Network-Based Protein-Protein Interaction Analysis

Bing Wang1,2,3,*, Hao Shen1, Peng Chen4 and Jun Zhang5

1School of Electronics and Information Engineering, Tongji University, China
2The Advanced Research Institute of Intelligent Sensing Network, Tongji University, China
3The Key Laboratory of Embedded System and Service Computing, Tongji University, China
4Institute of Health Sciences, Anhui University, China
5College of Electrical Engineering and Automation, Anhui University, China

Abstract

High-throughput experimental technologies in protein interaction continue to alter the study of current system biology, and a large-scale data can be available. Protein-protein interactions on these experimental platforms, however, present numerous production and bioinformatics challenges. Some issues like the functional modules identification, protein complexes prediction, protein function prediction and disease-related gene prioritization have become increasingly problematic in the analysis of protein-protein interaction networks. The development of powerful, efficient prediction methods for the structure and function analysis of protein interaction network is critical for the research community to accelerate research and publications. Currently, Network-based approaches are drawing the most attention in analyzing protein interactions. This review aims to describe the state-of-art of network-based strategies and applications to infer protein interactions.

ABBREVIATIONS

PPI: Protein-Protein Interaction; PPIN: Protein-Protein Interaction Network

INTRODUCTION

Cells and organs are very complex systems because the interactions and the relations between cells to cells, DNA to RNA and RNA to proteins are very multifaceted and large in volume and length [1]. Of the different types of biological interactions, protein-protein interactions (PPIs) are one of the most significant, interesting and complicated interactions because some protein may work as an individual entity, but usually two or more proteins bind together and form a complex to carry out their biological functions. Biological processes are largely dependent on protein-protein interactions which carry out numerous functions, from DNA replication, cell replication, protein synthesis, and energy production to molecule transport, to various inter- and intracellular signaling. Several experimental methods have been developed to analyze protein-protein interactions, including yeast two-hybrid assay [2-5] protein chips [6], and mass spectrometry of purified protein complexes [7,8], which produce a vast amount of information and make it possible for researchers to study the biological activities systematically.

Currently, many protein interaction databases had been developed, which can support the establishment of interaction networks. With comparison to the analysis technologies investigated PPIs in the interaction partner or interface level [9-12], protein-protein interaction network (PPIN) based-methods had caught researchers’ attentions for it can analyze the functions of proteins in a system biology level. For example, HParrish H et al., built a PPIN in 2007 for the bacterium Campylobacter jejuni, a food-borne pathogen and a major cause of gastroenteritis worldwide, and identified a number of conserved sub-networks, biological pathways and putative essential genes that may be used to identify potential new antimicrobial drug targets for C. jejuni and related organisms [13].

Recently, some comprehensive reviews provided insights into the analysis and applications of protein-protein interaction networks [14-18]. This up to date review specifically focuses on four aspects: the functional modules identification, protein complexes prediction, protein function prediction and disease-related gene prioritization.

Protein-protein interaction network (PPIN) analysis

Interaction networks can be represented as an interaction graph, where nodes represent proteins and edges represent pairwise interactions (an example can be found in Figure 1). The analysis of the network structure or topological properties of PPIN, such as distribution of node degree (number of incoming and outgoing edges per node), network diameter (average of the shortest distance between pairs of nodes), clustering coefficient (proportion of the potential edges between the neighbors of a node that are effectively observed in the graph), have led...
to the observation of some apparently recurrent properties of biological networks: power-law degree distribution, small world, high clustering coefficients, and modularity[19-26]. The network whose degree distribution follows a power law also has been called as scale-free network, in which the fraction $P(k)$ of nodes in the network having $k$ connections to other nodes goes for large values of $k$ as $P(k) \sim k^{-\gamma}$, where $\gamma$ is a parameter whose value typically ranges from 2 to 3. The character of small-world means that most nodes in PIN are not neighbors of one another, but they can be reached from every other by a small numbers of hops or steps. Modularity is another characteristic feature of PIN, where some protein groups are highly connected among them yet with lesser connections between modules.

**Important functional modules identification**

As one type of biological functional network, it is essential to understand the relationship between the organization of the network and its functions [19,27]. Therefore, clustering algorithms play an important role in the analysis of PIN, and can be used to uncover functional modules and obtain hints about cellular organization [28]. Brohee and Helden had evaluated four algorithms: Markov Clustering (MCL) [29], Restricted Neighborhood Search Clustering (RNSC) [30], Super Paramagnetic Clustering (SPC) [31], and Molecular Complex Detection (MCODE) [32] in 2006 and found that MCL and RNSC outperform SPC and MCODE in robustness where the test was implemented on unweighted graphs [19].

