Preface

The time has come for modern biology to move from the study of single molecules to networked systems. The recent technological advances in genomics and proteomics have resulted in the identification of many genes and proteins in various living organisms. However, simply knowing the existence of the genes and proteins in organisms does not provide insights about the molecular functions and biological processes in which they participate. Biologically, the functions of the bio-molecules and their participation in various cellular processes are mediated by networks of protein interactions and gene regulation. The new challenge for bioinformatics is to interrogate the biological data from the network perspectives, using new network-based data mining methodologies to uncover useful biological knowledge.

Recently, high-throughput methods (e.g. yeast-two-hybrid) for detecting protein-protein interactions (PPIs) *en masse* have enabled the construction of PPI networks on a genomic scale. A graphical map of an organism's interactome can be constructed from such experiments by considering individual proteins as the nodes, and the existence of a physical interaction between a pair of proteins as a link between two corresponding nodes. The growing flood of molecular interaction data, as biologists step up their efforts in mapping the interactomes of various species, is expected to surpass the data flux resulted from the development of genome sequencing in the past decade.

Such developments offer unprecedented opportunities to develop new bioinformatics methods for data mining the PPI networks. The study of protein interaction networks will provide invaluable insights about the inner working mechanisms of cells, which can lead to the uncovering of the underlying disease pathways and even the discovery of new drugs to benefit human beings. The subject area is therefore of interest to both biologists and computer scientists — to the biologists, the unraveling of PPI networks can unlock the many secrets of life; to the computer scientists, the PPI networks, being large-scale real-world graphical data, provide the ideal computational challenge for data mining research.

The objective of this book is to disseminate the research results and best practice from cross-disciplinary researchers and practitioners interested in, and working on bioinformatics, data mining, and proteomics. We aim to present the various methodologies in an accessible way, and we hope to bring better awareness of this interesting and challenging problem to inspire new computational solutions. The book is aimed at three overlapping audiences: (1) Researchers in the areas of bioinformatics, data mining, machine learning and data structure; (2) practitioners in the industry in these areas; (3) instructors and postgraduate students in colleges and universities. Upon reading the book, we hope that the molecular biologists would have acquired the necessary information to select the most appropriate methods and useful tools to apply, while the computer scientists would have a complete view of what have already been done and what are the new computational challenges that can serve as impactful starting points for furthering their data mining research.

Two main principles have been used to guide the writing of the book. Firstly, the chapters in the book are written in a comprehensive manner instead of focusing solely on a specific work. Each of

the chapters serves as a tutorial for network-based computational analysis of PPI networks. This book hopes to provide the reader a good understanding on how to mine the PPI networks by using effective computational algorithms as well as insights into the associated research challenges in data mining the PPI networks.

Secondly, the chapters are written with a balance of theory and practice to meet the needs of different categories of readers. We aim to present the various methodologies in an understandable way to our inter-disciplinary audience. Theoretical descriptions of the computational challenges and methods are provided in details for the computationally inclined readers. For the more practical minded readers, each chapter discusses and compares the experimental results and scientific findings obtained by using the methods. Where possible, each chapter lists and briefly describes online tools and database resources related to the topic at hand. Each chapter also provides a section of discussion on future trends from which researchers and postgraduate students can shape their research topics for further study.

As far as we know, this is the first book that is focused primarily on applying sophisticated data mining and graph mining techniques on protein interaction networks. We hope this book will be a useful resource for professionals and researchers who wish to learn how to apply advanced data mining techniques in protein interaction network. In addition, it can be used as a supplementary text book for Masters and PhD students studying bioinformatics.

This book is focused on presenting the current bioinformatics methods to interrogate the PPI networks. It is important to bear in mind that the PPI network is only one of the many components in the complex machinery of life. To fully understand the interplay of bio-molecules in carrying out critical life processes at the network, it is necessary to integrate and analyze other biological data.

