Table of Contents

Detailed Table of Contents

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.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.