

Table of Contents

Preface..... xix

Acknowledgment..... xxvii

Section I
Introduction

Chapter I

Molecular Biology of Protein-Protein Interactions for Computer Scientists..... 1

Christian Schönbach, Nanyang Technological University, Singapore

Chapter II

Data Mining for Biologists 14

Koji Tsuda, Max Planck Institute for Biological Cybernetics, Germany

Section II
PPI Network Construction and Cleansing

Chapter III

Domain-Based Prediction and Analysis of Protein-Protein Interactions..... 29

Tatsuya Akutsu, Kyoto University, Japan

Morihiro Hayashida, Kyoto University, Japan

Chapter IV

Incorporating Graph Features for Predicting Protein-Protein Interactions..... 45

Martin S. R. Paradesi, Kansas State University, USA

Doina Caragea, Kansas State University, USA

William H. Hsu, Kansas State University, USA

Chapter V

Discovering Protein-Protein Interaction Sites from Sequence and Structure.....	64
--	----

David La, Dept. of Biological Sciences, Purdue University, USA

*Daisuke Kihara, Dept. of Biological Sciences, Dept. of Computer Science,
Purdue University, USA*

Chapter VI

Network Cleansing: Reliable Interaction Networks	80
--	----

Paolo Marcatili, Sapienza University, Italy

Anna Tramontano, Sapienza University, Istituto Pasteur Fondazione Cenci Bolognetti, Italy

Section III

Knowledge Discovery from PPI Networks

Chapter VII

Discovering Interaction Motifs from Protein Interaction Networks.....	99
---	----

Hugo Willy, National University of Singapore, Singapore

Chapter VIII

Discovering Network Motifs in Protein Interaction Networks.....	117
---	-----

Raymond Wan, Kyoto University, Japan

Hiroshi Mamitsuka, Kyoto University, Japan

Chapter IX

Discovering Protein Complexes in Protein Interaction Networks.....	144
--	-----

Clara Pizzuti, ICAR, Consiglio Nazionale delle Ricerche, Italy

Simona Ester Rombo, Università della Calabria, Italy

Chapter X

Evolutionary Analyses of Protein Interaction Networks	169
---	-----

Takashi Makino, University of Dublin, Trinity College, Ireland

Aoife McLysaght, University of Dublin, Trinity College, Ireland

Section IV

Biological Applications Using PPI Analysis

Chapter XI

Discovering Lethal Proteins in Protein Interaction Networks.....	183
--	-----

Kar Leong Tew, Institute for Infocomm Research, Singapore

Xiao-Li Li, Institute for Infocomm Research, Singapore

Chapter XII	
Predicting Protein Functions from Protein Interaction Networks.....	203
<i>Hon Nian Chua, Institute for Infocomm Research, Singapore</i>	
<i>Limsoon Wong, National University of Singapore, Singapore</i>	

Chapter XIII	
Protein Interactions for Functional Genomics.....	223
<i>Pablo Minguez, Centro de Investigación Príncipe Felipe (CIPF), Spain</i>	
<i>Joaquin Dopazo, Centro de Investigación Príncipe Felipe (CIPF), Spain</i>	

Chapter XIV	
Prioritizing Disease Genes and Understanding Disease Pathways.....	239
<i>Xiaoyue Zhao, Bionovo Inc., USA</i>	
<i>Lilia M. Iakoucheva, Rockefeller University, USA</i>	
<i>Michael Q. Zhang, Cold Spring Harbor Laboratory, USA</i>	

Chapter XV	
Dynamics of Protein-Protein Interaction Network in Plasmodium Falciparum.....	257
<i>Smita Mohanty, Indian Institute of Science, India</i>	
<i>Shashi Bhushan Pandit, Georgia Institute of Technology, USA</i>	
<i>Narayanaswamy Srinivasan, Indian Institute of Science, India</i>	

