1

Introduction

1.1 RAPID GROWTH OF PROTEIN–PROTEIN INTERACTION DATA

Since the sequencing of the human genome was brought to fruition [154,310], the field of genetics now stands on the threshold of significant theoretical and practical advances. Crucial to furthering these investigations is a comprehensive understanding of the expression, function, and regulation of the proteins encoded by an organism [345]. This understanding is the subject of the discipline of proteomics. Proteomics encompasses a wide range of approaches and applications intended to explicate how complex biological processes occur at a molecular level, how they differ in various cell types, and how they are altered in disease states.

Defined succinctly, proteomics is the systematic study of the many and diverse properties of proteins with the aim of providing detailed descriptions of the structure, function, and control of biological systems in health and disease [241]. The field has burst onto the scientific scene with stunning rapidity over the past several years. Figure 1–1 shows the trend of the number of occurrences of the term "proteome" found in PubMed bioinformatics citations over the past decade. This figure strikingly illustrates the rapidly increasing role played by proteomics in bioinformatics research in recent years.

A particular focus of the field of proteomics is the nature and role of interactions between proteins. Protein–protein interactions (PPIs) regulate a wide array of biological processes, including transcriptional activation/repression; immune, endocrine, and pharmacological signaling; cell-to-cell interactions; and metabolic and developmental control [9,139,167,184]. PPIs play diverse roles in biology and differ based on the composition, affinity, and lifetime of the association. Noncovalent contacts between residue side chains are the basis for protein folding, protein assembly, and PPI [232]. These contacts facilitate a variety of interactions and associations within and between proteins. Based on their diverse structural and functional characteristics, PPIs can be categorized in several ways [230]. On the basis of their interaction surface, they may be homo- or hetero-oligomeric; as judged by their stability, they may be obligate or nonobligate; and as measured by their persistence, they may be transient or permanent. A given PPI can fall into any combination of these three categorical pairs. An interaction may also require reclassification under certain

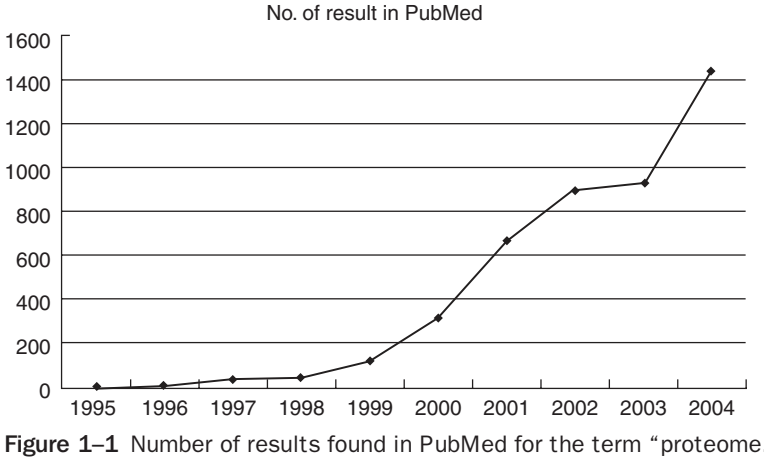

Figure 1–1 Number of results found in PubMed for the term "proteome." (Reprinted from [200] with permission of John Wiley & Sons, Inc.)

conditions; for example, it may be mainly transient in vivo but become permanent under certain cellular conditions.

It has been observed that proteins seldom act as single isolated species while performing their functions in vivo [330]. The analysis of annotated proteins reveals that proteins involved in the same cellular processes often interact with each other [312]. The function of unknown proteins may be postulated on the basis of their interaction with a known protein target of known function. Mapping PPIs has not only provided insight into protein function but also facilitated the modeling of functional pathways to elucidate the molecular mechanisms of cellular processes. The study of PPIs is fundamental to understanding how proteins function within the cell. Characterizing the interactions of proteins in a given cellular proteome will be the next milestone along the road to understand the biochemistry of the cell.

The result of two or more proteins interacting with a specific functional objective can be demonstrated in several different ways. The measurable effects of PPIs have been outlined by Phizicky and Fields [254]. PPIs can:

- alter the kinetic properties of enzymes; this may be the result of subtle changes at the level of substrate binding or at the level of an allosteric effect;
- act as a common mechanism to allow for substrate channeling;
- create a new binding site, typically for small effector molecules;
- inactivate or destroy a protein; or
- change the specificity of a protein for its substrate through interaction with different binding partners; for example, demonstrate a new function that neither protein can exhibit alone.

