Chapter 15
Network Analysis in Translational Research

Minlu Zhang, Jingyuan Deng, Lirong Tan, Ye Chen, and Long Jason Lu

Abstract Networks, where nodes/vertices denote entities and links/edges denote associations, provide a unified representation for a variety of complex systems, from social relationships to molecular interactions. In general, network analysis has proved useful in applications such as prediction of protein function, guiding the design of wet-lab experiments, and discovery of biomarkers of disease. Driven by the availability of large-scale data sets and appropriate informatics' tools, the research community has begun to apply network analysis of these data to define underlying causes of pediatric diseases. This will almost certainly lead to more effective strategies for prevention and treatment. In this chapter, we will introduce classic and the state-of-the-art network analysis methodologies, approaches and their applications. We then provide three examples of recent research, where network analysis is being applied in pediatrics. These include identification of targets

M. Zhang, Ph.D. • J. Deng • L. Tan
School of Computing Sciences and Informatics, University of Cincinnati College of Engineering & Applied Science, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA

Y. Chen
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA

School of Electronics and Computing Systems, University of Cincinnati College of Engineering & Applied Science, Cincinnati, OH, USA

L.J. Lu, Ph.D. (✉)
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA

Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA
e-mail: long.lu@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*, Translational Bioinformatics 2, DOI 10.1007/978-94-007-5149-1_15,

for development of new drugs to treat *Pseudomonas* infections, protein interactions in high-density lipoprotein particles that underlie development of cardiovascular disease, and analysis of networks in brain that may underlie conditions such as Attention Deficit Hyperactivity Disorder.

15.1 Introduction

Networks or graphs provide a unified representation for a variety of complex systems, from social relationships, e.g., co-authorship of different authors, to associations between molecular entities. Molecular networks, where each node denotes a molecule like a protein or gene and each edge denotes an association, form the foundation of contemporary systems biology. To date, network models are applied to depict a variety of biological systems. For example, protein-protein interaction (PPI) networks or maps, where each node is a protein and each edge is an interaction, are adopted to describe physical interactions or attachments between protein pairs in sets of proteins such as those found within a cell organelle, or particle (Rual et al. 2005). Gene co-expression networks, where each node is a gene and each edge indicates gene expression similarity, are applied to summarize similarities in expression patterns between gene pairs (Stuart et al. 2003). Transcriptional regulatory networks, where each node is either a transcription factor protein or a target gene, depict transcriptional regulatory relationships between transcription factors and target genes (Guelzim et al. 2002). Other common biological networks include metabolic networks of metabolites and their chemical reactions and signal transduction pathways of multiple interacting genes and proteins (Zhang et al. 2010). Analyzing large-scale networks may provide a systems-level view of these biological systems and thus advance understanding of the underlying biological processes. In recent years, applications of network analysis have proved useful in guiding the design of experiments, uncovering novel and effective prognostic biomarkers of disease, and facilitating drug discovery. Some basic features of networks are illustrated in Fig. 15.1.

In this chapter, we will first introduce tools and processes for state-of-the-art biological network analyses and provide a few examples of their application to basic research into causes of pediatric diseases. Next, we will introduce the global properties and characteristics of large-scale biological networks: scale-free topology, small-worldness, disassortativity, and modular structures. In the third section, we will discuss in brief three common types of network analyses and show how they have been applied: (1) topological analysis to identify proteins/genes of interest, (2) motif analysis to extract and identify small subnetwork motif structures, and (3) modular analysis to cluster large molecular networks to self-contained subnetworks as functional units or modules. Finally, examples of analyses will be provided, including: (1) human-*Pseudomonas* interaction networks and network-based identification of drug targets for treatment of *Pseudomonas* infections and,

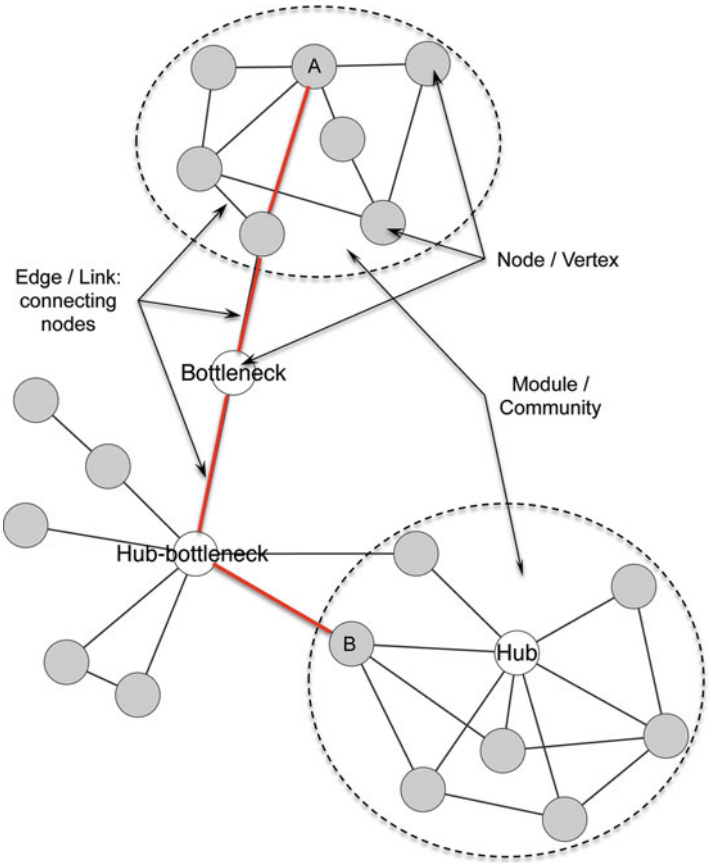

Fig. 15.1 Basic features of networks. Nodes or vertices are entities in a network, and edges or links connect pairs of nodes. The *red* edges form a path that connects nodes *A* and *B*. A hub node is defined as a node with a large number of neighbors, and a bottleneck is defined as a node that many shortest paths go through. In this sample network, the node in the *lower left* is both a hub and a bottleneck. Fifteen nodes in a network with 22 nodes form two different densely intra-connected and loosely inter-connected modular subnetwork structures called modules or communities (in *dashed ellipses*)

(2) analysis of proteins and protein-protein interactions in subspecies of high density lipoprotein particles that play a role in cardiovascular disease , and (3) construction of brain functional connectivity maps from neuroimaging data in children with a focus on attention-deficit/hyperactivity disorder (ADHD). Chapters 16 (Bioinformatics and Orphan Diseases) and 17 (Functional Genomics – Transcriptional Regulatory Networks) provide additional examples of the application of network analysis to studies of childhood diseases.