Section V

Tools for Analysis of PPI Networks

Chapter XVI	
Graphical Analysis and Visualization Tools for Protein Interaction Networks	286
<i>Sirisha Gollapudi, MyCIB, University of Nottingham, UK</i>	
<i>Alex Marshall, MyCIB, University of Nottingham, UK</i>	
<i>Daniel Zadik, MyCIB, University of Nottingham, UK</i>	
<i>Charlie Hodgman, MyCIB, University of Nottingham, UK</i>	

Chapter XVII	
Network Querying Techniques for PPI Network Comparison	312
<i>Valeria Fionda, Università della Calabria, Italy</i>	
<i>Luigi Palopoli, Università della Calabria, Italy</i>	

Chapter XVIII

Module Finding Approaches for Protein Interaction Networks.....	335
---	-----

Tero Aittokallio, University of Turku, Finland

Compilation of References	354
--	------------

About the Contributors	401
-------------------------------------	------------

Index.....	408
-------------------	------------

Detailed Table of Contents

Preface..... xix

Acknowledgment..... xxvii

Section I
Introduction

In Section I of the book, we will provide the two introductory chapters. Chapter I aims to provide a biological primer on protein-protein interactions for those readers who are computer scientists and who may not have adequate knowledge in biology. Chapter II is a primer on the key concepts and pattern mining algorithms in data mining useful for the readers who are biologists.

Chapter I
Molecular Biology of Protein-Protein Interactions for Computer Scientists..... 1
Christian Schönbach, Nanyang Technological University, Singapore

Advances in protein-protein interaction (PPI) detection technology and computational analysis methods have produced numerous PPI networks, whose completeness appears to depend on the extent of data derived from different PPI assay methods and the complexity of the studied organism. Despite the partial nature of human PPI networks, computational data integration and analyses helped to elucidate new interactions and disease pathways. The success of computational analyses considerably depends on PPI data understanding. Exploration of the data and verification of their quality requires basic knowledge of the molecular biology of PPIs and familiarity with the assay methods used to detect PPIs. Both topics are reviewed in this chapter. After introducing various types of PPIs the principles of selected PPI assays are explained and their limitations discussed. Case studies of the Wnt signaling pathway and splice regulation demonstrate some of the challenges and opportunities that arise from assaying and analyzing PPIs. The chapter is concluded with an extrapolation to human systems biology that offers a glimpse into the future of PPI networks.

Chapter II
Data Mining for Biologists..... 14
Koji Tsuda, Max Planck Institute for Biological Cybernetics, Germany

In this tutorial chapter, the author reviews basics about frequent pattern mining algorithms, including itemset mining, association rule mining, and graph mining. These algorithms can find frequently appearing substructures in discrete data. They can discover structural motifs, for example, from mutation data, protein structures, and chemical compounds. As they have been primarily used for business data, biological applications are not so common yet, but their potential impact would be large. Recent advances in computers including multicore machines and ever increasing memory capacity support the application of such methods to larger datasets. The author explains technical aspects of the algorithms, but do not go into details. Current biological applications are summarized and possible future directions are given.

Section II PPI Network Construction and Cleansing

We will describe the methods for constructing and cleansing PPI networks. Section II of this book will describe key bioinformatics methods for predicting/validating protein interactions, as well as network cleansing techniques to detect reliable protein interaction networks.

Chapter III

Domain-Based Prediction and Analysis of Protein-Protein Interactions..... 29

Tatsuya Akutsu, Kyoto University, Japan

Morihiro Hayashida, Kyoto University, Japan

Many methods have been proposed for inference of protein-protein interactions from protein sequence data. This chapter focuses on methods based on domain-domain interactions, where a domain is defined as a region within a protein that either performs a specific function or constitutes a stable structural unit. In these methods, the probabilities of domain-domain interactions are inferred from known protein-protein interaction data and protein domain data, and then prediction of interactions is performed based on these probabilities and contents of domains of given proteins. This chapter overviews several fundamental methods, which include association method, expectation maximization-based method, support vector machine-based method, and linear programming-based method. This chapter also reviews a simple evolutionary model of protein domains, which yields a scale-free distribution of protein domains. By combining with a domain-based protein interaction model, a scale-free distribution of protein-protein interaction networks is also derived.

