The Structure and Function of Biological Networks

A Thesis

Submitted to the Faculty

of

Drexel University

by

Daniel Duanqing Wu

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

June 2010

To my parents, my wife, and my children.

Acknowledgments

First and foremost, I am very grateful to my advisor Dr. Xiaohua (Tony) Hu from the bottom of my heart. Without his enduring support, encouragement, and guidance, this thesis would never become a reality. Tony has provided me intellectual freedom, deep insights, and patient guidance in my research. He has helped me open a door to study theoretical and technical details in data mining and bioinformatics. I have been privileged and fortunate to have him as my advisor.

I would also like to express my gratitude to the College of Information Science and Technology for providing me the opportunity to fulfill my dream of pursuing a Ph.D. degree here at Drexel.

My gratitude extends to Drs. Prudence W. Dalrymple, Jiexun Jason Li, Aleister J. Saunders, and Christopher C. Yang for their advice, time and efforts serving on my thesis committee.

Thanks also go to Drs. Chaomei Chen and Rosina Weber for their advice and support serving on my Portfolio Review Committee in the beginning years of my graduate studies at Drexel. I am very grateful to Dr. Weber for allowing me to work as her research assistant for two quarters.

I want to thank my friends and fellow graduate students at Drexel, especially Xiahua Zhou and Xiaodan Zhang, for the friendship and support.

The research related to this thesis was supported in part by NSF Career Grant IIS 0448023, NSF CCF 0514679, Pennsylvania Department of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and Pennsylvania Department of Health Grant (No. 239667). The College of Information Science and Technology generously provided financial support allowing me to attend several academic conferences over the years.

To my family, I thank them for their patience, their encouragement, and their love. I would not have made it this far in my life if it were not for the sacrifice and hard work of my parents. My deepest appreciation also goes to my parents-in-law for the trust they have conferred upon me and the support they have provided over the years. I am eternally indebted to my wife for her love and being there for me for almost two decades. My wonderful children, Nicholas and Jacqueline, are the sources of my inspiration and happiness. This thesis is dedicated to them.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Structure and Function of Biological Networks
Daniel Duanqing Wu
Xiaohua Hu

Biology has been revolutionized in recent years by an explosion in the availability of data. Transforming this new wealth of data into meaningful biological insights and clinical breakthroughs requires a complete overhaul both in the questions being asked and the methodologies used to answer them. A major challenge in organizing and understanding the data is the ability to define the structure in biological systems, especially high level structures. Networks are a powerful and versatile tool useful in bridging the data and the complex biological systems. To address the importance of the higher-level modular and hierarchical structure in biological networks, we have investigated in this thesis the topological structure of protein-protein interaction networks through a comprehensive network analysis using statistical and computational techniques and publicly available protein-protein interaction data sets. Furthermore, we have designed and implemented a novel and efficient computational approach to identify modules from a seed protein. The experiment results demonstrate the efficiency and effectiveness of this approach in finding a module whose members exhibit high functional coherency. In addition, toward quantitative studies of protein translation regulatory networks (PTRN), we have developed a novel approach to reconstruct the PTRN through integration of protein-protein interaction data and Gene Ontology annotations. We have applied computational techniques based on hierarchical random graph model on these reconstructed PTRN to explore their modular and hierarchical and to detect missing and false positive links from these networks. The identification of the high order structures in these networks unveils insights into their functional organization.

CHAPTER 1. INTRODUCTION

Networks are a natural, powerful, and versatile tool for representing the structure of complex systems and have been widely used in many disciplines, ranging from sociology to physics to biology [Strogatz 2001, Newman 2001, Girvan and Newman 2002, Newman 2003a]. Examples of such complex networks include those of personal or social contacts in sociology and epidemiology, citation networks underlying collections of published papers in information science, the Internet and the World Wide Web in computer science and information technology, and a growing number of biological networks, such as protein-protein interaction networks, metabolic networks, and genetic regulatory networks.

A network is a collection of network components representing its fundamental units and a set of connections featuring the relationship between these components. Traditionally, complex networks have been described by graph theory, in which the network components are represented by nodes (or nodes) and their relations by edges. Over the past decade, there have been strong interest and attention devoted toward better understanding the infrastructure underlying complex networks, particularly their topologies and the large-scale properties that can be derived. It is a fundamental belief that network functions are affected by network structures.

The functioning of complex biological systems demands the intricate coordination of various cellular processes and their participating components. Advances in technology and several crucial biological discoveries (such as the sequencing of human genome) have allowed experimental biology to provide much more detailed descriptions of these processes and components, resulting in large amount of biological data. This huge amount of data accompanied by the expectation that these data will provide detailed understanding of cellular processes has been driving the excitement in today's biology and computer science. One of the great tasks in the post-genomic era is to organize, digest these data, and ultimately translate this new wealth of data into meaningful biological insights and clinical breakthroughs [Kitano 2002].

However, studying complex biological networks has proved to be significantly challenging. Although useful, the data generated from high-throughput techniques are often incomplete and contains un-ignorable errors. Complicated with their large sizes, biological networks do not lend themselves to direct inspection and offer no single place from which a complete picture of topology can be obtained. As a consequence, the topology is often inferred from appropriate network measurements through various sophisticated approaches, each having its own strengths, weaknesses, and resulting in a distinct view of the network topology. Furthermore, the intrinsic dynamic and evolving nature of biological networks also makes the task more difficult.

The primary goal of this thesis is to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems, with focus on the modular and hierarchical architecture of biological networks, exemplified by protein translation regulatory networks (PTRN).

1.1 Biological networks

Biological networks have been used to model biological interactions at many different levels of detail, ranging from the atomic interactions in a folded protein structure to the relationship of organisms in a population or ecosystem. In the context of this thesis, we focus on molecular interaction networks when referring to biological networks.

The molecular interactions between individual constituents including genes, proteins, and metabolites are examined at the level of the cell, tissue, and organ to ultimately describe the entire organism or system. Therefore, biological networks provide an effective and important systems biology approach to understand how system properties emerge from these interactions.

In the multi-layered organization of organisms, molecular interactions form the bridge between individual molecules and macro-scale organization of the cell through functional modules [Oltvai and Barabási, 2002]. Biological networks representing these interactions may be in the form of metabolic networks, signal transduction pathways, genetic regulatory networks, and protein-protein interaction networks. These different types of networks provide complementary information useful in different contexts.

Metabolic networks have a relatively longer history compared to other biological networks [Jeong et al 2000]. They characterize the process of biochemical reactions performing a particular metabolic function. There have been successful attempts at modeling, synthesizing and organizing metabolic networks into public databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Goto 2000, Krishnamurthy et al 2003], in which the metabolic interactions are represented in terms of various binary relations.

Metabolic networks are chains of reactions linked to each other by chemical compounds (metabolites) through product-substrate relationships. A natural mathematical model for metabolic networks is a directed graph in which each node corresponds to a compound, and each edge corresponds to a reaction or an enzyme. The direction of an edge indicates whether the compound connected by the edge is a substrate or a product of the reaction/enzyme. It is also possible to replace this model by a directed graph if we are only interested in relations between enzymes. In such a model, the nodes of the graph represent the enzymes and a directed edge from one enzyme to another indicates that a product of the first enzyme is a substrate of the second.

Signal transduction pathways model another dynamic mechanism in which how biological information is transferred between and within cells [Weng et al 1999]. Cells are continually sensing and interacting with their environment, using signal transduction pathways and regulatory mechanisms to coordinate multiple functions so that they may respond and acclimate to an ever-changing environment [Ideker et al 2002].

Genetic regulatory networks, most frequently referred to as transcriptional regulatory networks or just genetic networks, represent regulatory interactions between pairs of genes and are generally inferred from gene expression data through microarray experiments. A simple and frequently used mathematical model for genetic regulatory networks is a Boolean network model in which nodes correspond to genes and a directed edge from one gene to the other represents the regulatory effect of the first gene on the second. The edge is often labeled by either a positive (+) or negative (-) sign to represent up- or down-regulation, respectively. More sophisticated models that capture the degree of regulation through weighted graphs and/or differential equations have also been proposed [Alm and Arkin 2003].