15.2 Biological Networks: Properties and Characteristics

Large-scale biological networks of various types are similar in a number of topological characteristics, namely scale-free, small-worldness, disassortativity, and modular structures. Similar to many other large-scale networks such as the World Wide Web and social networks, most large-scale biological networks of model organisms are observed to approximate a scale-free structure. Scale-free means that distributions of nodes within these networks follow a power-law distribution, denoted as $P(k) \sim k^{-\gamma}$, where $P(k)$ is the fraction of nodes that have k connections to other nodes (degree k), and the degree exponential constant γ is a constant usually smaller than 3 (Barabasi and Oltvai 2004). Figure 15.2 shows the scale-free topology of a large-scale human protein-protein interaction (PPI) network, using PPI data from the current release of the Human Protein Reference Database (Goel et al. 2012). Scale-free topology indicates that in a network, the higher the degree of the node, the greater the number of connections it has. A small fraction of nodes have a very high degree of connections (large number of connections) and are called hubs. Hubs are critically important to the overall functioning of the network. The scale-free topology contributes to the robustness of the network. For example, one

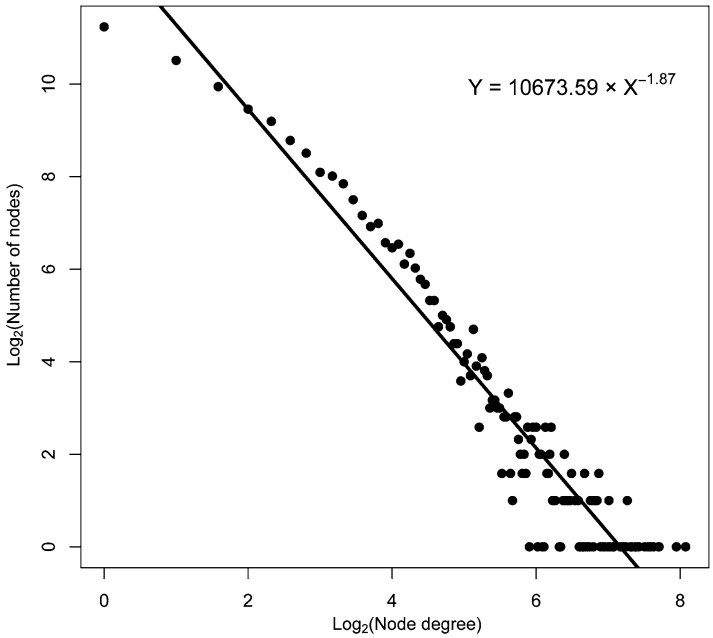

Fig. 15.2 The scale-free topology of a large-scale Protein-Protein Interaction (PPI) network. The node degree of a large-scale PPI network based on the ninth release of the Human Protein Reference Database (HPRD) follows a power-law distribution. A total of 9,517 proteins and 37,004 pairwise PPIs exist in the network

study showed that removing up to 80 % of randomly selected nodes in a network would not disconnect the rest of the network, due to the scale-free topology and sparseness of connections of most of the nodes (Albert et al. 2000). On the other hand, networks are more vulnerable, if there is an attack on the highly influential hub nodes (non-random attack).

Most large-scale biological networks have small-world properties as a consequence of their scale-free topology. Small-worldness is often defined and depicted as high clustering coefficient and low characteristic path length of a network, which means that nodes in a biological network are involved in self-contained community-like subnetworks and most node pairs are connected via short paths (Fig. 15.1). For example, in a *Escherichia coli* metabolic network of 287 nodes and 317 edges, where a node is a metabolite and an edge represents a reaction, a typical path between the most distant metabolites contains only four reactions (Wagner and Fell 2001). The small-world property also contributes to the robustness of a network, because when any perturbation occurs, most of the network components can conduct a timely response because of the small-worldness.

Disassortativity, which means hub and non-hub nodes are more likely to be connected than randomly expected, was discovered in large-scale PPI as well as transcriptional regulatory networks (Maslov and Sneppen 2002). The disassortativity, in accord with scale-free and small-world characteristics, strengthens the robustness of the network. The property also helps to separate subnetwork structures in the network, which may correspond to functional units that perform particular functions in the biological processes.

Large-scale networks are composed of functional modules that perform distinct, but relevant, functions (Hartwell et al. 1999). In biological networks, such functional modules correspond to modular subnetwork structures, which are highly intra-connected within themselves and loosely inter-connected with each other. The molecules within each module are likely annotated with the same or similar function. The modular structures are hierarchically organized, as observed in most large-scale biological networks, because of both modular and scale-free properties (Ravasz 2009).

15.3 Network Analysis and Applications: Topology, Motifs and Modules

The most common types of network analysis include topological, motif, and modular analysis. In this section, we introduce these methods of network analysis and their applications to knowledge discovery. Table 15.1 lists some tools commonly used in network analysis, their function, and how to access them. Typical outputs of these analyses are shown in Figs. 15.1, 15.4, 15.5, and 15.6.

Given a biological network, it is of interest to characterize its topology. In order to quantitatively measure the topological structure of a network, a number of classic

Table 15.1 Examples of tools (software applications) commonly used in network analysis

Tool	Function	How to access
Cytoscape (Shannon et al. 2003)	Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. Cytoscape supports many use cases in molecular and systems biology, genomics, and proteomics. Various plug-ins for analysis and visualization have been developed in addition to its core functionalities.	www.cytoscape.org/
Gephi (Bastian et al. 2009)	Gephi is an open source standalone software platform for visualizing networks and complex systems. Gephi provides topological and modular analysis for biological and social networks. A variety of plug-ins are also available for analysis, visualization layout, and applications of specific purposes.	www.gephi.org
VisANT (Hu et al. 2009)	VisANT provides a webstart application, a standalone Java application, and a batch mode for visualization and analysis. It has multiple layout options and is scalable to visualize genome-scale biological networks. Users can perform statistical, motif, and modular enrichment analysis using VisANT.	visant.bu.edu
Pajek (Batagelj and Mrvar 2004)	Pajek is a standalone application for network visualization. Pajek supports 3D layout and can perform modular analysis of networks to decompose them into modules.	vlado.fmf.uni-lj.si/pub/networks/pajek/
tYNA (Yip et al. 2006)	tYNA is a Web server that for biological network visualization. tYNA supports basic network analysis of major topological characteristics and motifs.	tyna.gersteinlab.org/
JUNG (O'Madadhain et al. 2003)	JUNG, short for Java universal network/graph framework, is a Java-based open-source software library package. JUNG provides various implemented algorithms for network modeling, visualization and analysis. It is a general library and is not specific to biological network applications.	jung.sourceforge.net/
PINA (Wu et al. 2009)	PINA is a Web server primarily for protein-protein interaction network visualization and statistical analysis. Important nodes such as hubs and bottlenecks can be identified. PINA can perform GO term enrichment analysis. It has good scalability.	cbg.garvan.unsw.edu.au/pina/

(continued)

Table 15.1 (continued)

Tool	Function	How to access
NeAT (Brohee et al. 2008)	NeAT is suite of Web-server tools for network analysis and visualization. NeAT supports topological analysis, pathway extraction, and modular analysis (multiple network clustering algorithms such as clique enumeration and MCL clustering). It also provides functionalities like network comparison, cluster comparison, and heat map generation.	rsat.ulb.ac.be/neat/
N-Browse (Kao and Gunsalus 2008)	N-Browse is both a webstart application and a Web-server for visualizing molecular networks. It is connected to and integrated with multiple databases such as modMine database for biological data mining.	http://aquila.bio.nyu.edu/NBrowse2/

network statistics from graph theory have been adopted. Commonly used network statistics include: degree, centrality, clustering coefficient, shortest path length, and eccentricity. These network statistics not only can describe single nodes in the network, but also characterize the network as a whole.