Chapter IV

Incorporating Graph Features for Predicting Protein-Protein Interactions..... 45

Martin S. R. Paradesi, Kansas State University, USA

Doina Caragea, Kansas State University, USA

William H. Hsu, Kansas State University, USA

This chapter presents applications of machine learning to predicting protein-protein interactions (PPI) in *Saccharomyces cerevisiae*. Several supervised inductive learning methods have been developed that treat this task as a classification problem over candidate links in a PPI network – a graph whose nodes represent proteins and whose arcs represent interactions. Most such methods use feature extraction from

protein sequences (e.g., amino acid composition) or associated with protein sequences directly (e.g., GO annotation). Others use relational and structural features extracted from the PPI network, along with the features related to the protein sequence. Topological features of nodes and node pairs can be extracted directly from the underlying graph. This chapter presents two approaches from the literature (Qi et al., 2006; Licamele & Getoor, 2006) that construct features on the basis of background knowledge, an approach that extracts purely topological graph features (Paradesi et al., 2007), and one that combines knowledge-based and topological features (Paradesi, 2008). Specific graph features that help in predicting protein interactions are reviewed. This study uses two previously published datasets (Chen & Liu, 2005; Qi et al., 2006) and a third dataset (Paradesi, 2008) that was created by combining and augmenting three existing PPI databases. The chapter includes a comparative study of the impact of each type of feature (topological, protein sequence-based, etc.) on the sensitivity and specificity of classifiers trained using specific types of features. The results indicate gains in the area under the sensitivity-specificity curve for certain algorithms when topological graph features are combined with other biological features such as protein sequence-based features.

Chapter V

Discovering Protein-Protein Interaction Sites from Sequence and Structure..... 64
David La, *Dept. of Biological Sciences, Purdue University, USA*
Daisuke Kihara, *Dept. of Biological Sciences, Dept. of Computer Science,
Purdue University, USA*

This chapter gives a comprehensive introduction of the sequence/structural features that are characteristic of protein-protein interaction sites and reviews state-of-the-art methodologies for protein-protein binding site prediction. Protein-protein interaction residues are largely responsible for mediating physical binding processes such as inhibitory effects through enzyme-inhibitor interaction, initiating immune response by an antibody-antigen interaction, and regulation of cell signaling proteins. Currently, various methods are available for predicting protein-protein interaction sites, which allow a residue-level understanding of the physical protein binding phenomena presented by the global construction protein-protein interaction networks. The overview of the discussed protein-protein binding site prediction strategies and detailed comparison of their weaknesses and strengths is aimed towards assisting protein researchers in gaining more insight to protein-protein interaction networks.

Chapter VI

Network Cleansing: Reliable Interaction Networks..... 80
Paolo Marcatili, *Sapienza University, Italy*
Anna Tramontano, *Sapienza University, Istituto Pasteur Fondazione Cenci Bolognetti, Italy*

This chapter provides an overview of the current computational methods for PPI network cleansing. The authors first present the issue of identifying reliable PPIs from noisy and incomplete experimental data. Next, they address the questions of which are the expected results of the different experimental studies, of what can be defined as true interactions, of which kind of data are to be integrated in assigning reliability levels to PPIs and which gold standard should the authors use in training and testing PPI filtering methods. Finally, Marcatili and Tramontano describe the state of the art in the field, presenting the different classes of algorithms and comparing their results. The aim of the chapter is to guide the

reader in the choice of the most convenient methods, experiments and integrative data and to underline the most common biases and errors to obtain a portrait of PINs which is not only reliable but as well able to correctly retrieve the biological information contained in such data.