Regularized MCL (R-MCL), an efficient and robust variation of MCL, was proposed by HSatuluri et al, which can improve the accuracy of identifying functional modules by R-MCL’s regularize operation and balance parameter [33]. Shih and Parthasarathy developed a ‘Soft’ R-MCL (SR-MCL) algorithm, a new variation of R-MCL, which can identify overlapped clusters within PIN [34].

Wang and Qian proposed a novel optimization formulation LCP$^2$ which can identify both dense and sparse modules simultaneous based on protein interaction patterns in given networks through searching for low two-hop conductance sets by Markov random walk on graphs [35]. Moreover, they presented another two algorithms, SLCP$^2$ and GLCP$^2$, to identify non-overlapping and overlapping functional modules. The authors also proposed a new joint network clustering algorithm, AS Model, which can combine both topology and homology information [36].

Jia et al. proposed a dense module searching (DMS) method to identify candidate sub-networks or genes for complex diseases by integrating the association signal from GWAS datasets into the human PIN [37]. This method extensively searches for sub-networks enriched with low $P$-value genes in GWAS datasets, and the experiments show the effectiveness of DMS by testing in two GWAS datasets for complex diseases, i.e. breast cancer and pancreatic cancer.

Wang et al., developed a fast algorithm, HC-PIN, based on the local metric of edge clustering value for hierarchical clustering, which can be used both in the unweighted network and in the weighted network [38]. The authors demonstrated that the usage of local metric in the algorithm HC-PIN not only improves its efficiency, but also enhances its robustness to the high rate of false positives in PIN. Meanwhile, HC-PIN can identify significant modules with low density.

**Network-based protein complex prediction**

A protein complex is a group of proteins that interact with each other at the same time and place, forming a single multi-molecular machine [16,39,40]. In a network-based way, the problem of identifying protein complexes from PPI data can be formulated as that of detecting dense regions containing many connections in PPI networks, or regions with large weights in weighted networks [41].

Nepusz et al., proposed an algorithm, named Cluster ONE, clustering with overlapping neighborhood expansion for detecting potentially overlapping protein complexes from protein-protein interaction data [41]. In Cluster ONE, there is a concept of the cohesiveness score, a measurement can determine how likely a group of proteins form a complex, had been calculated and it uses a greedy growth process to find protein complexes. The authors also found that taking into account network

---

**Figure 1** C. jejuni protein interaction networks. (a) The C. jejuni interaction dataset (CampyYTH v3.1) and (b) the higher confidence subset. In each case most of the proteins (square nodes) are connected into a single large network; the unconnected interactions are in the upper right of each panel. The networks in (a,b) connect over 79% (663 total) and 65% (548 total) of the unnamed and presumed poorly characterized proteins (yellow nodes), respectively [13].
weights, an estimation of the reliability of protein interactions and is included as edge labels in PPIN, can greatly improve the detection of protein complexes, although it is difficult to assess the reliability of the weights.

Zhang et al., constructed ontology augmented networks to predict protein complexes, which can combine the information from protein-protein interaction networks and gene ontology [42]. This method can formulate the topological structure of protein-protein interaction networks and the similarity of gene ontology annotations into a unified distance measure. The experimental results in this work showed that ontology augmented networks can get a higher F1 measure for predicting protein complexes.

Wu et al., presented a novel rough-fuzzy clustering (RFC) method to detect overlapping protein complexes in PPIN [43]. Rather than the graph models employed in previous approaches, this method applied fuzzy relation model by integrating fuzzy sets and rough sets to deal with overlapping complexes, and it determines whether the protein belongs to one or to multiple complexes by calculating the similarity between one protein and each complex. The work compared the RFC with several previous methods and show big performance improvement, i.e., the precision, sensitivity and separation are 32.4%, 42.9% and 81.9% higher than mean of the five methods in four weighted networks, and are 0.5%, 11.2% and 66.1% higher than mean of the six methods in five un weighted networks.

There are many studies focus on protein complexes identification from PPIN. Shen et al. proposed a complex mining algorithm called Multistage Kernel Extension (MKE) algorithm...
using a two-level kernel strategy based on the centrality-lethality rule [44]. Yang et al. applies a sophisticated natural language processing system, PPI Extractor, to extract PPI data from biomedical literature, and integrated PPI datasets to detect protein complexes [45]. Hanna and Zaki proposed another ranking algorithm, named Pro Rank+, which can figure out important proteins and complexes in the Bio GRID repository, and some of them had been demonstrated by previous studies [46].

