of Genomics of Development and Disease at Institute for Molecular Bioscience, The University of Queensland) under whom he conducted his postdoctoral research, a substantial portion of which was on identifying protein complexes in diseases; and Kum Kum Khanna (Senior Principal Research Fellow and Group Leader at QIMR-Berghofer Medical Research Institute) whose guidance played a significant part in his understanding of biological aspects of protein complexes. Sriganesh is grateful to Mark for passing him an original copy of a 1977 volume of *Progress in Biophysics and Molecular Biology* in which G. Rickey Welch makes a consistent principled argument that "multienzyme clusters" are advantageous to the cell and organism because they enable metabolites to be channeled within the clusters and protein expression to be co-regulated [Welch 1977]—a possession which Sriganesh will deeply cherish. Chern Han would like to thank his coauthors: Sriganesh for doing the heavy lifting in writing, editing, and

driving this project and Limsoon Wong for guiding him through his Ph.D. journey on protein complexes. He would also like to acknowledge the support of Bin Tean Teh (Professor with Program in Cancer and Stem Cell Biology, Duke-NUS Medical School), who currently oversees his postdoctoral research. Limsoon would like to acknowledge Chern Han and Sriganesh for doing the bulk of the writing for this book, and especially thank Sriganesh for taking the overall lead on the project. When he suggested the book to Chern Han and Sriganesh, he had not imagined that he would eventually be a co-author.

We are indebted also to the Editor-in-Chief of ACM Books, Tamer Özsu, Executive Editor Diane Cerra, Production Manager Paul C. Anagnostopoulos, and the entire team at ACM Books and Morgan & Claypool Publishers for their encouragement and for producing this book so beautifully.

Sriganesh Srihari
Chern Han Yong
Limsoon Wong
May 2017

Contents