Section III

Knowledge Discovery from PPI Networks

Section III of the book will focus on knowledge discovery – that is, how do we computationally interrogate the PPI networks to discover new useful biological knowledge such as interaction motifs, network motifs, and protein complexes using sophisticated graph data mining approaches previously inaccessible to the bench biologists. We will also describe the evolutionary analyses of protein interaction networks for gaining insights on molecular evolution and comparative genomics.

Chapter VII

Discovering Interaction Motifs from Protein Interaction Networks..... 99

Hugo Willy, National University of Singapore, Singapore

Recent breakthroughs in high throughput experiments to determine protein-protein interaction have generated a vast amount of protein interaction data. However, most of the experiments could only answer the question of whether two proteins interact but not the question on the mechanisms by which proteins interact. Such understanding is crucial for understanding the protein interaction of an organism as a whole (the interactome) and even predicting novel protein interactions. Protein interaction usually occurs at some specific sites on the proteins and, given their importance, they are usually well conserved throughout the evolution of the proteins of the same family. Based on this observation, a number of works on finding protein patterns/motifs conserved in interacting proteins have emerged in the last few years. Such motifs are collectively termed as the interaction motifs. This chapter provides a review on the different approaches on finding interaction motifs with a discussion on their implications, potentials and possible areas of improvements in the future.

Chapter VIII

Discovering Network Motifs in Protein Interaction Networks..... 117

Raymond Wan, Kyoto University, Japan

Hiroshi Mamitsuka, Kyoto University, Japan

This chapter examines some of the available techniques for analyzing a protein interaction network (PIN) when depicted as an undirected graph. Within this graph, algorithms have been developed which identify “notable” smaller building blocks called network motifs. The authors examine these algorithms by dividing them into two broad categories based on two definitions of “notable”: (a) statistically-based methods and (b) frequency-based methods. They describe how these two classes of algorithms differ not only in terms of efficiency, but also in terms of the type of results that they report. Some publicly-available programs are demonstrated as part of their comparison. While most of the techniques are generic

and were originally proposed for other types of networks, the focus of this chapter is on the application of these methods and software tools to PINs.

Chapter IX

Discovering Protein Complexes in Protein Interaction Networks..... 144

Clara Pizzuti, ICAR, Consiglio Nazionale delle Ricerche, Italy

Simona Ester Rombo, Università della Calabria, Italy

In this chapter a survey on the main graph-based clustering techniques proposed in the literature to mine protein-protein interaction networks (PINs) is presented. The detection of putative protein complexes is an important research problem in systems biology. In fact it may help in understanding the mechanisms regulating cell life, in deriving conservations across species, in predicting the biological functions of uncharacterized proteins, and, more importantly, for therapeutic purposes. Different kind of approaches are described and classified. Furthermore, some validation techniques commonly exploited in this context are illustrated. The goal of the chapter is to provide a useful guide and reference for both computer scientists and biologists. Computer scientists may have a complete vision of what has already been made and which are the new challenges about PINs clustering, taking them as a starting point for further researches and new proposals; on the other hand, biologists may find in the chapter the necessary material to select the most appropriate methods to apply for their specific purposes.

Chapter X

Evolutionary Analyses of Protein Interaction Networks 169

Takashi Makino, University of Dublin, Trinity College, Ireland

Aoife McLysaght, University of Dublin, Trinity College, Ireland

This chapter introduces evolutionary analyses of protein interaction networks and of proteins as components of the networks. The authors show relationships between proteins in the networks and their evolutionary rates. For understanding protein-protein interaction (PPI) divergence, duplicated genes are often compared because they are derived from a common ancestral gene. In order to reveal evolutionary mechanisms acting on the interactome it is necessary to compare PPIs across species. Investigation of co-localization of interacting genes in a genome shows that PPIs have an important role in the maintenance of a physical link between neighboring genes. The purpose of this chapter is to introduce methodologies for analyzing PPI data and to describe molecular evolution and comparative genomics insights gained from such studies.