PPIs are much more widespread than once suspected, and the degree of regulation that they confer is large. To properly understand their significance in the cell, one needs to identify the different interactions, understand the extent to which they take place in the cell, and determine the consequences of the interactions.

In recent years, PPI data have been enriched by high-throughput experimental methods, such as two-hybrid systems [155,307], mass spectrometry [113,144], and

protein chip technology [114,205,346]. Integrated PPI networks have been built from these heterogeneous data sources. However, the large volume of PPI data currently available has posed a challenge to experimental investigation. Computational analysis of PPI networks has become a necessary supplemental tool for understanding the functions of uncharacterized proteins.

1.2 COMPUTATIONAL ANALYSIS OF PPI NETWORKS

A PPI network can be described as a complex system of proteins linked by interactions. The computational analysis of PPI networks begins with the representation of the PPI network structure. The simplest representation takes the form of a mathematical graph consisting of nodes and edges [314]. Proteins are represented as nodes in such a graph; two proteins that interact physically are represented as adjacent nodes connected by an edge. Based on this graphic representation, various computational approaches, such as data mining, machine learning, and statistical approaches, can be designed to reveal the organization of PPI networks at different levels. An examination of the graphic form of the network can yield a variety of insights. For example, neighboring proteins in the graph are generally considered to share functions ("guilt by association"). Thus, the functions of a protein may be predicted by looking at the proteins with which it interacts and the protein complexes to which it belongs. In addition, densely connected subgraphs in the network are likely to form protein complexes that function as a unit in a certain biological process. An investigation of the topological features of the network (e.g., whether it is scale-free, a small network, or governed by the power law) can also enhance our understanding of the biological system [5].

In general, the computational analysis of PPI networks is challenging, with these major difficulties being commonly encountered:

- *The protein interactions are not reliable.* Large-scale experiments have yielded numerous false positives. For example, as reported in [288], high-throughput yeast two-hybrid (Y2H) assays are ~50% reliable. It is also likely that there are many false negatives in the PPI networks currently under study.
- *A protein can have several different functions.* A protein may be included in one or more functional groups. Therefore, overlapping clusters should be identified in the PPI networks. Since conventional clustering methods generally produce pairwise disjoint clusters, they may not be effective when applied to PPI networks.
- *Two proteins with different functions frequently interact with each other.* Such frequent, random connections between the proteins in different functional groups expand the topological complexity of the PPI networks, posing difficulties to the detection of unambiguous partitions.

Recent studies of complex systems [5,227] have attempted to understand and characterize the structural behaviors of such systems from a topological perspective. Such features as small-world properties [319], scale-free degree distributions [28,29], and hierarchical modularity [261] have been observed in complex systems, elements that are also characteristic of PPI networks. Therefore, topological methods can be

used to address the challenges mentioned earlier and to facilitate the efficient and accurate analysis of PPI networks.

1.2.1 Topological Features of PPI Networks

Barabasi and Oltvai [29] introduced the concept of degree distribution, $P(k)$, to quantify the probability that a selected node in a network will have exactly k links. Networks of different types can be distinguished by their degree distributions. For example, a random network follows a Poisson distribution. In contrast, a scale-free network has a power-law degree distribution, $P(k) \sim k^{-\gamma}$, indicating that a few hubs bind numerous small nodes. When $2 \leq \gamma \leq 3$, the hubs play a significant role in the network [29]. Recent publications have indicated that PPI networks have the features of a scale-free network [121,161,198,313]; therefore, their degree distribution approximates a power law, $P(k) \sim k^{-\gamma}$. In scale-free networks, most proteins participate in only a few interactions, while a small set of hubs participate in dozens of interactions.

PPI networks also have a characteristic property known as the "small-world effect," which states that any two nodes can be connected via a short path of a few links. The small-world phenomenon was first investigated as a concept in sociology [217] and is a feature of a range of networks arising in both nature and technology, including the Internet [5], scientific collaboration networks [224], the English lexicon [280], metabolic networks [106], and PPI networks [284,313]. Although the small-world effect is a property of random networks, the path length in scale-free networks is much shorter than that predicted by the small-world effect [74,75]. Therefore, scale-free networks are "ultra-small." This short path length indicates that local perturbations in metabolite concentrations could permeate an entire network very quickly. In PPI networks, highly connected nodes (hubs) seldom directly link to each other [211]. This differs from the assortative nature of social networks, in which well-connected individuals tend to have direct connections to each other. In contrast, biological networks have the property of disassortativity, in which highly connected nodes are only infrequently linked.