Network-based protein function prediction

In the past two decades, the vigorous development in sequencing technologies posed a novel challenge that is how to elucidate protein function from wealth of genomics data generated [47]. A few years ago, Sharan et al., showed in their work that, even for the most well-studied organisms such as yeast, about one-fourth of the proteins remain uncharacterized, and this high percentage does not drop evidently now [figure 2] [48]. Fortunately, protein interaction networks for many species provide a special view to predict the functions of proteins in a computational way.

Wu et al., systematically identified apoptotic/cell cycle related key proteins using a Naïve Bayesian model based on a modified apoptotic/cell cycle related PPI networks [49]. Their work not only identified some already known key proteins such as p53, Rb, Myc and Src but also found that the proteasome, Cullin family members, kinases and transcriptional repressors play important roles in regulating apoptosis and the cell cycle. Meanwhile, they found some proteins were enriched in some pathways such as those of cancer, the proteasome, the cell cycle and Wnt signalling, which can provide further new clues towards future anticancer drug discovery.

Davis et al., predicted protein functions from the conservation of topology-function relationships in protein-protein interaction network [50]. They developed a statistical framework that is built upon canonical correlation analysis where the graphlet degrees represented the wiring around proteins in PINs and gene ontology (GO) annotations described the protein functions. Their method can characterize statistically significant topology-function relationships, and uncover the functions that have conserved topology in PINs. Applications to the PINs of yeast and human show that their proposed framework had identified seven biological processes and two cellular components GO terms to be topologically orthologous.

Saha et al., proposed a software, named FunPred, to predict protein functions based on network neighborhood properties [51]. There are two approaches in FunPred, one applies a combination of three simple-yet-effective scoring techniques: the neighborhood ratio, the protein path connectivity and the relative functional similarity. Another is a heuristic approach using the edge clustering coefficient to reduce the search space by identifying densely connected neighborhood regions. Wu et al., developed a regularized non-negative matrix factorization (RNMF) algorithm for protein functional properties prediction where attribute features, latent graph, and unlabeled data information in PPI networks had been used [52]. Peng et al., predicted protein functions using an unbalanced Bi-random walk (UBiRW) algorithm on PPI network and functional interrelationship network by considering the topological and structural difference between them [53].

Network-based diseases-related gene prioritization

Elucidating the underlying molecular mechanisms of diseases has become increasingly important in disease prevention, diagnosis, and drug design [54]. PPIN-based analysis approaches have been recently developed and applied to diseases analysis [54-60]. Candidate gene prioritization is one of important application of network-based knowledge. Studies on the properties of disease genes in protein interaction networks have shown that two genes sharing higher-order topological similarities are likely to interact with each other and may be associated with the same or similar phenotypes [61,62]. Wu et al., established a regression model that measures the correlation between gene closeness and phenotype similarities in the PPI network to prioritize potential candidate genes for inherited diseases on the basis of correlation scores [63]. Dezso et al., applied a modified shortest path between’s to prioritize candidate genes in PPI networks, where a candidate gene has high relevant score to the disease of interest if it laid more on significantly shorter paths connecting nodes of known disease genes than other genes in the network [H59H]. Recently, Luo and Liang proposed a random walk-based algorithm on the reliable heterogeneous network (RWRHN) to prioritize potential candidate genes for inherited diseases, in which a PPI network reconstructed by topological similarity, a phenotype similarity network and known associations between diseases and genes [figure 3][54]

DISCUSSION AND CONCLUSION

The development of powerful high-throughput experimental technologies has fundamentally changed the study of current system biology [64]. However, huge data produced by these different platforms also presents some serious challenges, such as the high false positive rate in current ‘wet’ experiments and the validation of the analytical results from ‘dry’ methods. Network-based analysis, a kind of computational tools, can adopt graph theory to address the inherent knowledge within the protein interaction data. For example, it can score the importance of proteins using degree information of nodes within PPIN no matter the network is weighted or unweighted. One of biggest advantage of network-based PPI approaches is it can analyze the interactions at a system biology level. Also, it can easily combine other information, such as GO term, protein subcellar location, gene regulation, and so on, into the processing framework, which in turn makes the PPI networks modeling more precise. Especially, network-based analysis will become more and more important for some complex diseases for many of them can be seen as network diseases, that is to say, the root of these diseases is not one or few molecules. In this work, we only focus on the new progresses published on four aspects, i.e., the functional modules identification, protein complexes prediction, protein function prediction and disease-related gene prioritization. It is clear that network-based methods hold incredible promise for protein interaction research in many other applications, and their capabilities in the hands of investigators will undoubtedly accelerate our understanding of the mechanism of cell to perform their functions.
ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (Nos. 61300058, 61472282 and 61374181), and Anhui Provincial Natural Science Foundation (No.1508085MF129).

REFERENCES