The ultimate success of the PPI network analysis will depend on the parallel improvements both in the biological experimental techniques from the biologists which provide rich biological datasets for data mining community, as well as in the data mining techniques from the computer scientists which provide efficient ways to exploit the protein interaction data to help biologists better understand the life processes. We hope this book will play a role in helping to realize the wonderful magic of what a perfect marriage of biology with computer science can bring.

THE CHALLENGES

It is a well-known fact that protein interactions play a central role at virtually every level of cellular activities. Analysis of the PPI networks is necessary to decipher the underlying cellular mechanisms as well as the behavior of various biological systems. We will discuss the computational challenges for mining PPI networks.

Noisy PPI Data: Alarmingly High False Positive and False Negative Rates

While high-throughput methods such as yeast-two-hybrid (Y2H) (Fields & Song, 1989) and tandem affinity purification-mass spectrometry (TAP-MS) (Puig et al., 2001; Rigaut et al., 1999) have enabled comprehensive detection of protein interactions, the quality of detected protein interactions is far from satisfactory. On the one hand, the experimental conditions in which the detection methods are carried out may cause a bias towards detecting interactions that do not occur under physiological conditions, resulting in false positive detection rates that could be alarmingly high. In other words, the experimental data may not be very accurate and contain interaction data that do not occur in the cell. On the other hand, the high-throughput methods can also fail to detect various types of interactions, for example,

loss of weak transient interactions, loss of post-translational modification, and bias against soluble or membrane proteins (Lalonde et al., 2008; Tarassov et al., 2008). This results in false negative detection and low experimental coverage of the interactomes.

How serious is the situation — specifically, what are the false positive and false negative rates in current protein interaction data? Many researchers have attempted to answer this question recently (G. D. Bader & Hogue, 2002; J. S. Bader, Chaudhuri, Rothberg, & Chant, 2004; Gentleman & Huber, 2007; Hart, Ramani, & Marcotte, 2006; Von Mering et al., 2002), by comparing the overlapping between the collected protein interactions from different large-scale biological experiments, or checking the consistency of the function or location information between the interaction partners to estimate the accuracy and coverage of the PPI data. The results of all these research consistently showed that the quality of the protein interaction data is indeed very problematic, with accuracy values ranging from 10% to 50% (Gentleman & Huber, 2007; Sprinzak, Sattath, & Margalit, 2003; Von Mering et al., 2002) and coverage values lower than 50%, even for the most studied and curated interactome of *Saccharomyces cerevisiae* (yeast) (Hart et al., 2006). For many other species, one can expect the accuracy and coverage of the PPIs to be even lower.

How to deal with the false positive and false negative interactions? Computational methods can help to validate the existing protein interactions (to address false positive interaction issue) and predict novel protein interactions (to address false negative interaction issue). Our book provides bioinformatics solutions to predict/validate protein interactions by using various biological features/properties, such as protein domains, protein sequences (e.g., amino acids composition etc), protein functions, biological processes, cellular locations, structural information, and topological features extracted from the PPI network. Network cleansing techniques to improve the quality of the entire protein interaction networks are also surveyed in this book.

Mining PPI Network: A Computationally Challenging Graph Mining Problem

As mentioned earlier, a PPI network is typically modelled as an undirected graph (sometimes with weighted links) where the nodes represent unique proteins and the links denote interactions between two proteins. Such a network is a very large graph with thousands of vertices and tens of thousands of edges, even for a simple model organism such as yeast. One can only imagine the insurmountable complexity of the PPI networks for the more complicated species such as the human being. If we wish to investigate the evolution of various protein interaction networks or to alignment the PPI networks, the computational challenge is even more overwhelming since we will need to handle multiple humongous networks.

Graph theory is an important tool to facilitate the efficient analyses of the large scale interaction networks (Barabasi & Oltvai, 2004). In our book, we have surveyed several state-of-the-art graph mining techniques and their applications in discovering such useful biological objects from PPI networks as interaction motifs, network motifs, lethal proteins, protein complexes/functional modules, as well as in the evolutionary analyses PPI network and also network alignment/querying tasks. Although mining these objects from protein interaction networks are computationally challenging problems, it is possible to reduce the search space and time complexity and obtain better mining results by exploiting biological knowledge coupled with the development of novel efficient graph mining techniques.