A number of recent publications have proposed the use of centrality indices, including node degree, pagerank, clustering coefficient, betweenness centrality, and bridging centrality metrics, as measurements of the importance of components in a network [47,53,103,110,226,268,319]. For instance, betweenness centrality [225] was proposed to detect the optimal location for partitioning a network [122,145]. The modified betweenness cut approach has been suggested for use with weighted PPI networks that integrate gene expression [61]. Jeong's group has espoused the degree of a node as a key basis for the identification of essential network components [161]. In this model, power-law networks are very robust to random attacks but highly vulnerable to targeted attacks [7]. Hahn's group identified differences in degree, betweenness, and closeness centrality between essential and nonessential genes in three eukaryotic PPI networks (yeast, worm, and fly) [131]. Estrada's group introduced a new subgraph centrality measure to characterize the participation of each node in all subgraphs in a network [103,102]. Palumbo's group sought to identify lethal nodes by arc deletion, thus facilitating the isolation of network subcomponents [239]. Guimera's group devised a clustering method to identify functional modules

in metabolic pathways and categorized the role of each component in the pathway according to its topological location relative to detected functional modules [129].

As we will subsequently discuss in greater detail, the unique topological features found to be characteristic of PPI networks will play significant roles in the computational analysis of these networks.

1.2.2 Modularity Analysis

The idea of functional modules, introduced in [139], offers a major conceptual tool for the systematic analysis of a biological system. A functional module in a PPI network represents a maximal set of functionally associated proteins. In other words, it is composed of those proteins that are mutually involved in a given biological process or function. A wide range of graph-theoretic approaches have been employed to identify functional modules in PPI networks. However, these approaches have tended to be limited in accuracy due to the presence of unreliable interactions and the complex connectivity of the networks [288]. In particular, the topological complexity of PPI networks, arising from the overlapping patterns of modules and cross talks between modules, poses challenges to the identification of functional modules. Because a protein generally performs different biological processes or functions in different environments, real functional modules are overlapping. Moreover, the frequent, dynamic cross connections between different functions are biologically meaningful and must be taken into account [274].

In an attempt to parse this complexity, the hierarchical organization of modules in biological networks has been recently proposed [261]. The architecture of this model is based on a scale-free topology with embedded modularity. In this model, the significance of a few hub nodes is emphasized, and these nodes are viewed as the determinants of survival during network perturbations and as the essential backbone of the hierarchical structure. This hierarchical network model can plausibly be applied to PPI networks because cellular functionality is typically hierarchical in nature, and PPI networks include a few hub nodes that are biologically lethal.

The identification of functional modules in PPI networks or modularity analysis can be successfully accomplished through the use of cluster analysis. Cluster analysis is invaluable in elucidating network topological structure and the relationships among network components. Typically, clustering approaches focus on detecting densely connected subgraphs within the graphic representation of a PPI network. For example, the maximum clique algorithm [286] is used to detect fully connected, complete subgraphs. To compensate for the high-density threshold imposed by this algorithm, relatively dense subgraphs can be identified in lieu of complete subgraphs, either by using a density threshold or by optimizing an objective density function [56,286]. A number of density-based clustering algorithms using alternative density functions have been presented [12,24,247].

As noted, hierarchical clustering approaches can plausibly be applied to biological networks because of the hierarchical nature of functional modules [261,297]. These approaches iteratively merge nodes or recursively divide a graph into two or more subgraphs. To merge nodes iteratively, the similarity or distance between two nodes or two groups of nodes is measured and a pair is selected for merger in each iteration [17,263]. Recursive division of a graph involves the selection of nodes

or edges to be cut. Partition-based approaches have also been applied to biological networks. One partition-based clustering approach, the Restricted Neighborhood Search Clustering (RNSC) algorithm [180], determines the best partition using a cost function. In addition, other approaches have been applied to biological networks. For example, the Markov Clustering Algorithm (MCL) finds clusters using iterative rounds of expansion and inflation that, respectively, prefer the strongly connected regions and weaken the sparsely connected regions [308]. The line graph generation method [250] transforms a network of proteins connected by interactions into a network of connected interactions and then uses the MCL algorithm to cluster the PPI network. Samantha and Liang [272] applied a statistical approach to the clustering of proteins based on the premise that a pair of proteins sharing a significantly greater number of common neighbors will have a high functional similarity. The recently introduced STM algorithm [148] votes a representative of a cluster for each node.

Topological metrics can be incorporated into the modularity analysis of PPI networks. From our studies, we have observed that the bridging nodes identified in PPI networks serve as the connecting nodes between protein modules; therefore, removing the bridging nodes preserves the structural integrity of the network. Such findings can play an important role in the modularity analysis of PPI networks. Removal of the bridging nodes yields a set of components disconnected from the network. Thus, using bridging centrality to remove the bridging nodes can be an excellent preprocessing procedure to estimate the number and location of modules in the PPI network. Results of this research [151,152] have shown that such approaches can generate larger modules that discard fewer proteins, permitting more accurate functional detection than other current methods.