In an undirected network such as a PPI network, node degree is defined as the number of connections linked to a node. In a directed network such as a transcriptional regulatory network, the number of out-going and in-coming links of a node can be measured separately as out-degree and in-degree. The average degree of all nodes in a network is an indication of how dense or sparse the connections are. Similar to other real-life networks, biological networks are sparse, which is defined as having far fewer than the possible maximum links $\frac{n\times(n-1)}{2}$, where n is the number of nodes. The clustering coefficient of a node ranges from 0 to 1 and can be defined as the fraction of links compared with all possible links for its neighboring nodes. The average clustering coefficient for all nodes in a network is an indication of whether these nodes are involved in densely connected subnetwork clusters. The higher the average clustering coefficient of a network, the more likely that its nodes form self-contained subnetwork modules. The shortest path length between two nodes is the minimum number of links needed to connect two nodes. Characteristic path length for a network is defined as the median of the means of the shortest path lengths of all nodes taken together. Small-worldness is defined as having high clustering coefficient and low characteristic path length.

In biological networks, hubs and bottlenecks, are of special interest. A hub is an influential local connector, while a bottleneck is a bridge-like connector between different communities (Fig. 15.1). A number of studies have reported enriched functional essentiality in hub or bottleneck nodes of protein-protein interaction networks of model organisms, or high correlation between topological connectivity and functional essentiality (Jeong et al. 2001; Yu et al. 2007). Based on these observations, one application of retrieving hubs and bottlenecks is to identify

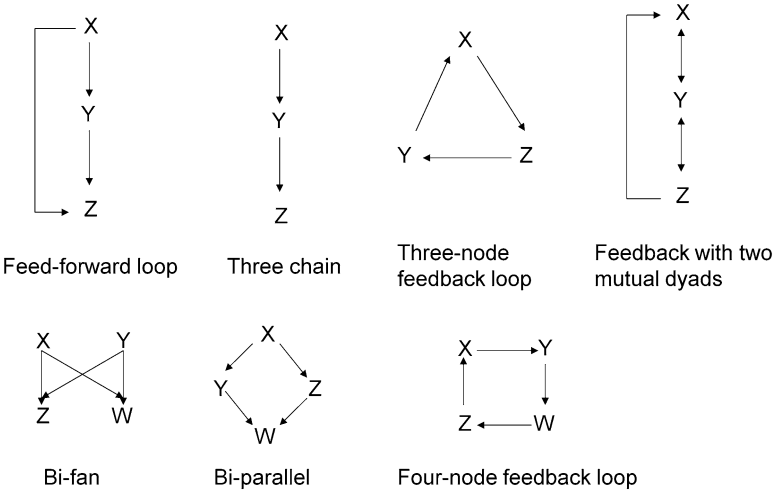

Fig. 15.3 Commonly identified motifs in networks. Each letter "x", "y", "z", or "w" indicates a node. Each motif consists of 3–4 nodes in these commonly identified motifs

important proteins or genes in biological networks. For example, hubs or bottlenecks in the PPI networks of human pathogens may help identify essential proteins for the survival of these pathogens. Such proteins are potential targets for development of drugs to treat infections caused by the pathogens.

A network motif is defined as an interconnected subnetwork structure of several nodes that occurs much more frequently than random expectation. Motifs have been identified in a wide range of networks, such as transcriptional regulatory networks, social networks, electrical circuits, the World Wide Web, and ecological food webs. Figure 15.3 shows several types of commonly identified motifs. Despite different sizes, types, and complexity of networks, the approaches to identify network motifs are generally similar. A common method involves scanning a network for all patterns with several interconnected nodes, recording frequencies of their occurrence, and comparing pattern frequencies with those in networks that have randomly rewired links. A subnetwork pattern is considered a motif, if it has a much higher frequency than randomly expected, often defined as two or more standard deviations greater than the average number of occurrences in random networks. Different types of motifs are over-represented in networks with different types of functions. For example, in transcriptional regulatory networks, the "feed-forward loop" motif is common.

Biological networks are highly modular. Structurally, they are composed of densely interrelated nodes of proteins or genes. These subnetworks have high connectivity density and are loosely interconnected (Fig. 15.1). Functionally, modules in biological networks are those interconnected molecules that together perform specific functions. Structural modules are believed to correspond to functional modules. In the past decade, the research community has made efforts to uncover functional

modules through the extraction of structural modules from large-scale biological networks. Various graph clustering algorithms and methods have been developed or adopted for modular analysis. These methods can be roughly categorized into four types based on the underlying methodologies, including density-based, partition-based, centrality-based, and hierarchical clustering approaches (Zhang et al. 2010). Many of the algorithms are implemented in the tools listed in Table 15.1.

A density-based clustering method seeks to identify densely connected subnetwork structures in a network, for example fully connected or near-fully connected subgraphs. Clique enumeration is a simple and straightforward density-based method, which can identify all cliques of up to n nodes (Spirin and Mirny 2003). Other density-based methods may seek to find densely connected subnetworks with less stringent connectivity criteria than fully connected cliques. For example, Palla et al. developed a clustering method to identify k-clique subnetwork structures, which are defined as unions of all cliques of k nodes that share $k - 1$ nodes (Palla et al. 2005). In a biological network such as a PPI network, proteins in a clique or clique-like module likely correspond to those belonging to the same protein complex. Modules identified by density-based methods can overlap with each other. Such a property is favored in biological networks because proteins and genes may participate in different functional units under different contexts. One limitation of density-based methods is that a module may never cover loosely connected nodes in a network.

Unlike density-based clustering methods, partition-based methods seek to identify an optimal partition of all nodes in a network based on a certain cost function. Cost function is an important concept in the theory of networks. It is a measure of how far away a particular solution is from an optimal solution to the problem to be solved. The cost function is dependent on what is being modeled and the *a priori* assumptions about the optimal solution. Such a method often starts with a random partition of a network, followed by iterative reassignment of nodes from different previous partitions until an optimal cost is achieved. In each iteration, a cost function is calculated, and the method outputs a partition when an optimal cost is achieved. One example of such cost function is "modularity", a score between 0 and 1, where a higher score indicates many more within-module links and many less between-module links than expected. A limitation of such methods is that network partitioning assigns nodes to distinct modules and does not allow overlapping modules.