Proteins are executors of cellular functions. They play critical roles in cell structure, biochemical activity and dynamic behavior, mostly by interacting with other proteins. A better understanding of the protein-protein interaction networks is a crucial step toward deciphering the structure, function, and dynamics of biological systems [Ideker and Sharan 2008]. So far, protein-protein interactions represent the largest and most diverse data sets available [Uetz et al

2000, Walhout et al 2000, McGraith et al 2000, Rain et al 2001, Ito et al 2001, Ho et al 2002, Giot et al 2003, Li et al 2004]. Some of the data have been accumulated for decades obtained from dispersed literature done by small scale traditional laboratory experiments. The first large-scale maps were generated using yeast two-hybrid systems. High-throughput studies later used affinity purification followed by mass spectrometry. Various computational approaches have also been used to predict functional relations between proteins as well as physical protein-protein interactions [Bock and Gough 2001, Aloy and Russell 2002, Kim et al 2002, Aloy and Russell 2003, Han et al 2004, Huang et al 2004, Espadaler et al 2005, Ogmen et al 2005, Martin et al 2005, Pitre et al 2006, Najafabadi and Salavati 2008]. Therefore, PPI networks provide a natural starting point toward deciphering the structure and function of biological networks. In this thesis, I use protein-protein interactions as the primary data sets.

1.2 Network models

In order to capture features observed in real world networks, many theoretical network models have been proposed. These models play an important role in shaping our understanding of complex networks and help to explain the origin of observed network characteristics [Barabási and Oltvai 2004].

Each of these models is characterized by the way in which networks are created as well as by a few statistical features that networks display, such as degree distribution, average path length between pairs of nodes, and clustering coefficient. Among these models, there are four, namingly random networks, small-world networks, scale-free networks, and hierarchical networks, that have a direct impact on our understanding of biological networks and deserve a further introduction here.

Random networks are the simplest, straightforward, yet mathematically elegant realization of a complex network, first studied from a pure mathematics point of view by [Erdös and Rényi 1959]. In a random network of N nodes, each pair of nodes is connected by an edge with uniform probability p, resulting in a graph with approximately $p\frac{N(N-1)}{2}$ randomly placed edges. The node degrees follow a Poisson distribution, indicating that most nodes have approximately the same number of links. The mean path length in a random network is proportional to the logarithm of the network size, demonstrating the characterized small-world property.

The small-world model was first proposed by [Watts and Strogatz 1998] who started a large area of research related to the small-world topology. In a small-world network, even if it may have a large number of nodes, the typical distance between two nodes is very small. Small-world networks have low average path lengths and high clustering coefficients. The small-world property seems pervasive in almost all networks [Fell and Wagner 2000, Newman 2000, Wagner and Fell 2001].

Scale-free networks have attracted much attention for more than a decade [Barabási and Albert 1999]. It is a theoretic model characterized by a connectivity distribution which decays as a power-law. This feature is suggested to be a direct consequence of two generic mechanisms: network evolution and preferential attachment. In a scale-free network, there are large variations in the number of links per node but there are also a few nodes that have many links, forming the so-called "hubs". Like small-world networks, scale-free networks have low average path lengths and high clustering coefficients. It can be shown that a scale-free network has the small-world properties, but not all small-world networks are scale-free. Also, most biological networks approximate a scale-free topology, but not all biological networks in cell are scale-free [Pastor-Satorras and Vespignani 2001, Albert and Barabási 2002, Dorogovtsev and Mendes 2002].

Hierarchical networks describe the topology in which a root node (i.e. the top level) is connected to one or more other nodes that are one level lower in the hierarchy (i.e. the second level) with a link between each of the second level nodes and the top level root node, while each of the second level nodes that are connected to the top level root node also have one or more other nodes that are one level lower in the hierarchy (i.e., the third level) connected to it.

The modular structure has been found in many biological networks [Ravasz et al 2002, Ravasz and Barabási 2003, Holme et al 2003]. Modules (or communities, a term often used in sociology) can be loosely defined as groups within a network where links within modules are much denser than those across modules. Studying the modularity of a large network is potentially very useful [Lu et al 2006, Wang and Zhang 2007]. It not only provides structural information about the network, but also reveals the underlying mechanisms that determine the network structure and dynamics.

While modular structure in a network concerns mostly the grouping (clustering) of network components, hierarchical structure goes beyond simple grouping by including organization at all scales in the network. Cellular functions have been widely believed to be organized in a modular manner hierarchically where each module at each hierarchical level performs a relatively independent task.

The study of the modular and hierarchical structure in networks has received a lot of attention in recent years. Yet, many questions still remain to be answered.

1.3 Motivations

With the goal of deciphering and utilizing the structure of biological networks, this thesis attempts to address the following research questions:

1) What are the modular and hierarchical structures of biological networks?

2) What are the efficient computational approaches (or How) to determine these structures?
3) What are the biological significances of these structures?

Network topology plays an important role in understanding network structure and performance. Several of the most important and commonly used topological features include degree, clustering coefficient, and average path length. As the first step in this endeavor, I will perform a comprehensive comparative study of these global topological features on protein-protein interaction networks from different species. I will use computational and statistical techniques and pre-existing data from publicly available sources.

Protein translation is a vital cellular process for any living organism. There has been an extensive effort using computational methods in deciphering the transcriptional regulatory networks. However, research on translation regulatory networks has caught little attention in the bioinformatics and computational biology community probably due to the nature of available data and the bias of the conventional wisdom. In this thesis, I will reconstruct protein translation networks in yeast and perform a global network analysis of these reconstructed networks [Wu and Hu 2006c, Wu and Hu 2007]. This work attempts to facilitate the elucidation of the structure and properties of translation networks.

Modular structure is a property common to many networks. This structure results from modules of densely connected nodes within a network.

Module discovery in biological networks may help us better understand the organizational principles of the biological systems. By separating the network into modules which may be functional groups could simplify the functional analysis considerably. Many methods have been proposed for module detection based on a variety of distinct approaches. However, most of the approaches either suffer from high computational cost or sacrificed quality. I have developed and implemented a novel algorithm for fast module identification by hierarchical growth [Wu and Hu 2005]. Since we deal with a data set with incompleteness and noise, I will also explore using hierarchical random graph model to infer the modular and hierarchical structures of PPI networks.

CHAPTER 2. BACKGROUND AND LITERATURE REVIEW

In order to discuss networks formally, I first present the notion of a graph and an introduction of the most basic graph theoretic concepts in this chapter. The notations and naming conventions used in this thesis, however, are by no means the rule or unequivocal. One may often see other different notations in the literature.

2.1 Graph theoretic definitions

A *network* can be described in more formal mathematical language as a *graph*. However, we will use both terms interchangeably in this thesis. First, we define some basic concepts of graph theory that will allow us to compare and characterize different complex networks.

The graph representation can be differentiated based on the level of organization. In the context of biological networks, the graph nodes are the network components which can be proteins, genes, metabolites, or modules. An edge between two nodes indicates an interaction between the corresponding molecules.

A graph is usually denoted with $G(V,E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges connecting the nodes. An edge $uv \in E$ is a link between the pair of nodes u and v where $u,v \in V$.

The most common type of graph is called a *simple* graph. In simple graphs, there is at most one edge between any two nodes. If multiple edges are allowed between nodes, the graph is known as a *multi-graph*. A self-loop is an edge whose end nodes are the same node. A graph that may contain multiple edges and graph loops is called a *pseudo* graph. In this thesis, I will mainly use simple graphs.

A *sub-graph* of a graph *G* is a graph whose node and edge sets are subsets of those of *G*. A *super-graph* of a graph *G* is a graph that contains *G* as a *sub-graph*.

A graph is *directed* if its edges are directed (pointing toward either one of the ends) and *undirected* otherwise. A graph is *complete* (or called a *clique*) if every node has a connecting edge to every other node. The complete graph of *n* nodes is often denoted by k_n where k_n would have $\frac{n(n-1)}{2}$ edges.

Nodes that share a common edge are adjacent. The degree of a node is the number of edges incident with it, i.e. a measure of immediate adjacency. In directed graphs, the in-degree of a node is the number of edges ending at the node, whereas the out-degree is the number of edges beginning at the node. A node of degree zero is an *isolated* node. If the edge set is finite, then the total sum of node degrees is equal to twice the number of edges. A *degree sequence* is a list of degrees of a graph in non-increasing order. A sequence of non-increasing integers is *realizable* if it is a degree sequence of some graph.

The degree distribution, defined as the fraction of nodes in the network that have a specified degree, is a function describing the probability of a node having the specified degree.

The set of neighbors, called the (*open*) *neighborhood* $N_G(v)$, for a node *v* in a graph *G*, consists of all nodes adjacent to *v* but not including *v*. When *v* is also included, it is called a *closed neighborhood*, denoted by $N_G[v]$. In this thesis, when stated without any qualification, a neighborhood is assumed to be open.

The *diameter* of a network, <*l*>, is defined as the average distance between any two nodes. The *distance* between two nodes is defined as the number of edges along their shortest path.