Section IV

Biological Applications Using PPI Analysis

Section IV of the book will present the various potential biological applications of protein interaction networks, e.g. the identification of lethal proteins, the functional annotation of unknown proteins, and the enrichment of functional genomics analyses. We will discuss how to use the protein interaction network and other biological resources for drug discovery, how to uncover the underlying disease pathways, and how to exploit the PPI networks to understand the complex dynamics of cellular processes.

Chapter XI

Discovering Lethal Proteins in Protein Interaction Networks.....	183
--	-----

Kar Leong Tew, Institute for Infocomm Research, Singapore

Xiao-Li Li, Institute for Infocomm Research, Singapore

This chapter introduces state-of-the-art computational methods which discover lethal proteins from Protein Interaction Networks (PINs). Lethal proteins are an interesting subject in understanding the minimal condition for cellular development and survival. A dysfunctional research subject or absence of a lethal protein would result in fatality of the cell. Biological experiments have been conducted to systematically detect such proteins. However, such processes are time consuming and requires huge amount of effort to conduct. The researchers have developed a series of computational methods which take advantage of the network properties of individual proteins to detect lethal proteins in PINs. In this chapter, each computational method is studied in depth with an analysis on its pros and cons. Finally, a discussion on the possible further research directions will conclude the chapter.

Chapter XII

Predicting Protein Functions from Protein Interaction Networks.....	203
---	-----

Hon Nian Chua, Institute for Infocomm Research, Singapore

Limsoon Wong, National University of Singapore, Singapore

Functional characterization of genes and their protein products is essential to biological and clinical research. Yet, there is still no reliable way of assigning functional annotations to proteins in a high-throughput manner. In this chapter, the authors provide an introduction to the task of automated protein function prediction. They discuss about the motivation for automated protein function prediction, the challenges faced in this task, as well as some approaches that are currently available. In particular, they take a closer look at methods that use protein-protein interaction for protein function prediction, elaborating on their underlying techniques and assumptions, as well as their strengths and limitations.

Chapter XIII

Protein Interactions for Functional Genomics.....	223
---	-----

Pablo Minguez, Centro de Investigación Príncipe Felipe (CIPF), Spain

Joaquin Dopazo, Centro de Investigación Príncipe Felipe (CIPF), Spain

Here the authors review the state of the art in the use of protein-protein interactions (ppis) within the context of the interpretation of genomic experiments. They report the available resources and methodologies used to create a curated compilation of ppis introducing a novel approach to filter interactions. Special attention is paid in the complexity of the topology of the networks formed by proteins (nodes) and pairwise interactions (edges). These networks can be studied using graph theory and a brief introduction to the characterization of biological networks and definitions of the more used network parameters is also given. Also a report on the available resources to perform different modes of functional profiling using ppi data is provided along with a discussion on the approaches that have typically been applied into this context. They also introduce a novel methodology for the evaluation of networks and some examples of its application.

Chapter XIV

Prioritizing Disease Genes and Understanding Disease Pathways.....	239
--	-----

Xiaoyue Zhao, Bionovo Inc., USA

Lilia M. Iakoucheva, Rockefeller University, USA

Michael Q. Zhang, Cold Spring Harbor Laboratory, USA

Genetic factors play a major role in the etiology of many human diseases. Genome-wide experimental methods produce an increasing number of genes associated with such diseases. This chapter introduces data sources, bioinformatics tools, and computational methods for prioritizing disease candidate genes and identifying disease pathways. The main strategy is to examine the similarity among the candidate genes and known disease genes at the functional level. The authors review different similarity measures and prevailing methods for integrating results from different functional aspects. They hope this chapter will help advocate many useful resources that the researchers can use to investigate diseases of their interest.