Centrality-based clustering methods are based on network centrality statistics. One commonly used measurement of centrality-based clustering is edge betweenness, which is defined as the number of paths (edges that are connected by nodes) of the shortest lengths between all possible pairings of node pairs that pass through an edge. Edges with high betweenness correspond to bridge-like connectors in a network, the disconnection of which would result in self-contained node clusters. A classic edge betweenness based clustering method takes all the links in a network as input, and iteratively removes an edge with the highest edge betweenness (Girvan and Newman 2002). A network can be clustered into subnetwork modules after a certain percentage of edge removal. In a biological network, the resulting modules

from a centrality-based clustering method often form a hierarchy because of the hierarchical modularity properties of biological networks.

Classic hierarchical clustering methods have also been applied to clustering biological networks. In a PPI network, a distance measure between a pair of proteins in a network can be as simple as their pairwise distance (shortest path length d) or a measurement based on the shortest path length, e.g. d^2. The proteins are grouped together from the closest in the network to the farthest. The grouping forms a hierarchy, and given a cutoff of the distance, proteins can be put into different clusters. One limitation of hierarchical clustering is that different linkage methods and different cutoffs would result in different output modules. In addition, classic hierarchical clustering algorithms do not consider overlapping clusters.

Prediction of the function of proteins is also a major application of network modular analysis in PPI networks. In PPI networks, subnetworks identified via network clustering likely correspond to functional modules. Despite the variety of clustering methods and algorithms, protein members in each subnetwork tend to have the same or similar functional annotations, e.g., Gene Ontology annotations (Berardini et al. 2010). Based on grouping of proteins from clusters, those proteins with unknown functions are predicted to have functional annotations that are highly enriched for the group (Brun et al. 2004; King et al. 2004).

15.4 Applications of Network Analysis

15.4.1 Network Pharmacology – Identification of Targets for Development of New Drugs to Treat Pseudomonas Infections

Network analysis can be applied to the identification of potential drug targets for treating *Pseudomonas* infections. *Pseudomonas aeruginosa* (*PA*), a common bacterium, is a ubiquitous opportunistic pathogen that can cause chronic infections in patients with damaged tissues or reduced immunity. For example, *PA* often infects the lungs of children with cystic fibrosis. *PA* possesses a remarkable capacity to resist multiple front-line antibiotics, and more effective drugs to treat *PA* infection are urgently needed.

The philosophy in traditional drug design, i.e., the “one gene, one drug, one disease” paradigm, focuses on the individual properties of a protein, for example, whether the deletion of a gene is lethal to an organism and if it is, then inhibition of the product of the gene by a drug should also be lethal. On the other hand, many effective drugs have been found to affect a handful of targets instead of a single protein (Huang 2002). Recently, understanding of biological networks and molecule functionality has given rise to a new discipline, network pharmacology, that provides the opportunity to identify new drug targets in an infectious organism, while avoiding targets that cause toxicity in the host (Zhang et al. 2012).

In order to utilize the approach advocated by network pharmacology for drug target identification to treat *PA* infection, the establishment of a genome-scale PPI network is a prerequisite. Currently, the knowledge of *PA* PPI networks is limited. Experimentally verified PPIs in *PA* are limited to a handful of protein pairs from individual small-scale studies (Goll et al. 2008). In a recent study, we computationally predicted a genome-scale PPI network in *PA* using machine learning (Zhang et al. 2012). Specifically, we first constructed a positive reference set of *PA* PPIs by homology mapping based on PPI data from three closely related species of bacteria: *C. jejuni, E. coli, and H. pylori*. Experimentally verified PPIs of the three organisms were retrieved from DIP, BIND, and literature (Xenarios et al. 2002; Peregrin-Alvarez et al. 2009; Hu et al. 2009). We then collected and compiled a set of eight genomic features of *PA* genes or their protein products, each of which may be highly relevant to protein physical interactions and thus have predictive power of PPI. We performed training and testing by a random forest classifier, and conducted PPI prediction using the trained classifier. As the output, a confidence score is associated with each predicted PPI representing the probability of the physical interaction or the co-involvement in a protein complex.

To verify these predicted PPIs, we performed three types of computational validations and confirmations. We first used an independent testing set of 35 experimentally verified *PA* PPIs from MPIDB (Goll et al. 2008), the majority of the PPIs were predicted positive in our results, while none of these PPIs can be identified when using homology mapping. In addition, we compared the network structure of the predicted network with established networks and found that they were largely the same. For example, the node degree distribution of the predicted network follows a power-law with a degree exponent of 1.69. Furthermore, we extracted 20 known *PA* drug target proteins in the network from DrugBank (Wishart et al. 2006) and observed that 12 of them hubs. This is consistent with the hypothesis that drug targets tend to be topologically important hubs and bottlenecks and functionally annotated with essentiality.

After extensive computational validation of the predicted PPI network, we performed subsequent analysis to identify putative drug target proteins based on their topological importance and functional essentiality. We also identified a set of essential functional modules that are highly enriched with these essential proteins (Zhang et al. 2012). The essential hubs and modules that are not targets of presently available drugs provide potential targets for development of new drugs that are likely to be effective in treatment of *PA* infections.

The infectious process of bacterial pathogenesis often involves PPIs between bacterial and host proteins. A map of human-*PA* protein interactions should help elucidate the disease mechanisms of *PA* infections in cystic fibrosis patients. We extracted 12 human-*PA* protein interactions between 11 human proteins and three *PA* proteins from pathogen interaction gateway (PIG) (Driscoll et al. 2009). Combining these human-*PA* interactions with the human PPIs from human protein reference database (HPRD) (Keshava Prasad et al. 2009) and *PA* PPIs from the high-confidence PPI network that contain proteins involved in human-*PA* protein interactions, we constructed a map of human-*PA* interactions (Fig. 15.4). Based