The *clustering coefficient* is used to quantify the probability that an edge exists between two neighboring nodes of a node. For example, in a network, if node A and node C are neighbors of node B, i.e. B connects to both A and C, then the clustering coefficient defines how probable that node A directly connects to node C. Formally, the clustering coefficient is defined as

$$C_i = \frac{2e_i}{k_i(k_i - 1)},$$

where k_i is the number of neighbors of node *i*, e_i is the number of edges connecting these neighbors, and $\frac{k_i(k_i - 1)}{2}$ denotes the maximum number of possible edges connecting these neighbors. A global measurement related to C_i

is the average clustering coefficient $\langle C \rangle$ over all nodes in the network, characterizing the overall tendency of nodes to form clusters or groups.

2.2 Topology of protein-protein interaction networks

In their pioneering work, [Barabási and Albert 1999] describe a highly heterogeneous protein-protein interaction (PPI) network with scale-free connectivity properties in yeast. The signature of scale-free networks, as opposing to random networks, is that the degrees (or connectivity) of nodes follow a power-law, i.e.

$$P(k) \approx k^{-\gamma},$$

where $P(k)$ is the probability of a node having a degree of k and $\gamma > 0$.

Power law feature has been observed in many networks, such as PPI networks of *S. cerevisiae*, *H. pylori*, *C.elegans*, and *D. melanogaster* [Walhout et al 2000, Li et al 2004, Giot et al 2003].

In PPI networks, not only the degree distribution exhibits power-law dependence, other topological properties have also been shown to be scale-free. One such property is the clustering coefficient. Yook and colleagues [Yook et al 2004] observe that the clustering coefficient of *S. cerevisiae* follows a power-law. [Ng and Huang 2004] confirms the popular scale-free topology across six different species based on degree distribution and diameter.

However, not all research agrees on the power-law behavior in all PPI networks. [Thomas et al 2003] find that the connectivity distribution in a human PPI network does not follow power law. They argue that current belief of power law distribution may reflect a behavior of a sampled sub-graph. Since we only have an incomplete and low coverage sample of an entire protein interactome, the behavior in a sampled sub-graph does not necessarily imply the same behavior for the whole graph. They call for the attention to the importance of assessing the accuracy of the observed degree distribution in reference to the full proteome. From a slightly different angle, [Tanaka et al 2005] report that some PPI networks do not follow power law if using a rank-degree plot instead of regularly used frequency-degree plot. [Colizza et al 2005] evaluate three PPI networks constructed from yeast data sets. Although they observe that the connectivity distribution follows power law, only one of the three networks approximates power law behavior for the clustering coefficient. [Soffer and Vazquez 2004] find that the power law dependence of the clustering coefficient is to some extent caused by the degree correlations of the networks, with high degree nodes preferentially connecting with low degree ones.

A more recent study by [Yang et al 2008] provides an improved method for extracting accurate topological information about real PPI networks from experimentally-obtained sub-networks. They found that random sampling of networks preserves topological information, regardless of the type of network analyzed. Their results indicate that the degree distribution of the original network may not be scale-free, but in fact exhibit an exponential distribution. As mentioned earlier, PPI data obtained from high-throughput techniques have unavoidable limitations including false positive, false negative, and assumed binary interactions [Gandhi et al 2006]. [Yang et al 2008] argue that these false positives may contribute to the observed power-law behavior of the PPI networks based on the following rationale: (i) the high confidence *Drosophila* network (purportedly containing fewer false positives [Bader et al 2004]) has a stronger exponential component (also observed by [Przulj et al 2004]); (ii) many proteins preferentially behave as either baits or preys but not both, suggesting an experimentally-introduced preferential attachment phenomenon (introduction of hubs by experimental bias) which, as shown by [Barabási and Albert 1999], is a key factor for occurrence of power-law distributions; and (iii) the degree distribution of a mammalian PPI network obtained by [Ma'ayan et al 2005] from the literature, which should have a much lower rate of false positives, exhibits an almost purely exponential distribution.

Therefore, even though some questions in this fundamental area have been addressed, many important ones still resist complete resolution. We will discuss a comprehensive evaluation of the topological structure of a variety of PPI networks in Chapter 3.

2.3 Protein translation regulatory networks (PTRN)

The central dogma of molecular biology describes that the genetic information is transferred from DNA to mRNA through transcription and from mRNA to protein via translation. In every living organism, translation is a vital cellular process in which the information contained in the mRNA sequence is translated into the corresponding protein by the complex translation machinery.

There are three major steps in protein biosynthesis: initiation, elongation, and termination. Initiation is a series of biochemical reactions leading to the binding of ribosome on the mRNA and the formation of the initiation complex around the start codon [Pain 1996]. This process involves various regulatory proteins (the so-called initiation factors). Eukaryotic protein synthesis exploits various mechanisms to initiate translation, including cap-dependent initiation, re-initiation, and internal initiation. For the majority of mRNAs in the cell, translation is carried out through the cap-dependent pathway. Although debatable, it is widely believed that some cellular mRNAs contain internal ribosome entry sites (IRES) and there exists a cap-independent, IRES mediated translation [Merrick 2004]. During elongation, codon-specific tRNAs are recruited by the ribosome to grow the polypeptide chain one amino acid at a time while the ribosome moves along the mRNA template (one codon at a time). This process also involves various elongation factors and proceeds in a cyclic manner. In termination phase, the termination codon is recognized by the ribosome. The newly synthesized peptide chain and eventually the ribosomes themselves are released.

Recent years have witnessed the breakthrough in high-throughput technologies that have been used in monitoring the various components of the transcription and translation machineries. DNA microarrays enable the estimation of the copy number for every mRNA species within a single cell and the changes in gene expression temporally or under different physiological conditions [Lockhart and Winzeler 2000]. Two-dimensional gel electrophoresis coupled with tandem mass spectrometry makes it possible to measure simultaneously specific protein levels for thousands of proteins in the cell. These high-throughput technologies and the success of several genome projects are rapidly generating an enormous amount of data about genes and proteins that govern such cellular processes as transcription and translation. Analyzing these data is providing new insights into the regulatory mechanisms in many cellular systems. One of the major goals in post-genomic era is to elucidate in a holistic manner the mechanisms by which sub-cellular processes at the molecular level are manifest at the phenotypic level under physiological and pathological conditions.

The complexity and the large sizes of the transcription and translation machineries make computational approaches attractive and necessary in facilitating our understanding the design principles and functional properties of these cellular systems. Transcriptional regulation, used by cells to control gene expression, has been a focus in a variety of computational methods to infer the structure of genetic regulatory networks or to study their high level properties [de Jong 2002]. However, research on translational regulatory networks has caught little attention in the bioinformatics and computational biology community. This contrast may partly due to two factors. Firstly, transcriptional control, other than translational control, has long been regarded by conventional wisdom as the primary control point in gene expression. Secondly, the success of genome projects and the application of high-throughput technologies provide tremendous amount of data about transcriptional regulation that are readily available for computational analysis. On the contrary, data about translational control are still very limited and probably too specialized so that they are consumed primarily by biologists.

Proteins, rather than DNAs or mRNAs, are the executors of the genetic program. They provide the structural framework of a cell and perform a variety of cellular functions such as serving as enzymes, hormones, growth factors, receptors, and signaling intermediates. Biological and phenotypic complexity eventually derives from changes in protein concentration and localization, post-translational modifications, and protein-protein interactions. Expression levels of a protein depend not only on transcription rates but also on such control mechanisms as nuclear export and mRNA localization, transcript stability, translational regulation, and protein degradation. Results from biological research have demonstrated that translational regulation is one of the major mechanisms regulating gene expression in cell growth, apoptosis, and tumorigenesis [Holland 2004]. Therefore, study of protein translation networks, especially from systems biology perspective, may provide new insights into our understanding of this important cellular process.

Very limit work has been done in this regard. Mehra and colleagues [Mehra et al 2003] develop a genome-wide model for the translation machinery in *E. coli* that provides mapping between changes in mRNA levels and changes in protein levels in response to environmental or genetic perturbations. They also propose a mathematical and computational framework [Mehra and Hatzimanikatis 2006] that can be applied to the analysis of the sensitivity of a translation network to perturbation in the rate constants and in the mRNA levels in the system.

However, much more research is needed in this area. Towards the goal of understanding how translation machinery functions from a system's perspective, it is imperative that we have a better understanding of the global properties of protein translation regulatory networks, especially integrated with functional perspectives. We will present our work on reconstructing and analyzing protein translation regulatory networks in Chapter 5 and Chapter 6.