Integrating Various Biological Evidences for System-Level Understanding

Recently, high throughput experimental technologies have been developed and as a result, an increasing number of large datasets at the various biological levels from genomics, proteomics to metabolomics

have been generated. Many of the datasets are deposited in centralized databases that are publicly accessible by all the researchers. At the same time, there are increasing efforts by biologists to provide annotations of the biological knowledge. Projects such as the Gene Ontology have been initiated to enable collective and systematic functional annotations of genes and proteins. All these efforts are motivated by the need to integrate various biological evidences for system-level understanding of the inherently complex cellular processes.

The benefits of integrating various biological evidences are two-fold. Firstly, as mentioned, the PPI data generated from the high-throughput methods are inevitably noisy and incomplete, so it is important to weight the links (i.e. the protein interactions) in the PPI networks by using appropriate confidence measures. For example, we can employ metrics from biological evidences such as reproducibility of the interactions from multiple experimental methods, support from such other non-interaction data as co-expression, co-localization and shared functions, as well as the conservation of the protein interactions across other genomes, and so forth. For example, a Bayesian network model had been developed to predict protein interactions by integrating noisy experimental interaction data with the weighted genomic features which are only weakly associated with interactions, for example, messenger RNA coexpression, coessentiality, and colocalization, and so forth (Jansen et al., 2003). The results were positive, indicating that it is useful to integrate biological evidences in addressing the limitations in the current quality of PPI data. Similarly, it is also possible to use machine learning methods, such as kernel methods (Ben-Hur & Noble, 2005), to integrate different biological resources into high-dimensional vector space to do classification and thus better predict protein interactions.

Secondly, integrating the PPI networks with additional biological evidences enables a systems-level understanding of biological processes and human diseases (Ideker & Sharan, 2008) by utilizing the rich information and strengths from different sources at different biological levels (Ghazalpour, Doss, Zhang, Wang, & Plaisier, 2006; Kelley & Ideker, 2005; Mootha et al., 2003). Obtaining quantitative and dynamic PPI data across different tissue cells and their integration with gene expression, functional, structural, and metabolic pathway data is the key towards successful development of diagnostics and therapeutics that target disease-relevant protein interactions. Numerous chapters in this book (e.g. the two book chapters that focus on disease related research) therefore describe methods that involved integration of various biological evidences.

On the other hand, one should be mindful that integration may not necessarily result high-quality data and results—the intrinsic noisiness of data and integration methods does really matter. Computational methods, such as machine learning approaches, probabilistic and statistical tools, should be carefully designed to maximally exploit the knowledge from various biological evidences and minimize the side effect from the irrelevant and noisy sources. At the same time, the users of the computational methods should be aware of the limitations and assumptions made by the designers of the computational methods. We hope such a message is clear in this book to both the computer scientists (designers of computational methods) and the biologists (users of the computational methods).

ORGANIZATION OF THE BOOK

The book is divided into five major sections, described below, covering the topics of: introduction (Section I), PPI network construction and cleansing (Section II), knowledge discovery from PPI networks (Section III), biological applications using PPI analysis (Section IV) and tools for analysis of PPI networks (Section V).

Section I: Introduction

In Sectino I, we will provide two introductory chapters, namely Chapter I "*Molecular Biology of Protein-Protein Interactions for Computer Scientists*" and Chapter II "*Data Mining for Biologists*".

Chapter I is intended for the readers from computer science who may not have adequate background about protein-protein interactions in biology. This introductory chapter will provide a brief tour of protein interaction types, experimental detection methods and their limitations. Biologically interesting examples of the Wnt signaling pathway and splice regulation will be used to demonstrate the challenges and opportunities that arise from assaying and analyzing protein interactions.