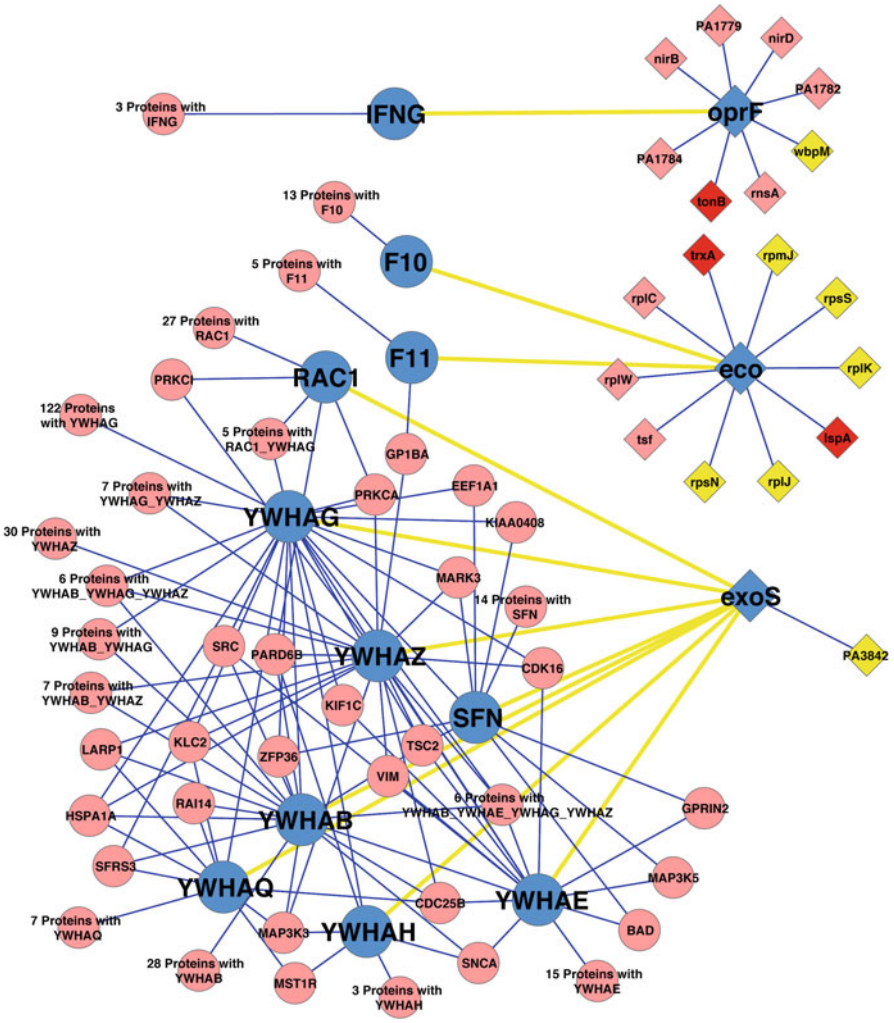

Fig. 15.4 A map of human-*Pseudomonas aeruginosa*(*PA*)PPIs. A *round node* is a human protein, and a *diamond node* is a *PA* protein. *Blue nodes* are proteins involved in human-*PA* PPIs that are denoted by *yellow edges*. *Yellow* and *red PA* proteins are essential proteins, and *red* ones are predicted to be potential drug targets. *Blue* edges denote corresponding PPIs in a human interactome and the high-confidence *PA* PPI network

on this human-*PA* interaction network, we found that *Cellular protein metabolic process* (GO:0044267) was enriched among *PA* proteins (p-value = 7.9e-6). In addition, top significantly enriched GO annotations by human proteins were identified, such as *ATP binding* (GO:0005524) of molecular functions, and *nerve growth factor receptor signaling pathway* (GO:0048011), *blood coagulation* (GO:0007596), *intracellular signaling pathway* (GO:0023034), and *platelet activation* (GO:0030168)

(p-value < 1e-12). Interestingly, 10 of 22 *PA* proteins in the map of human-*PA* interactions are essential proteins, and three proteins (TonB, thioredoxin TrxA, and lipoprotein signal peptidase LspA) were predicted as potential drug targets by this study, both numbers being significantly higher than randomly expected (p-value < 1e-3). Such intensified relevance of the proteins in the map supports the validity of their being candidate drug targets for further investigations.

15.4.2 Protein-Protein Interaction Networks in High Density Lipoprotein Particles

In this section, we will illustrate the application of network analysis to characterization of protein-protein interactions (PPIs) in complex mixtures of proteins within subcellular particles using as an example our recent study on high-density lipoprotein (HDL) subspecies. Databases and tools used include Gene Ontology (GO) (Berardini et al. 2010), Human Protein Reference Database (HPRD) (Keshava Prasad et al. 2009) and The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang da et al. 2009).

HDLs are blood-borne complexes consisting of proteins and lipids that play critical roles in cardiovascular disease (CVD) (Boden 2000; Lewis and Rader 2005; Cuchel and Rader 2006). There is growing controversy about whether and/or how HDL prevents CVD. A widely accepted mechanism is called reverse cholesterol transport, where HDL promotes cholesterol efflux from peripheral cells such as macrophage-derived foam cells in the vessel wall to transport excess cholesterol and other lipids back to the liver for catabolism (Franceschini et al. 1991). HDL is also involved in other CVD-protective functions, including anti-oxidation, anti-inflammation and endothelial relaxation (Watson et al. 1995; Naqvi et al. 1999; Nofer et al. 2002; Barter et al. 2004; Negre-Salvayre et al. 2006; Mineo et al. 2006).

There is evidence that HDL is composed of numerous distinct particle subpopulations, each containing a unique protein makeup that plays critical physiological roles (Gordon et al. 2010a, b). The major activities of HDLs rely on the cooperative interactions among its protein components. Recent proteomic studies on HDL have identified upwards of 75 distinct HDL associated protein components (Gordon et al. 2010a; Heller et al. 2005; Karlsson et al. 2005; Rezaee et al. 2006; Vaisar et al. 2007; Davidson et al. 2009). We fractionated HDL particles by gel filtrations and identified the proteins within subpopulations using mass spectrometry (Gordon et al. 2010a). In total, 106 HDL-associated proteins were identified, 17 of which have not been reported previously to be associated with HDL particles (Gordon et al. 2010a). To elucidate the potential roles of these newly identified HDL proteins, we first constructed a PPI network among these HDL associated proteins using data retrieved from the HPRD (Keshava Prasad et al. 2009) (Fig. 15.5). Eight out of the 17 new HDL proteins have a direct interaction with known HDL proteins, and an additional three can be connected to known HDL proteins by one intermediate