2.4 Identifying network modular structure

The study of modular (or community, both terms will be used interchangeable in this thesis) structure in a network is not new. It is closely related to the graph partitioning in graph theory and computer science as well as the hierarchical clustering in sociology [Newman 2003a]. However, recent years have witnessed an intensive activity in this field partly due to the dramatic increase in the scale of networks being studied. Many algorithms for finding communities in networks have been proposed. They can be roughly classified into two categories, divisive and agglomerative.

The divisive approach takes the route of recursive removal of nodes (or edges) until the network is separated into its components or communities, whereas the agglomerative approach starts with isolated individual nodes and joins together small communities.

One important algorithm is proposed by Girvan and Newman (the GN algorithm) [Girvan and Newman 2002]. The GN algorithm is based on the concept of betweenness, a quantitative measure of the number of shortest paths passing through a given node (or edge). The nodes (or edges) with the highest betweenness are believed to play the most prominent role in connecting different parts of a network. The GN algorithm detects communities in a network by recursively removing these high betweenness nodes (or edges). It has produced good results and is well adopted by different authors in studies of various networks [Newman 2003a], but has a major disadvantage which is its computational cost. For sparse networks with n nodes, the GN algorithm is of $O(n^3)$ time. Various alternative algorithms have been proposed [Newman 2004a, Newman 2004b, Newman and Girvan 2004, Donetti and Munoz 2004, White and Smyth 2005, Lancichinetti et al 2009, Mucha et al 2010], attempting to improve either the quality of the community structure or the computational efficiency of finding communities.

The GN algorithm has been applied to a number of metabolic networks from different organisms to detect communities that relate to functional units in the networks [Holme et al 2003]. It has also been adapted to analyze a network of gene relationships as established by co-occurrence of gene names in published literature and to detect communities of related genes [Wilkinson and Huberman 2004].

A slightly different approach is to identify community structure in protein-protein interaction network by growing from a given seed protein or proteins. This may also be used to answer question such as what is the community a given protein (or proteins) belongs to [Maraziotis et al 2007]. Due to the complexity and modularity of biological networks, it may be more feasible computationally to study a community containing one or a few proteins of interest. Hashimoto and colleagues [Hashimoto et al 2004] have used such an approach to growing genetic regulatory networks from seed genes. Their work is based on probabilistic Boolean networks and sub-networks are constructed in the context of a directed graph using both the coefficient of determination and the Boolean function influence among genes. The similar approach is also taken by Flake and colleagues [Flake et al 2002] to find highly topically related communities in the Web based on the self-organization of the network structure and on a maximum flow method.

Related works also include those that predict co-complex proteins. Jansen and colleagues [Jansen et al 2002] use a procedure integrating different data sources to predict the membership of protein complexes for individual genes based on two assumptions: first, the function of any protein complex depends on the functions of its subunits; and second, all subunits of a protein complex share certain common properties. Bader and Hogue [Bader and Hogue 2003] report a Molecular Complex Detection (MCODE) clustering algorithm to identify molecular complexes in a large protein interaction network. MCODE is based on local network density – a modified measure of the clustering coefficient. Bu and colleagues [Bu et al 2003] use a spectral analysis method to identify the topological structures such as quasi-cliques and quasi-bipartites in a protein-protein interaction network. These topological structures are found to be biologically relevant functional groups. In our previous work, we developed a spectral-based clustering method using local density and node neighborhood to analyze the chromatin network [Hu et al 2004, Hu 2005].

Additional works along this line of research are based on the concept of network modularity introduced by Hartwell and colleagues [Hartwell et al 1999]. The works by [Spirin and Mirny 2003] and [Rives and Galitski 2003] both used computational analyses to cluster the yeast protein-protein interaction network and discovered that molecular modules are densely connected with each other but sparsely connected with the rest of the network.

Comprehensive comparisons of different approaches to community structure identification or clustering in terms of robustness, sensitivity and computational cost can be found in [Danon et al 2005, Brohee and van Helden 2006].

Another active research frontier in this area is the identification of functional modules in PPI networks, who share common cellular function beyond the scope of classical pathways, by means of detecting differentially expressed regions in PPI networks. This requires on the one hand an adequate scoring of the nodes in the network to be identified and on the other hand the availability of an effective algorithm to find the maximally scoring network regions. [Ideker et al 2002] have proposed to identify interaction modules in this setting by devising firstly an adequate scoring function on networks and secondly an algorithm to find the high-scoring sub-networks. The underlying combinatorial problem has been proven to be NP-hard for additive score functions defined on the nodes of the network. [Ideker et al 2002] use a heuristic strategy based on simulated annealing and develop a score to measure the significance of a sub-network that includes the integration of multivariate P-values. This score has been extended by [Rajagopalan and Agarwal 2005] to incorporate an adjustment parameter in conjunction with a greedy search algorithm in order to obtain smaller subgraphs. This approach however, excludes the possibility to combine multiple P-values. Variants of greedy search strategies have also been used by [Sohler et al 2004, Cabusora et al 2005, Nacu et al 2007]. An alternative edge scoring based on correlation of gene expression has been proposed by [Guo et al 2007]. All the former methods are heuristic approaches that cannot guarantee to identify the maximally scoring subgraph. Some of these often computationally demanding approaches tend to deliver large high-scoring networks, which may be difficult to interpret. [Dittrich et al 2008] have proposed a new approach that is characterized by a modular scoring function, based on signal-noise decomposition implemented as a mixture model. This solution permits the smooth integration of multivariate P-values derived from various sources, delivers provably optimal and suboptimal solutions to the maximal scoring sub-graph problem by integer-linear programming (ILP) in reasonable running time, and allows to control the resultant sub-network size by an adjustment parameter, which is statistically interpretable as false-discovery rate (FDR).

We will present a novel and efficient approach for detecting protein modules in PPI networks in Chapter 4.

2.5 Hierarchical structure in complex networks

This line of research is somewhat related to work on the modular structure identification, however, at a higher level beyond the modular structure, dealing with uncovering some homogeneity in the heterogeneity of nodes in complex networks [Shen et al 2009]. From statistics perspectives, finite mixture distributions appear to be the framework to choose.

[Daudin et al 2008] propose a finite mixture model for random graphs and use an EM algorithm to estimate the parameters. The joint distribution of the random variables that describe each group to which a node belongs is approximated (in the E-step of EM algorithm) by the product of the conditional distribution of each variable given the rest. They resort to a heuristic criterion to obtain the number of groups. [Handcock et al 2007] report latent position cluster model where each node is assigned a latent position in a Euclidean space. They use two methods to determine the latent position and the cluster membership of the nodes, a two-stage maximum-likelihood estimation (EM at second stage) and a standard Markov chain.

To address the important questions such as “How can we tell what a network looks like, when we can’t actually look at it and don’t even have a clue what it looks like?” [Newman and Leicht 2007] propose a method using mixture model and EM to explore networks where nodes are classified to groups based on the observed patterns of connections between them.

Many studies indicate that there exists a high degree of clustering or modularity in biological networks. [Ravasz et al 2002] report that the network modules combined with each other in a hierarchical manner form a hierarchical network, which accounts for the coexistence of modularity, local clustering, and scale-free topology in the network. More evidences [Ravasz and Barabasi 2003, Barabasi and Oltvai 2004, Ma et al 2004] have been reported for the existence of hierarchical structure in many biological and non-biological networks. In these analyses, emphasis is put on detecting global signatures of a hierarchical architecture, such as the clustering coefficient. However, this scaling is neither necessary nor sufficient for a network to be hierarchical [Soffer and Vazquez 2004].

To assess whether a network is organized in a hierarchical architecture and to identify the different levels in the hierarchy, [Sales-Pardo et al 2007] propose an unsupervised method using hierarchical random graphs and show its ability of extracting the hierarchical organization of complex biological, social, and technological networks.

More recently, [Clauset et al 2008] present a technique, also using hierarchical random graphs, to infer hierarchical structure from network data. They use the hierarchical random graph model to gain insight into the structure of real networks. Starting with a real network, they estimate what dendrogram is most likely to fit that particular network. The parameters of the hierarchical graph model contain condensed information about the actual network. One caveat of this approach is a high number of variables that require fitting. In a model with $|V|$ nodes, this model requires ($|V|$-1) probability fittings. To solve this problem, they determine the right model by using Monte Carlo sampling of hierarchical random graphs with a probability proportional to the likelihood that the model results in the observed network.

By analyzing three real networks, the metabolic network of *Treponema pallidum*, a network of associations between terrorists, and a food web of grassland species, [Clauset et al 2008] demonstrate that their method can detect the hierarchical structure in these real-world networks and the existence of hierarchy can simultaneously explain and quantitatively reproduce many commonly observed topological properties of networks. Additional features of this method include its robustness against noisy data and its ability to predict missing connections in partly known networks with high accuracy by using knowledge of hierarchical structure.