Chapter II is useful for the biologists who may not have adequate background knowledge on data mining algorithms. This chapter reviews the basics about various relevant data mining algorithms and their applications to biology. The chapter will focus on frequent pattern mining algorithms, including itemset mining, association rule mining, and graph mining. Originally developed for applications in totally different domains, these pattern mining algorithms can be used to automatically detect frequently appearing patterns/substructures from biological data. The chapter will summarize data mining's current biological applications and discuss the new directions of applying data mining techniques in biology.

Section II: PPI Network Construction and Cleansing

Chapter III, "*Domain-Based Prediction and Analysis of Protein Interaction Network*", will review domain-based models both for prediction of protein-protein interactions and for explaining the scale-freeness of protein-protein interaction networks. The chapter will describe the use computational methods such as the association method, EM method, SVM-based method and LP-based method, to infer domain-domain interactions from known protein-protein interaction data. These domain-domain interactions can then be used to predict new protein interactions based on the domains in the given proteins. This chapter will also review an evolutionary model of protein domains to explain how to derive a scale-free distribution of protein-protein interaction networks.

Chapter IV, "*Incorporating Graph Features for Prediction of Protein-Protein Interactions*", will review the machine learning methods which have been designed to predict protein interactions. Many of the current techniques either used the features extracted from protein sequences or other biological properties, or incorporated relational and structural features extracted from the PPI network. This chapter will describe predicting protein interactions using the graph features extracted from a PPI network along with other available biological features of the proteins and their interactions. The future trends in this area of predicting protein interactions based on information from PPI network will be highlighted.

Chapter V, "*Discovering Protein-Protein Interaction Sites from Sequence and Structure*", will first introduce both sequence and structural features which characterize protein-protein interaction sites and then review and compare the methodologies for predicting protein-protein binding site. The chapter will show how protein-protein interaction data can serve as a critical platform for advanced molecular recognition research and experimental design.

Chapter VI, "*Network Cleansing: Reliable Protein Interaction Networks*", will provide an overview of the existing methods for network cleansing. Different classes of cleansing algorithms will be described and their results are compared. The chapter will provide the information to guide the readers in the choice of the most appropriate methods, experiments and integrative data to obtain a portrait of reliable protein interaction network, at the same time highlighting the common biases and errors.

Section III: Knowledge Discovery from PPI Networks

Chapter VII, "*Discovering Interaction Motifs from Protein Interaction Networks*", will describe how to discover protein interaction motifs that are conserved in interacting proteins. Such knowledge discovery can help understand the mechanisms of protein interactions. It is based on the observation that protein interactions usually occur at some specific sites/motifs on the proteins and that these motifs are well conserved throughout the evolution among the proteins of the same family. The chapter will provide a review on the different approaches on mining for interaction motifs in PPI data, together with their implications, potentials, and possible areas of improvements in the future.

Chapter VIII, "*Discovering Network Motifs in Protein Interaction Networks*", will examine the methods to mine network motifs from PPI networks. Network motifs are potentially the basic building blocks of the protein interaction networks. By discovering the network motifs, researchers can gain insights on the general structure of the overall network, categorize different PPI networks into so-called "super-families", and even formulate hypothesis on how the network was formed through evolution. Both statistically-based methods and frequency-based methods will be described and the experimental results are compared in this chapter.

Chapter IX, "*Discovering Protein Complexes in Protein Interaction Networks*", will survey different state-of-the-art graph-based clustering techniques for detecting putative protein complexes from PPI networks. The discovered putative protein complexes are useful biological knowledge. For example, the knowledge of the complexes 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.

Chapter X, "*Evolutionary Analyses of Protein Interaction Networks*", will describe the methodologies for analyzing PPI data to understand molecular evolution and gain comparative genomics insights from such studies. In order to reveal the evolutionary mechanisms acting on the interactomes, it is necessary to compare protein interactions across species. This chapter will show that the evolution of proteins as the components of protein interaction networks can be understood through the evolutionary rates of the PPI networks. It will also show that protein interactions can influence the genomic locations of genes during evolution.