Chapter XV

Dynamics of Protein-Protein Interaction Network in Plasmodium Falciparum.....	257
---	-----

Smita Mohanty, Indian Institute of Science, India

Shashi Bhushan Pandit, Georgia Institute of Technology, USA

Narayanaswamy Srinivasan, Indian Institute of Science, India

Integration of organism-wide protein interactome data with information on expression of genes, cellular localization of proteins and their functions has proved extremely useful in developing biologically intuitive interaction networks. This chapter highlights the dynamics in protein interaction network across different stages in the lifecycle of Plasmodium falciparum, a malarial parasite, and the implication of the network dynamics in different physiological processes. The main focus of the chapter is the integration of information on experimentally derived interactions of P.falciparum proteins with expression data and analysis of the implications of interactions in different cellular processes. Extensive analysis has been made to quantify the interaction dynamics across various stages, as well as correlating it with the dynamics of the cellular pathways involving the interacting proteins. The authors' analysis demonstrates the power of strategic integration of genome-wide datasets in extracting information on dynamics of biological pathways and processes.

Section V

Tools for Analysis of PPI Networks

We will conclude the book by three chapters that discuss state-of-the-art software tools that provide visualization and analysis for protein interaction networks.

Chapter XVI

Graphical Analysis and Visualization Tools for Protein Interaction Networks	286
---	-----

Sirisha Gollapudi, MyCIB, University of Nottingham, UK

Alex Marshall, MyCIB, University of Nottingham, UK

Daniel Zadik, MyCIB, University of Nottingham, UK

Charlie Hodgman, MyCIB, University of Nottingham, UK

Software for the visualisation and analysis of protein-protein interaction (PPI) networks can enable general exploration, as well as providing graph-theoretic algorithms for specific tasks. Analyses can include reduction of complexity or the scope of the network in order to make it more manageable, or increase in complexity by integration with other datasets, to represent biology more accurately. Two software approaches are outlined in this chapter: desktop applications and web services. Desktop applications have attractive user interfaces with a wide range of analysis tools, and often capabilities for integration of other bio-molecular data. Web services provide a newer approach to network analysis. They have the advantages of a broader range of potential functionalities and a more extensible framework than standalone desktop tools. However, their relative infancy means that they are not as well developed. This chapter provides an evaluation of some common desktop applications, compared to and contrasted with several examples of web services.

Chapter XVII

Network Querying Techniques for PPI Network Comparison	312
--	-----

Valeria Fionda, Università della Calabria, Italy

Luigi Palopoli, Università della Calabria, Italy

The aim of this chapter is that of analyzing and comparing network querying techniques as applied to protein interaction networks. In the last few years, several automatic tools supporting knowledge discovery from available biological interaction data have been developed. In particular, network querying tools search a whole biological network to identify conserved occurrences of a query network module. The goal of such techniques is that of transferring biological knowledge. Indeed, the query subnetwork generally encodes a well-characterized functional module, and its occurrences in the queried network probably denote that this function is featured by the associated organism. The proposed analysis is intended to be useful to understand problems and research issues, state of the art and opportunities for researchers working in this research area.

Chapter XVIII

Module Finding Approaches for Protein Interaction Networks.....	335
---	-----

Tero Aittokallio, University of Turku, Finland

This chapter provides an overview of the computational approaches developed for exploring the modular organization of protein interaction networks. A special emphasis is placed on the module finding tools implemented in three freely available software packages, VisANT, Cytoscape and MATISSE, as well as on their biomedical applications. The selected methods are presented in the broader context of module discovery options, ranging from approaches that rely merely on topological properties of the underlying network to those that take into account also other complementary data sources, such as the mRNA levels of the proteins. The author will also highlight some current limitations in the measured network data that

should be understood when developing and applying module finding methodology, and discuss some key future trends and promising research directions with potential implications for clinical research.

Compilation of References 354

About the Contributors 401

Index..... 408