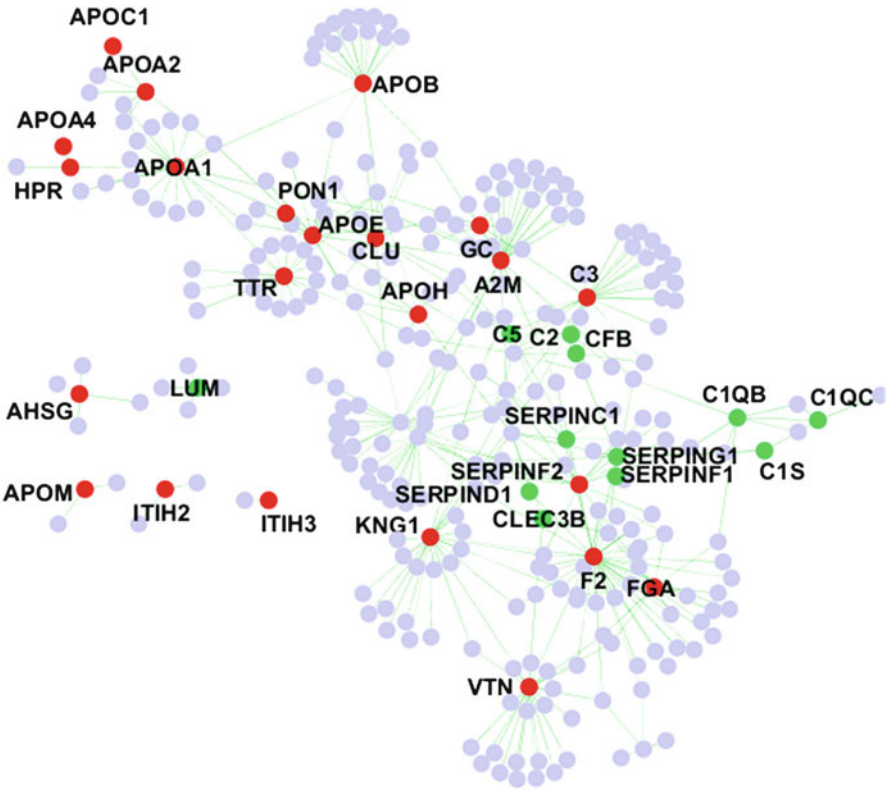

Fig. 15.5 Protein interaction networks between newly identified lipid associated proteins and known HDL associated proteins. In this PPI network, *red nodes* denote known HDL proteins; *green nodes* denote new phospholipid associated proteins

protein. This supports the possibility that these new proteins play a role in HDL function. We also examined the functions of the proteins in this network based on Gene Ontology (GO) (Berardini et al. 2010). GO annotations provide information about the distribution of proteins in different biological processes and molecular functions. We were able to assign new functional roles to several of these newly identified proteins, such as complement function and serine protease inhibitors (Gordon et al. 2010a).

In addition to discovering new HDL-associated proteins, the proteomic data generated in gel filtration has also allowed us to identify distinct HDL particles using a network approach. We measured two proteins' likelihood to co-occur in the same particle by quantifying the similarity between the migration patterns exhibited in gel filtration. The result is a similarity network that contains 90 nodes and 278 edges (Fig. 15.6). In the network, each node is an HDL-associated protein and each edge is an interaction between two proteins that have highly similar co-migration patterns. Given the similarity network, next we identified cliques that may correspond to the

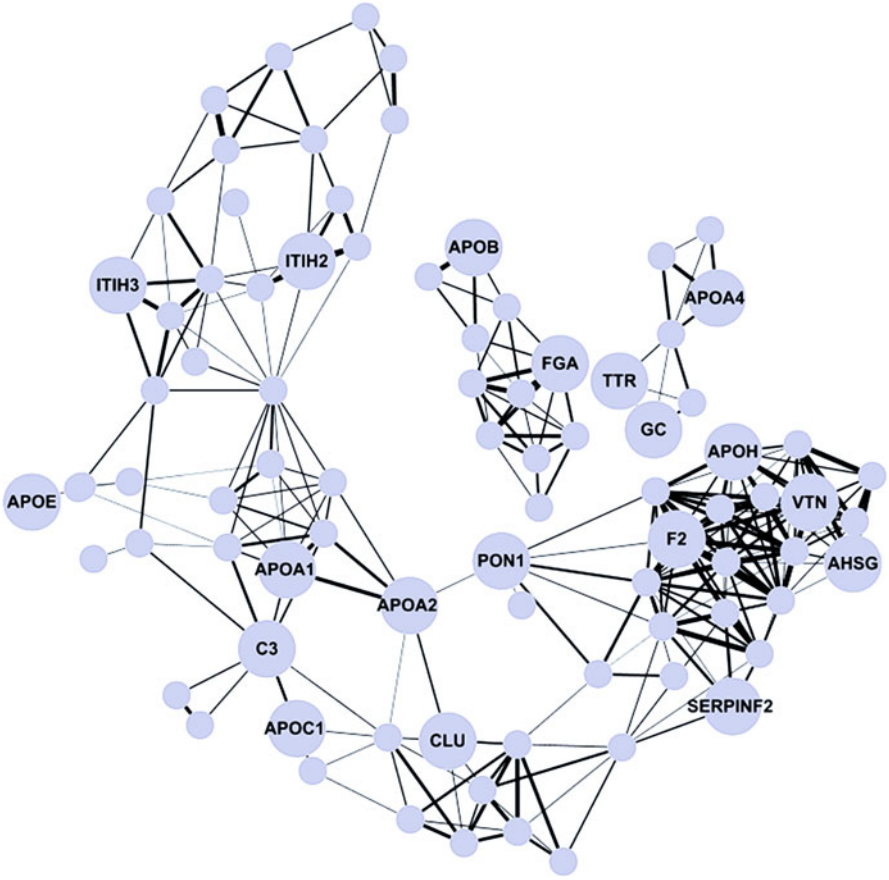

Fig. 15.6 The co-migration similarity network of HDL proteins, where each node is a HDL protein and the width of the edge is proportional to the co-migration similarity score between the connected two proteins

distinct HDL subspecies. A clique in a network is a subset of its nodes, where every two nodes in this subset are connected by an edge; a maximal clique is a clique that cannot be extended by including one more adjacent node. Thus, a maximal clique in the similarity network may correspond to the entirety or partial of a HDL subspecies, where its components have quite similar co-migration patterns. We identified all the maximal cliques from the similarity network, which generated 52 maximal cliques sizing from 3 to 10 nodes (proteins).

We elucidated the functions of each clique by performing functional enrichment analysis on these cliques based on GO annotations. We found 35 out of the 52 maximal cliques have significantly enriched functions. The most enriched functions are regulation of cholesterol esterification, complement activation, coagulation, innate immune response, and regulation of acute inflammatory response. These results are consistent with previously known HDL functions such as reverse

cholesterol transport, anti-oxidation, anti-inflammation and endothelial relaxation, but suggest that subpopulations of the HDL particles differ in functional profiles. These findings are guiding the design of mouse gene knockout experiments to validate the existence of these cliques and their proposed functions.

15.4.3 Detection of Brain Functional Connectivity Map in Children with Attention-Deficit/Hyperactivity Disorder (ADHD)

Imaging techniques, such as magnetoencephalography (MEG), positron emission tomography (PET), and magnetic resonance imaging (MRI), provide opportunities for analyzing both structural and functional organization of human brains. Traditional analysis of brain images focuses on the study of individual brain regions. The differences between the brains of two groups of humans, e.g., normal versus diseased, are compared on a region-by-region basis. There are limitations to this approach. First, some disorders may affect a large number of brain regions, and the significance of changes in any individual region is difficult to establish. Second, the functional impairment may be caused by abnormalities in the interconnections among several brain regions. To overcome these limitations, the different anatomic regions of the brain must be analyzed as an interconnected whole.

Network analysis takes into account patterns of connections among brain regions or voxels in digital images. Different from molecular networks presented in the previous sections, the nodes in brain networks represent voxels or brain regions, depending on the expected granularity of the analysis, and the edges represent the structural or functional connections among the different voxels or brain regions.