In Chapter 6, we will discuss our work on analyzing the hierarchical structure and predicting missing links in reconstructed protein translation regulatory networks.

CHAPTER 3. MINING AND ANALYZING THE TOPOLOGICAL STRUCTURE OF PROTEIN-PROTEIN INTERACTION NETWORKS

3.1 Introduction

Proteins are important players in executing the genetic program. When carrying out a particular biological function or serving as molecular building blocks for a particular cellular structure, proteins rarely act individually. Rather, biological complexity is encapsulated in the structure and dynamics of the combinatorial interactions among proteins (as well as other biological molecules) at different levels, ranging from biochemical pathways to ecological phenomena [Barabási and Oltvai 2004]. Therefore, one of the key challenges in the post genomic era is to understand these complex molecular interactions that confer the structure and dynamics of a living cell.

The development of high-throughput data collection techniques has generated tremendous amount of data about protein-protein interactions (PPI). This provides a rich data source for further investigations including those employing computational approaches in an attempt to understand and model the structure and dynamics of biological systems [Bork et al 2004].

Modeling protein-protein interactions often takes the form of graphs or networks, where nodes represent proteins and edges represent the interactions between pairs of proteins. Research on such networks so far has revealed a number of distinctive topological properties, including the "small world effect", the power-law degree distribution, and clustering (or network transitivity), and the community structure [Girvan and Newman 2002]. These properties are shared by many biological networks and appear to be of biological significance. Examples of such biological relevance include the correlation between gene knock-out lethality and the connectivity of the encoded protein [Jeong et al 2001], and between the evolutionary conservation of proteins and the connectivity [Fraser et al 2002, Fraser et al 2003, Wuchty 2004]. Consequently, topological information has been exploited in the predictive functional assignment of uncharacterized proteins and the theoretical modeling for the evolution of PPI networks [Pei and Zhang 2005, Bu et al 2003, Hu et al 2004, Hu 2005, Valente and Cusick 2006].

Different data sets of protein-protein interactions, however, contain information gathered from different experimental systems where interactions are detected under different conditions. One caveat is that there is a surprisingly small overlap among different data sets [Deane et al 2002]. Moreover, these data sets are constantly being updated. Therefore, it is important to evaluate these available PPI data sets and to validate any conclusions drawn from these data.

In this chapter, we present a comprehensive evaluation of the topological structure of PPI networks across different species, with different confidence levels, and from different experimental systems.

3.2 Method

We represent a network as a simple graph, meaning that it is undirected, unweighted, and without self-loops. Each node of the graph represents a protein and each edge represents an interaction between the two proteins connected by it.

3.2.1 Data sets

We analyze the topology of three sets of PPI networks.

1) The species-specific set includes *E. coli, H. pylori, S. cerevisiae, D. melanogaster, C. elegans, M. musculus*, and *H. sapiens* PPI networks. The data sets were downloaded from the Database for Interacting Proteins (DIP) [Xenarios et al 2000].

2) The experimental systems-specific set includes fly and yeast PPI networks. The data sets were downloaded from the General Repository for Interactions Datasets (GRID) [Breitkreutz et al 2003]. From fly data set, we constructed three individual PPI networks, each representing the protein interactions detected by one of the following experimental systems: Enhancement, Suppression, and Two Hybrid. From yeast data set, three individual PPI networks constructed represent three different experimental systems: Affinity Precipitation, Synthetic Lethality, and Two Hybrid. For each data set, we also constructed a network representing the entire set of protein interactions.

3) The third set contains PPI networks with different confidence levels. The data set was downloaded from the Biomolecular Interaction Network Database (BIND) [Bader et al 2003]. We used the fly data set to build three PPI networks: the first one containing interactions with confidence score >= 0.5 (high confidence), the second on with confidence score >= 0.3 (medium confidence), and the third one containing all interactions.

3.2.2 Measurements of network topology

We measure the basic topological properties of each PPI network, including:

- The number of proteins, measured by the number of nodes.
- The number of interactions, measured by the number of edges.
- The number of connected nodes within the network.
- The size of the largest (or giant) component, measured by the size of the largest connected sub-graph.

We also measure three degree related metrics:

- The maximum degree ($K_{\max}$).
- The average degree ($< k >$), defined as

$$< k >= 2|E|/|V|,$$

Where $|E|$ is the total number of edges and $|V|$ is the total number of nodes.

- The degree distribution ($P(k)$) which defines the frequency of a node in the network with degree *k*.
- The diameter of a network, $< l >$, defined as the average distance between any two nodes. The distance between two nodes is defined as the number of edges along their shortest path.
- The clustering coefficient C_i, defined as

  $$C_i = \frac{2e_i}{k_i(k_i - 1)},$$

  where e_i is the number of edges connecting the neighbors of node i and $k_i(k_i - 1)/2$ denotes the maximum number of possible edges connecting these neighbors.
- The average clustering coefficient, defined as

  $$\langle C \rangle = \frac{\sum_i C_i}{|V|},$$

  where $|V|$ is the total number of nodes in the network. Assuming the same degree distribution, we use the following to obtain an average clustering coefficient of a random network [Newman 2003b]:

  $$\langle C_{rand} \rangle = \frac{(\langle k \rangle^2 - \langle k \rangle)^2}{(n \langle k \rangle^3)}.$$

We also calculate a local property called node density $< D >$. The definition of node density is inspired by [Bader and Hogue 2003] who define a local density by expanding the definition of the clustering coefficient for node *i* to include *i* itself in the formula when calculating C_i.

All statistical analyses are performed using SPSS software package.

3.3 Results

3.3.1 Basic properties of the PPI networks

Table 3-1, Table 3-2, and Table 3-3 list the basic properties of all PPI networks used for our analysis. The sizes of networks vary significantly across species, indicating the varied status in data collecting and documenting for the specific data source and virtually our understanding of PPI for these organisms. Table 3-1 shows the small sizes of so called giant components for *H. sapiens* and especially for *M. musculus*, meaning that we have a fairly large number of unconnected small sub-graphs in these two networks. As one can expect, the size of the giant component decreases in higher confidence networks while the number of unconnected sub-graphs increases.

Table 3-1 PPI networks of different species.

Species	Proteins	Interactions	Components	Giant Component(*)
E. coli	1640	6658	200	1396 (85.1%)
H. pylori	702	1359	9	686 (97.7%)
S. cerevisiae (Core)	2614	6379	66	2445 (93.5%)
D. melanogaster	7441	22636	52	7330 (98.5%)
C. elegans	2629	3970	99	2386 (90.8%)
M. musculus	327	274	79	49 (15.0%)
H. sapiens	1059	1318	119	563 (53.2%)

*Number inside the parenthesis: percentage of the size of the giant component in the entire network.

Table 3-2 PPI networks of different confidence levels.

Network	Confidence	Proteins	Interactions	Components	Giant Component
Fly00	> 0	7064	21111	68	6929 (98.1%)
Fly30	>= 0.3	6382	9157	213	5881 (92.1%)
Fly50	>= 0.5	4689	4877	590	3068 (65.4%)

Table 3-3 PPI networks of different experimental systems.

Network	Experimental Systems	Proteins	Interactions	Components	Giant Component
Fly	Combined	7938	25827	72	7793 (98.2%)
Fly-E	Enhancement	1054	1819	56	902 (85.6%)
Fly-S	Suppression	1121	2247	44	1020 (91.0%)
Fly-TH	Two Hybrid	5614	17544	12	5591 (99.6%)
Yeast	Combined	4918	18119	48	4824 (98.1%)
Yeast-AP	Affinity Precipitation	2388	7405	39	2292 (96.0%)
Yeast-SL	Synthetic Lethality	1468	4773	44	1343 (91.5%)
Yeast-TH	Two Hybrid	3937	6358	138	3632(92.3%)

3.3.2 Average global topological properties of PPI networks

In Table 3-4, we report the average global topological properties of PPI networks. Across species, PPI networks all exhibit small values of average degree and diameters, even though the absolute values differ significantly. Also, except for *C. elegans*, PPI networks for all other species have larger average clustering coefficient comparing to the corresponding random clustering coefficient, indicating a non-random and hierarchical structure within these networks.

As shown in Table 3-4, networks with higher confidence level have higher diameters, higher average clustering coefficient, a lower average degree, and shifting further away from random structure. Because we can reasonably assume that the higher the confidence level, the closer a proposed networks to the real one, this makes it plausible to postulate the presence of organizational architecture in PPI networks.