1.2.3 Prediction of Protein Functions in PPI Networks

Predicting protein function can be, in itself, the ultimate objective of the analysis of a PPI network. Despite the many extensive studies of yeast that have been undertaken, there are still a number of functionally uncharacterized proteins in the yeast database. The functional annotation of human proteins can provide a strong foundation for the complete understanding of cell mechanisms, information that is invaluable for drug discovery and development. The increased interest in and availability of PPI networks have catalyzed the development of computational methods to elucidate protein functions.

Protein functions may be predicted on the basis of modularization algorithms. If an unknown protein is included in a functional module, it is expected to contribute toward the function that the module represents. The generated functional modules may thus provide a framework within which to predict the functions of unknown proteins. Each generated module may contain a few uncharacterized proteins along with a larger number of known proteins. It can be assumed that the unknown proteins play a positive role in realizing the function of the generated module. However, predictions arrived at through these means may be inaccurate, since the accuracy of the modularization process itself is typically low. For greater reliability, protein functions should be predicted directly from the topology or connectivity of PPI networks.

Several topology-based approaches that predict protein function on the basis of PPI networks have been introduced. At the simplest level, the "neighbor counting

method" predicts the function of an unknown protein by the frequency of known functions of the immediate neighbor proteins [274]. The majority of functions of the immediate neighbors can be statistically assessed [143]. The function of a protein can be assumed to be independent of all other proteins, given the functions of its immediate neighbors. This assumption gives rise to a Markov random field model [85,196]. Recently, the number of common neighbors of the known protein and the unknown protein has been taken as the basis for the prediction of function [201].

Machine learning has been widely applied to the analysis of PPI networks, and, in particular, to the prediction of protein functions. A variety of methods have been developed to predict protein function on the basis of different information sources. Some of the inputs used by these methods include protein structure and sequence, protein domain, PPIs, genetic interactions, and gene expression analysis. The accuracy of prediction can be enhanced by drawing upon multiple sources of information. The Gene Ontology (GO) database [84] is one example of such semantic integration.

1.2.4 Integration of Domain Knowledge

As noted, the accuracy of results obtained from computational approaches can be compromised by the inclusion of false connections and the high complexity of networks. The reliability of this process can be improved by the integration of other functional information. Initially, the identification of similarities in gene sequence can be a primary indicator of a functional association between two genes. Additionally, genome-level methods for functional inference, such as gene fusion events and phylogenetic profiling, can generate useful data pointing to functional linkages. Beyond this, we know that genes with correlated expression profiles determined through microarray experiments are likely to be functionally related. Many studies [65,66,153,304] have investigated the integration of PPI networks with gene expression data to improve the accuracy of the functional modules identified. Finally, as briefly noted earlier, GO [18,301] can be a useful data source to combine with the PPI networks. GO is currently one of the most comprehensive and well-curated ontology databases in the bioinformatics community. It represents a collaborative effort to address the need for consistent descriptions of genes and gene products. The GO database includes GO terms and their relationships. The former are well-defined biological terms organized into three general conceptual categories that are shared across different organisms: biological processes, molecular functions, and cellular components. The GO database also provides annotations to each GO term, and each gene can be annotated on one or more GO terms. The GO database and its annotations can thus be a significant resource for the discovery of functional knowledge. These tools have been employed to facilitate the analysis of gene expression data [89,105,147] and have been integrated with unreliable PPI networks to accurately predict functions of unknown proteins [84] and identify functional modules [68,70].

1.3 SIGNIFICANT APPLICATIONS

The systematic analysis of PPIs can enable a better understanding of cellular organization, processes, and functions. Functional modules can be identified from the

PPI networks that have been derived from experimental data sets. There are many significant applications following this analysis. In this book, the following principal applications to which this analysis can be applied will be discussed:

- *Predicting protein function.* As noted earlier, the most basic application of PPI networks is the use of topological analysis to predict protein function. The generated functional modules can serve as a framework within which to predict the functions of unknown proteins. Each generated module may contain a few uncharacterized proteins. By associating unknown proteins with the known proteins, we can suggest that those proteins participate positively in performing the functions assigned to the modules.
- *Lethality analysis.* The topological analysis of PPI networks can be used to systematically assess the biological importance of bridging and other nodes in a PPI network [65,66,70,148]. Lethality, a crucial factor in characterizing the biological indispensability of a protein, is determined by examining whether a module is functionally disrupted when the protein is eliminated. Information regarding lethality is compiled in most PPI databases. For example, the MIPS database [214] indicates the lethality or viability of each included protein. Such sources allow the researcher to compare the lethality of nodes with high bridging-score values to that associated with other competing network parameters in the PPI networks. These comparisons reveal that nodes with the highest bridging scores are less lethal than both randomly selected nodes and nodes with high degree centrality. However, the average lethality of the neighbors of the nodes with the highest bridging scores is greater than that of a randomly selected subset. Our research has indicated that bridging nodes have relatively low lethality; interconnecting nodes are characterized by higher lethality; and modular nodes and peripheral nodes have, respectively, the highest and lowest proportion of lethal proteins. These results imply that many of the bridging nodes do not perform tasks critical to biological functions [151,152]. As a result, these nodes would serve as good targets for drugs, as discussed later.
- *Assessing the druggability of molecular targets from network topology.* Translating the societal investments in the Human Genome Project and other similar large-scale efforts into therapies for human diseases is an important scientific imperative in the post–human-genome era. The efficacy, specificity/selectivity, and side-effect characteristics of well-designed drugs depend largely on the appropriate choice of pharmacological target. For this reason, the identification of molecular targets is a very early and critical step in the drug discovery and development process. The goal of the target identification process is to arrive at a very limited subset of biological molecules that will become the principal focus for the subsequent discovery research, development, and clinical trials. Pharmacological targets can span the range of biological molecules from DNA and lipids to metabolites. In fact, though, the majority of pharmacological targets are proteins. Effective pharmacological intervention with the target protein should significantly impact the key molecular processes in which the protein participates, and the resultant perturbation should be successful in modulating the pathophysiological process of interest. Another important consideration that is sometimes overlooked during the target identification step is the potential for side effects.

Ideally, an appropriate balance should be found among efficacy, selectivity, and side effects. In practice, however, compromises are often required in the areas of specificity/selectivity and side effects, since pharmacological interventions with proteins that are central to key processes will likely affect many biological pathways. We have observed that the biological correlates of the nodes with the highest bridging scores indicate that these nodes are less lethal than other nodes in PPI networks. Thus, they are promising drug targets from the standpoints of efficacy and side effects.

1.4 ORGANIZATION OF THIS BOOK

This book is intended to provide an in-depth examination of computational analysis as applied to PPI networks, offering perspectives from data mining, machine learning, graph theory, and statistics. The remainder of this book is organized as follows:

- Chapter 2 introduces the three principal experimental approaches that are currently used for generating PPI data: the Y2H system [121,156,307], mass spectrometry (MS) [113,120,144,187,210,303], and protein microarray methods [114,346].
- Chapter 3 discusses various computational approaches to the prediction of protein interactions, including genomic-scale, sequence-based, structure-based, learning-sequence-based, and network topology-based techniques.
- Chapter 4 introduces the basic properties of and metrics applied to PPI networks. Basic concepts in graphic representation employed to characterize various properties of PPI networks are defined for use throughout the balance of the book.
- Chapter 5 discusses the modularity analysis of PPI networks. Various modularity analysis algorithms used to identify modules in PPI networks are discussed, and an overview of the validation methods for modularity analysis is presented.
- Chapter 6 explores the topological analysis of PPI networks. Various metrics used for assessing specific topological features of PPI networks are presented and discussed.
- Chapter 7 focuses on greater detail on one type of modularity algorithm, specifically, the distance-based modularity analysis of PPI networks.
- Chapter 8 focuses on greater detail on graph-theoretic approaches for modularity analysis of PPI networks.
- Chapter 9 discusses the flow-based analysis of PPI networks.
- Chapter 10 examines statistical- and machine learning-based analysis of PPI networks.
- Chapter 11 discusses the integration of domain knowledge into the analysis of PPI networks.
- Chapter 12 presents some of the more recent approaches that have been developed for incorporating diverse biological information into the explorative analysis of PPI networks.
- Chapter 13 offers a synthesis of the methods and concepts discussed throughout the book and reflections on potential directions for future research and applications.

1.5 SUMMARY

The analysis of PPI networks poses many challenges, given the inherent complexity of these networks, the high noise level characteristic of the data, and the presence of unusual topological phenomena. As discussed in this chapter, effective approaches are required to analyze PPI data and the resulting PPI networks. Recently, a variety of data-mining and statistical techniques have been applied to this end, with varying degrees of success. This book is intended to provide researchers with a working knowledge of many of the advanced approaches currently available for this purpose. (Some of the material in this chapter is reprinted from [200] with permission of John Wiley & Sons, Inc.)