There are three main types of brain networks: structural networks, functional networks, and effective networks. The edges in these three brain networks represent physical connections, functional correlations, and causal functional relations, respectively. Brain networks can be represented as weight matrices or graphs. Figure 15.7 shows an example of a functional brain network. The forms of representation for the structural, functional, and effective networks are similar to one another.

Topological analysis can be performed on the brain networks to reveal the connection patterns of different groups of brains, e.g., healthy versus diseased. The results of analysis can reveal the topological differences that cannot be revealed by traditional non-network based approaches. The topological differences can provide new insights into the mechanism of the diseases.

Small-world property is one of the topological features found in healthy brains. It is reported that a variety of psychiatric disorders, such as Alzheimer's disease in adults and ADHD in children, are associated with disrupted small-worldness. In the following, we use a study of brain network of ADHD as an example to illustrate how the brain network approach can be applied to study the mechanisms of the diseases.

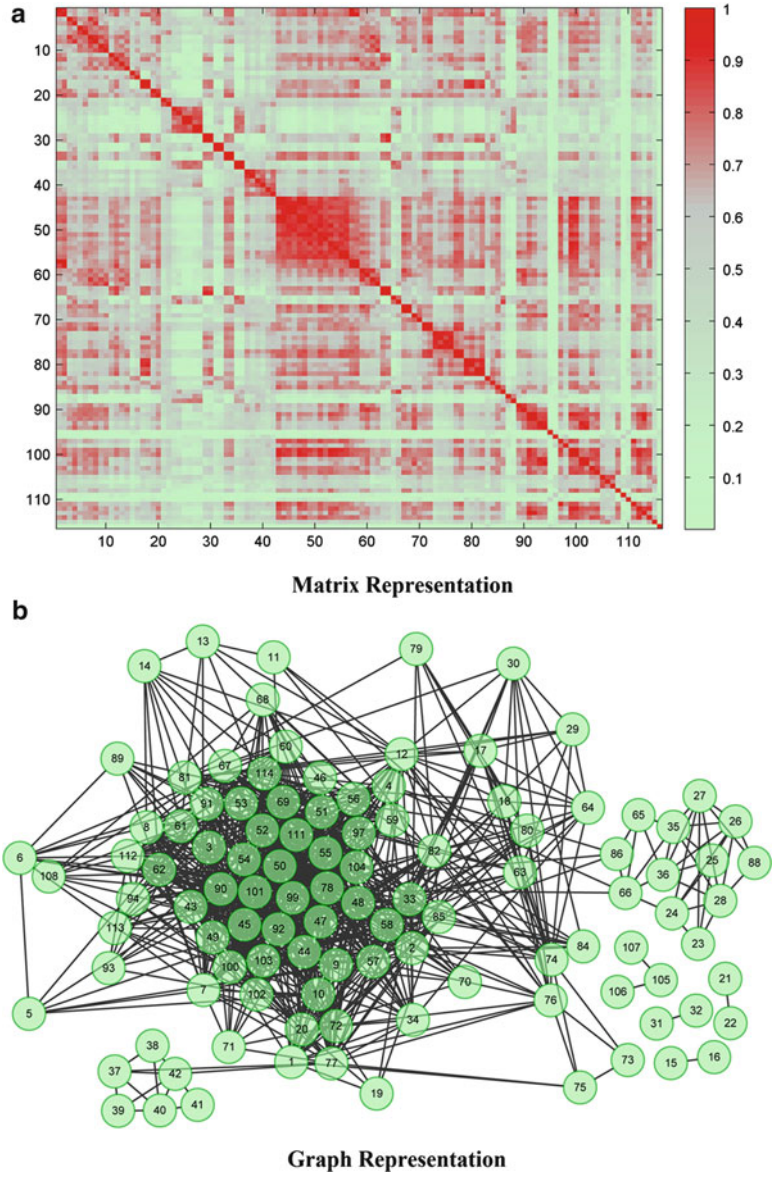

Fig. 15.7 An example of a functional brain network generated from the fMRI data reported in (Wang et al. 2009) and reanalyzed using tools described in the text of this chapter. The whole brain is parcellated into 116 brain regions based on an automated anatomical labeling (AAL) template. The correlation between every pair of brain regions is calculated and shown in the correlation matrix in (**a**). The correlations are thresholded at 0.8 and the resulting graph is shown in (**b**)

ADHD is a psychiatric disorder that is widely studied in children. Wang et al. (2009) investigated the topological changes of the functional brain networks in children with ADHD. The whole brain was parcellated into 90 regions using the anatomical automatic labeling (AAL) template, and a mean time course was calculated for each region. Pair-wised Pearson's correlation between the regional time courses was calculated to construct the 90×90 connectivity matrix, which was subsequently thresholded at different levels of cost to generate the binary graph (network G), with nodes representing brain regions and edges representing undirected connections. Unlike other investigators, they used the global and local efficiency instead of clustering coefficient and characteristic path length, to evaluate the small-worldness of the networks. Then, they generated populations of regular networks $\left(G_{reg}\right)$ and random networks (G_{rand}) that have the same number of nodes and edges as network G, respectively. The network G was considered to be small-world, if it has a global efficiency larger than G_{reg} while smaller than G_{rand}, and a local efficiency larger than G_{rand} while smaller than G_{reg}.

For the topological analysis, they first compared the global efficiencies and local efficiencies of brain networks with those of the random networks, in which nodes are randomly connected, and regular networks, in which nodes are connected to a fixed number of other nodes. Results demonstrated that brain networks in both ADHD group and control group exhibited small-world topology. Comparing the ADHD group with the control group, no significant difference in global efficiency was found, whereas the local efficiency was significantly different between the two groups at a range of cost ($0.12 < \text{cost} < 0.16$). At the cost of 0.15, when the between-group difference in local efficiency was most significant, the ADHD group exhibited an increased local efficiency (p-value $= 0.016$) and a decreased global efficiency (p-value $= 0.144$). The patterns of the changes in the network efficiency supported the notion that ADHD promoted a shift from small-world networks toward regular networks. Furthermore, it has been suggested that the small-world configuration of brain networks maximizes the efficiency of information processing at a relatively low wiring cost. Disease-related shift toward either random or regular networks may reflect disrupted local/global communication in the brain, although the biological causes underlying the shift remain unclear.

Despite several recent successes, there are still many challenges to applying network approaches to study brain diseases. Further studies are necessary to address questions such as why there is a network topological alteration, when do changes in network topology first occur, how changes in topology progress, and how differences among groups of brain networks can be measured more accurately? There are also specific challenges in applying network approaches to the study of psychiatric disorders in the brains of children. Children's brains have a larger functional reorganization capability than adults' brains. The plasticity of the children's brains could cause difficulty in identifying the abnormality of the brain networks.