Section IV: Biological Applications Using PPI Analysis

Chapter XI, "*Discovering Lethal Proteins in Protein Interaction Networks*", will describe current methods for detecting lethal proteins from protein interaction networks. These methods exploit either the network properties alone or integrate with biological various biological information and properties. The study of lethal proteins is useful for understanding the minimal condition for cellular development and survival.

Chapter XII, "*Predicting Protein Functions from Protein Interaction Networks*", will investigate in detail the popular methods of using protein-protein interactions to predict protein functions. In particular, both local prediction methods and global optimization methods will be described and compared.

Chapter XIII, "*Protein Interactions for Functional Genomics*", will review the use of protein-protein interactions for the interpretation of genomic experiments. As an example, this chapter will describe the available resources and methodologies which are used to create a curated compilation of protein interactions and a novel approach to filter interactions.

Chapter XIV, "*Prioritizing Disease Genes and Understanding the Disease Pathways*", will describe a variety of useful data sources and bioinformatics tools and methods that can help prioritize disease

genes and identify disease pathways. The main strategy is to examine the similarity among the candidate genes and known disease genes at the functional level. The chapter will review different similarity measures and prevailing methods for integrating results from different functional aspects, and advocate many useful resources that the readers can use to investigate various diseases of their interest.

Chapter XV, "*Dynamics of Protein-Protein Interaction Network in Plasmodium Falciparum*", will present an application to integrate gene expression data with the protein interaction network acting in the different cellular process for *Plasmodium falciparum*, a malarial parasite, to reveal the dynamics in protein interaction network across different stages in the lifecycle of *Plasmodium falciparum*. The analysis hopes to demonstrate the power of strategic integration of genome-wide datasets for unraveling the dynamic complexity of biological pathways and processes.

Section V: Tools for Analysis of PPI Networks

Chapter XVI, "*Graphical Analysis and Visualization Tools for Protein Interaction Networks*", will present several visualisation and analysis software tools for protein interaction networks. The major advantages and disadvantages of each tool will be given, along with the analyses offered, and the capability for integration of other types of data. This chapter will also outline and evaluate two typical software approaches, namely desktop applications and web services.

Chapter XVII, "*Network Querying Techniques for PPI Network Comparison*", will describe and compare network querying techniques and tools applied to protein interaction networks. Network querying tools can be used to search a whole biological network to identify conserved occurrences of a query network module.

Chapter XVIII, "*Module Finding Approaches for Protein Interaction Networks*", will review the computational approaches for finding functional modules from protein interaction networks. The chapter will focus on those module finding tools implemented in freely available software packages. As in the other chapters in this book, this chapter will also discuss the key future trends and promising research directions with potential implications for clinical research.

REFERENCES

Bader, G. D., & Hogue, C. W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol, 20*(10), 991-997.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., & Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol, 22*(1), 78-85.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet, 5*(2), 101-113.

Ben-Hur, A., & Noble, W. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics, 21 Suppl.*

Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature, 340*(6230), 245-246.

Gentleman, R., & Huber, W. (2007). Making the most of high-throughput protein-interaction data. *Genome Biol, 8*(10), 112.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., & Plaisier, C. (2006). *Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight.*

Hart, G. T., Ramani, A. K., & Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol, 7*(11), 120.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome Res, 18*(4), 644-652.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., et al. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science, 302*(5644), 449-453.

Kelley, R., & Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology, 23*(5), 561-566.

Lalonde, S., Ehrhardt, D. W., Loque, D., Chen, J., Rhee, S. Y., & Frommer, W. B. (2008). Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J, 53*(4), 610-635.

Mootha, V. K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., & Hjerrild, M. (2003). Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A, 100*(2), 605-610.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., & Bragado-Nilsson, E. (2001). The tandem affinity purification (TAP) method: A general procedure of protein complex purification. *Methods, 24*(3), 218-229.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol, 17*(10), 1030-1032.

Sprinzak, E., Sattath, S., & Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *J Mol Biol, 327*(5), 919-923.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M., & Shames, I. (2008). An in vivo map of the yeast protein interactome. *Science, 320*(5882), 1465-1470.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., & Fields, S. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature, 417*(6887), 399-403.