Also shown in Table 3-4 is the significant impact of different experimental systems on the topological structure of the resulting networks. In fly data set, PPI networks obtained from “enhancement” and “suppression” systems have an average clustering coefficient dramatically larger than that of networks built from “two hybrid” system. Similar results are also shown in yeast PPI networks, with significantly smaller average clustering coefficient for “two hybrid” networks. The diameters of networks tend to be insensitive to differed experimental systems.

Contrary to the average clustering coefficient, the average node density shows much lesser variability across species. It is less susceptible to the changes in confidence levels and in experimental methods.

Table 3-4 Average global topological properties of PPI networks.

Network	K_{max}	<*k*>	<*l*>	<*D*>	<*C*>	<C_{rand}>
E. coli	152	8.12	3.73	0.7053	0.5889	0.1168
H. pylori	54	3.87	4.14	0.4514	0.0255	0.0403
S. cerevisiae (Core)	111	4.88	5.00	0.5609	0.2990	0.0103
D. melanogaster	178	6.08	4.39	0.3920	0.0159	0.0097
C. elegans	187	3.02	4.81	0.4885	0.0490	0.0462
M. musculus	12	1.68	3.57	0.6082	0.1011	0.0062
H. sapiens	33	2.49	6.80	0.5703	0.1658	0.0098
Fly00	178	5.98	4.45	0.4002	0.0281	0.0095
Fly30	59	2.87	7.06	0.4989	0.0518	0.0015
Fly50	42	2.08	9.42	0.5636	0.0793	0.0008
Fly	178	6.51	4.39	0.4077	0.0675	0.0104
Fly-E	110	3.45	4.44	0.6206	0.3441	0.0725
Fly-S	124	4.01	4.30	0.6004	0.3459	0.0735
Fly-TH	144	6.25	4.23	0.3813	0.0093	0.0123
Yeast	288	7.37	4.12	0.4659	0.1538	0.0240
Yeast-AP	69	6.20	4.43	0.5177	0.2646	0.0163
Yeast-SL	157	6.50	3.84	0.5402	0.2324	0.1600
Yeast-TH	288	3.23	4.96	0.5372	0.0869	0.0368

3.3.3 Degree distribution

Degree distribution, $P(k)$, is the probability that a selected protein has exactly degree *k*. We evaluate the distribution of degrees $P(k)$ as a function of *k*. Figure 3-1, Figure 3-2, and Figure 3-3 show the degree distribution for all the networks we evaluate. The log-log plot clearly demonstrates the power law dependence of $P(k)$ on degree *k*. For our analysis, we select to use directly the raw data, instead of following [Jeong et al 2001] with exponential cutoff. The results of statistical analysis are listed in Table 3-5. Without exponential cutoff, our regression analysis yields power-law exponents γ between 1.31 and 2.76, in fairly good agreement with previously reported results.

Even though the regression analysis and figures clearly show strong power-law degree distribution, we want to conduct further statistical analysis to test if the power law model adequately captures all the features in the testing data. Using SPSS software package, we create a scatter plot of residues by fit values for the power law model. The result is shown in Figure 3-4, which clearly indicates a pattern in the data that is not captured by the power law model. This means that the power law is a model that has excellent fit statistics, but has poor residuals, indicating its inadequacy.

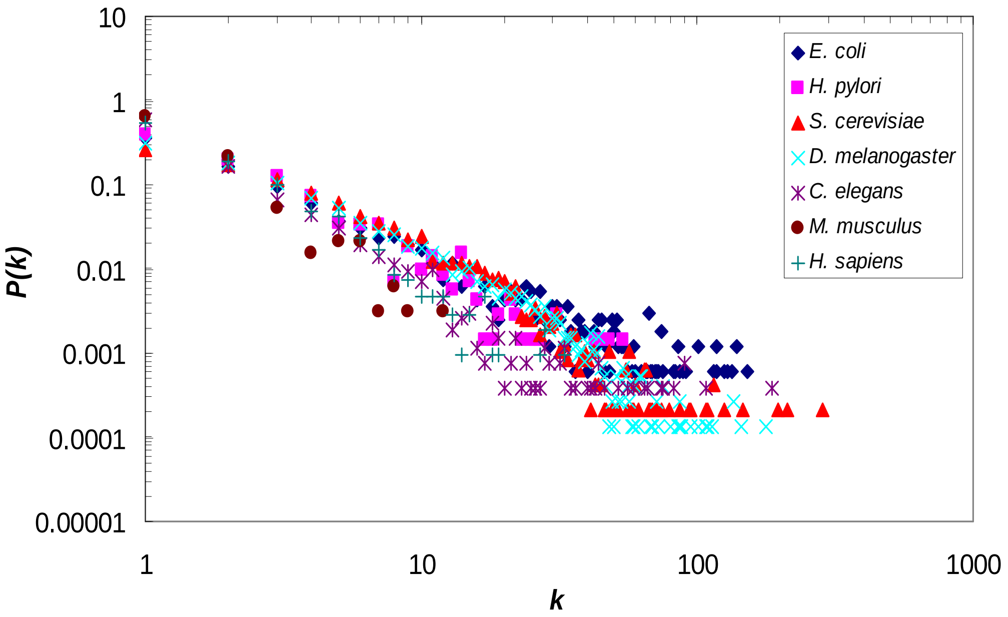

Figure 3-1 Degree distribution of PPI networks of different species.

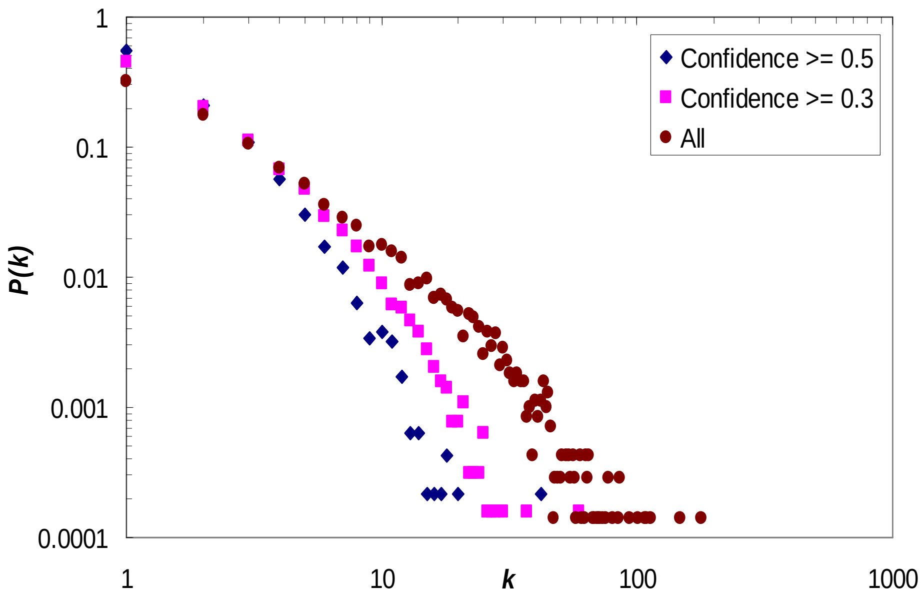

Figure 3-2 Degree distribution of PPI networks of different confidence levels.

3.3.4 The average clustering coefficient distribution

We have shown results of average clustering coefficient for PPI networks in a previous section. The clustering coefficient spectrum has been used to characterize quantitatively the hierarchical organization of the network structure [Colizza et al 2005]. We now take a closer look at the distribution of the clustering coefficient by averaging the clustering coefficient over nodes with degree k

$$C(k) = \frac{\sum_i C_i \delta_{k_i,k}}{n_k},$$

where n_k is the number of proteins with degree *k* and $\delta_{k_i,k}$ is the discrete delta function.

The results, as shown in Figure 3-5, Figure 3-6, and Figure 3-7, indicate that while *E. coli* and *S. cerevisiae* (also shown in Table 3-4) PPI networks show somewhat weak power law distribution, networks of other species do not follow a power law. Different experimental systems and different confidence levels do not seem to change this non-scale-free behavior.

3.3.5 The average node density distribution

Finally, we evaluate the distribution of the average node density over the nodes with degree *k*. The results for the node density spectrum ($D(k)$ over degree *k*) display consistent power law behavior for all the networks (Figure 3-8 for different species, Figure 3-9 for different experimental systems, and Figure 3-10 for different confidence levels).