15.5 Summary

In this chapter, we discussed general biological network analysis with a focus on network topology and structures of subnetworks. We also provided three examples of applications of network analysis in translational research. In general, biological network analysis has proved useful in applications such as predicting the function of proteins, guiding the design of wet-lab experiments, and identifying markers of various stages of disease. Applications of network analysis have gone beyond molecular entities to characterization of many other complex relationships among entities, e.g., regions of the brain. However, they have not been fully exploited in studies of pediatric diseases. Chapters 16 (Bioinformatics and Orphan Diseases) and 17 (Functional Genomics – Transcriptional Regulatory Networks) provide additional examples of how tools of network analysis can be applied to identification of potential new therapies of relatively rare diseases of childhood and to research on development and maturation of human organ systems.

References

Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature. 2000;406(6794):378–82.

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.

Barter PJ, et al. Antiinflammatory properties of HDL. Circ Res. 2004;95(8):764–72.

Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: International AAAI conference on weblogs and social media. 2009;361–362.

Batagelj V, Mrvar A. Pajek – analysis and visualization of large networks. In: Graph drawing software. Berlin/New York: Springer; 2004. p. 77–103.

Berardini TZ, et al. The gene ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010;38(Database issue):D331–5.

Boden WE. High-density lipoprotein cholesterol as an independent risk factor in cardiovascular disease: assessing the data from Framingham to the Veterans Affairs High–Density Lipoprotein Intervention Trial. Am J Cardiol. 2000;86(12A):19L–22.

Brohee S, et al. Network analysis tools: from biological networks to clusters and pathways. Nat Protoc. 2008;3(10):1616–29.

Brun C, Herrmann C, Guenoche A. Clustering proteins from interaction networks for the prediction of cellular functions. BMC Bioinformatics. 2004;5:95.

Cuchel M, Rader DJ. Macrophage reverse cholesterol transport: key to the regression of atherosclerosis? Circulation. 2006;113(21):2548–55.

Davidson WS, et al. Proteomic analysis of defined HDL subpopulations reveals particle-specific protein clusters: relevance to antioxidative function. Arterioscler Thromb Vasc Biol. 2009;29(6):870–6.

Driscoll T, et al. PIG–the pathogen interaction gateway. Nucleic Acids Res. 2009;37(Database issue):D647–50.

Franceschini G, Maderna P, Sirtori CR. Reverse cholesterol transport: physiology and pharmacology. Atherosclerosis. 1991;88(2–3):99–107.

Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci U S A. 2002;99(12):7821–6.
Goel R, et al. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. Mol Biosyst. 2012;8(2):453–63.
Goll J, et al. MPIDB: the microbial protein interaction database. Bioinformatics. 2008;24(15):1743–4.
Gordon SM, et al. Proteomic characterization of human plasma high density lipoprotein fractionated by gel filtration chromatography. J Proteome Res. 2010a;9(10):5239–49.
Gordon S, et al. High-density lipoprotein proteomics: identifying new drug targets and biomarkers by understanding functionality. Curr Cardiovasc Risk Rep. 2010b;4(1):1–8.
Guelzim N, et al. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet. 2002;31(1):60–3.
Hartwell LH, et al. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl): C47–52.
Heller M, et al. Mass spectrometry-based analytical tools for the molecular protein characterization of human plasma lipoproteins. Proteomics. 2005;5(10):2619–30.
Hu Z, et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 2009;37(Web Server issue):W115–21.
Huang S. Rational drug discovery: what can we learn from regulatory networks? Drug Discov Today. 2002;7(20 Suppl):S163–9.
Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
Jeong H, et al. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2.
Kao HL, Gunsalus KC. Browsing multidimensional molecular networks with the generic network browser (N-Browse). Curr Protoc Bioinformatics. 2008;Chapter 9:Unit 9 11.
Karlsson H, et al. Lipoproteomics II: mapping of proteins in high-density lipoprotein using two-dimensional gel electrophoresis and mass spectrometry. Proteomics. 2005;5(5):1431–45.
Keshava Prasad TS, et al. Human protein reference database–2009 update. Nucleic Acids Res. 2009;37(Database issue):D767–72.
King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20(17):3013–20.
Lewis GF, Rader DJ. New insights into the regulation of HDL metabolism and reverse cholesterol transport. Circ Res. 2005;96(12):1221–32.
Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science. 2002;296(5569):910–13.
Mineo C, et al. Endothelial and antithrombotic actions of HDL. Circ Res. 2006;98(11):1352–64.
Naqvi TZ, et al. Evidence that high-density lipoprotein cholesterol is an independent predictor of acute platelet-dependent thrombus formation. Am J Cardiol. 1999;84(9):1011–17.
Negre-Salvayre A, et al. Antioxidant and cytoprotective properties of high-density lipoproteins in vascular cells. Free Radic Biol Med. 2006;41(7):1031–40.
Nofer JR, et al. HDL and arteriosclerosis: beyond reverse cholesterol transport. Atherosclerosis. 2002;161(1):1–16.
O'Madadhain J, et al. The JUNG (Java Universal Network/Graph) framework. Technical Report UCI-ICS 03-17. Irvine: UC Irvine; 2003.
Palla G, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435(7043):814–18.
Peregrin-Alvarez JM, et al. The modular organization of protein interactions in *Escherichia coli*. PLoS Comput Biol. 2009;5(10):e1000523.
Ravasz E. Detecting hierarchical modularity in biological networks. Methods Mol Biol. 2009;541:145–60.
Rezaee F, et al. Proteomic analysis of high-density lipoprotein. Proteomics. 2006;6(2):721–30.
Rual JF, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437(7062):1173–8.

Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003;100(21):12123–8.
Stuart JM, et al. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–55.
Vaisar T, et al. Shotgun proteomics implicates protease inhibition and complement activation in the antiinflammatory properties of HDL. J Clin Invest. 2007;117(3):746–56.
Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci. 2001;268(1478):1803–10.
Wang L, et al. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. Hum Brain Mapp. 2009;30(2):638–49.
Watson AD, et al. Protective effect of high density lipoprotein associated paraoxonase. Inhibition of the biological activity of minimally oxidized low density lipoprotein. J Clin Invest. 1995;96(6):2882–91.
Wishart DS, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–72.
Wu J, et al. Integrated network analysis platform for protein-protein interactions. Nat Methods. 2009;6(1):75–7.
Xenarios I, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30(1):303–5.
Yip KY, et al. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics. 2006;22(23):2968–70.
Yu H, et al. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3(4):e59.
Zhang M, et al. Molecular network analysis and applications. In: Alterovitz G, Ramoni M, editors. Knowledge-based bioinformatics: from analysis to interpretation. Chichester: Wiley; 2010.
Zhang M, et al. Prediction and analysis of the protein interactome in *Pseudomonas aeruginosa* to enable network-based drug target selection. PLoS One. 2012;7(7):E41202.