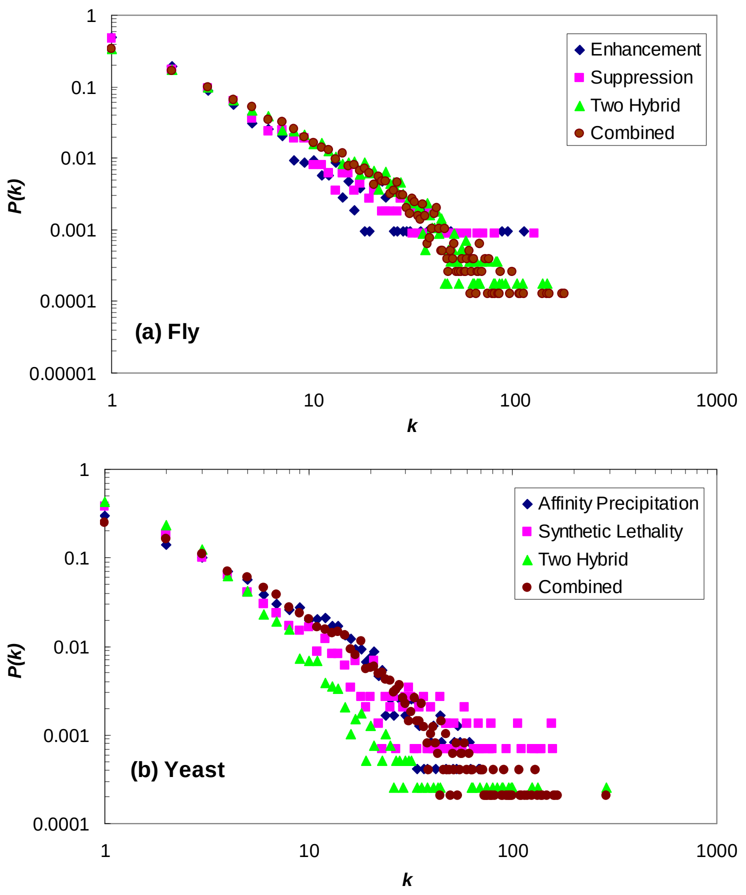

Figure 3-3 Degree distribution of PPI networks of different experimental systems.

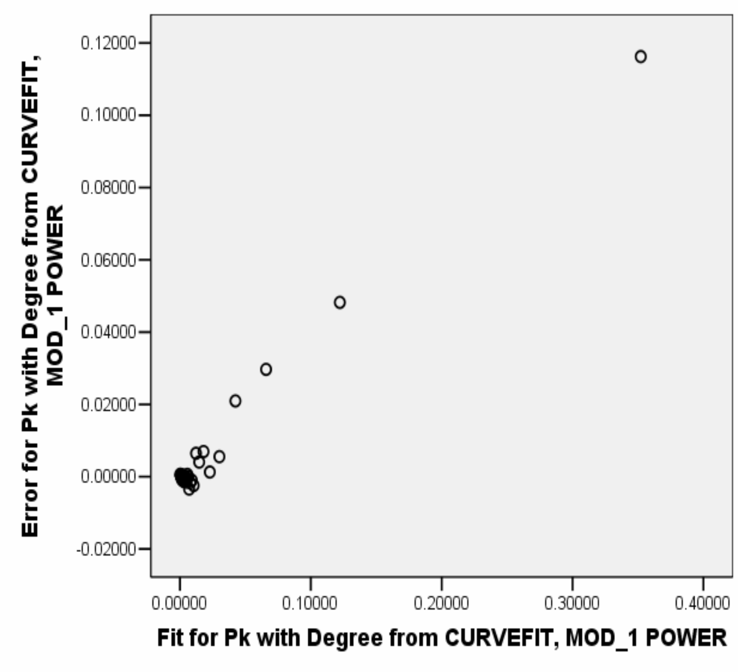

Figure 3-4 Residuals vs fit values.

Table 3-5 Statistical analysis of PPI networks.

Networks	γ† (R^2)	α† (R^2)	β† (R^2)
E. coli	1.355 (0.882)	0.562 (0.656)	0.536 (0.756)
H. pylori	1.651 (0.899)	0.495 (0.373)	0.826 (0.985)
D. melanogaster	1.945 (0.923)	3.050 (0.311)	0.836 (0.989)
S. cerevisiae (Core)	1.977 (0.911)	0.893 (0.721)	0.759 (0.867)
C. elegans	1.599 (0.839)	0.625 (0.362)	0.833 (0.976)
M. musculus	2.360 (0.931)	0.598 (0.431)	0.689 (0.965)
H. sapiens	2.025 (0.931)	0.657 (0.190)	0.626 (0.699)
Fly00	1.980 (0.930)	0.382 (0.194)	0.789 (0.913)
Fly30	2.540 (0.931)	0.698 (0.265)	0.780 (0.918)
Fly50	2.763 (0.915)	0.791 (0.375)	0.783 (0.920)
Fly	1.947 (0.934)	0.555 (0.334)	0.758 (0.865)
Fly-E	1.518 (0.858)	1.020 (0.539)	0.769 (0.886)
Fly-S	1.527 (0.936)	0.879 (0.513)	0.747 (0.893)
Fly-TH	1.912 (0.923)	0(0)	0.783 (0.867)
Yeast	1.761 (0.919)	0.752 (0.326)	0.728 (0.698)
Yeast-AP	1.819 (0.904)	0.635 (0.301)	0.619 (0.664)
Yeast-SL	1.311 (0.830)	0.650 (0.342)	0.674 (0.734)
Yeast-TH	1.614 (0.843)	1.453 (0.664)	0.947 (0.918)

† $P(k) \sim k^{-\gamma}$, $C(k) \sim k^{-\alpha}$, $D(k) \sim k^{-\beta}$

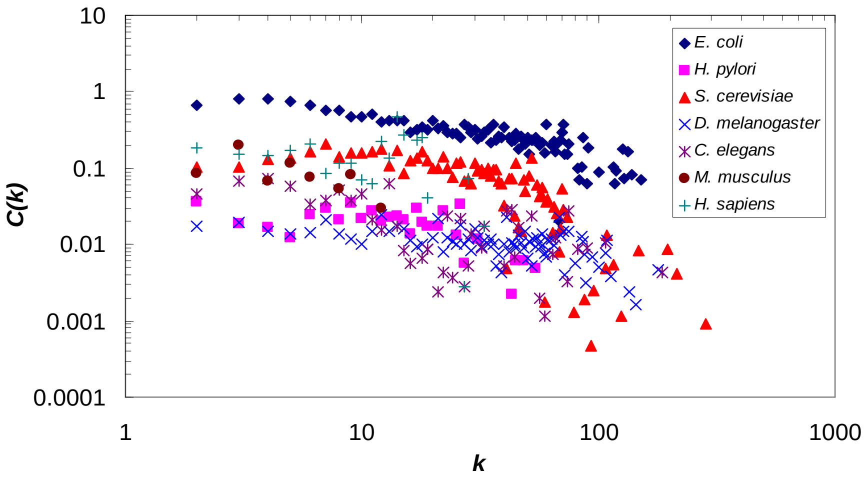

Figure 3-5 Average clustering coefficient *C(k)* as a function of degree *k* in PPI networks across different species.

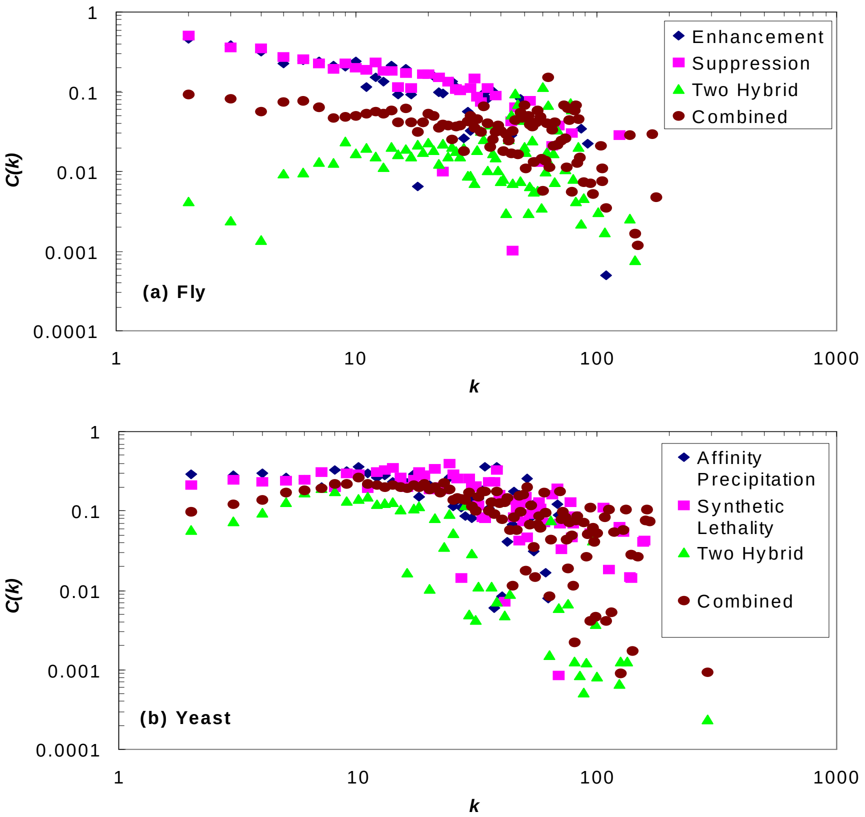

Figure 3-6 Average clustering coefficient *C(k)* as a function of degree *k* in PPI networks derived from different experimental systems.

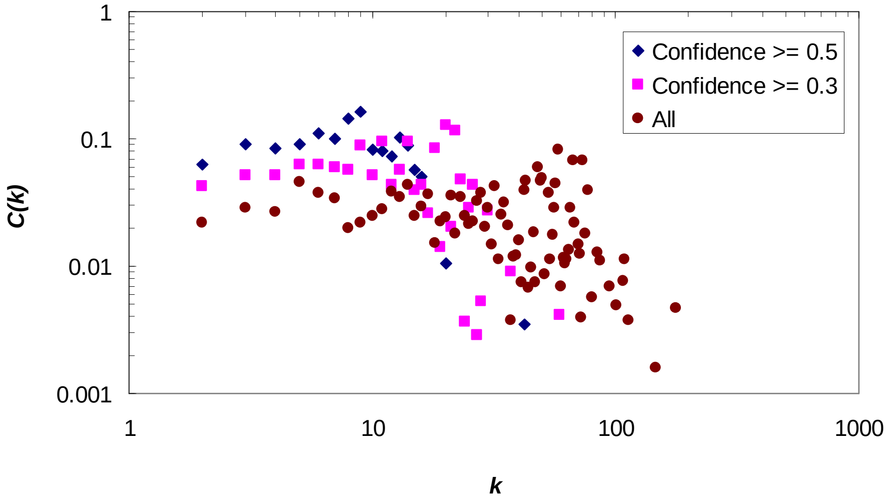

Figure 3-7 Average clustering coefficient *C(k)* as a function of degree *k* in PPI networks with different confidence levels.

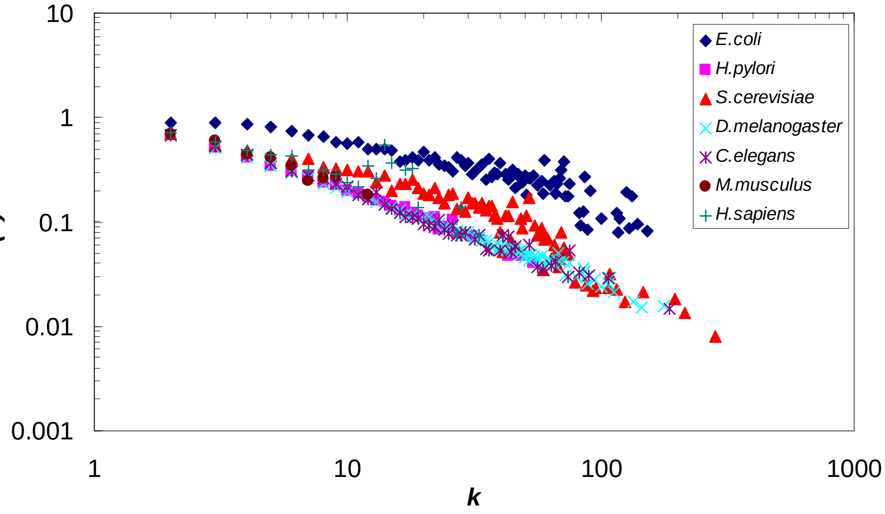

Figure 3-8 Average node density *D(k)* as a function of degree *k* in PPI networks across different species.

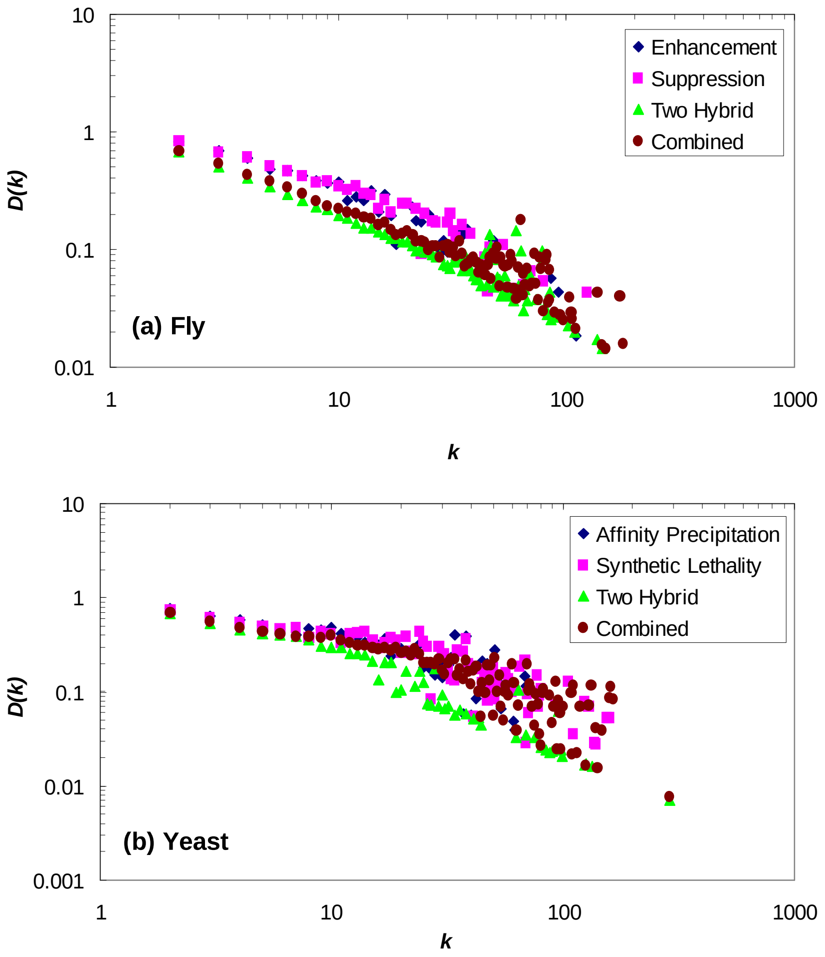

Figure 3-9 Average node density *D(k)* as a function of degree *k* in PPI networks derived from different experimental systems.

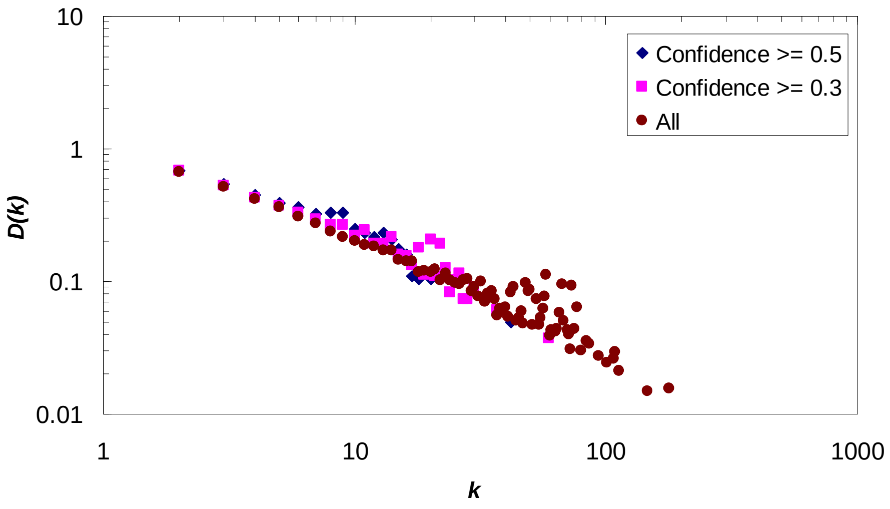

Figure 3-10 Average node density *D*(*k*) as a function of degree *k* in PPI networks with different confidence levels.

3.4 Discussion

In this chapter, we used the graph theory and statistical approaches to analyzing the topological structure of protein-protein interaction networks across different species. We also evaluated the impacts of different confidence levels and different experimental systems on the topology of PPI networks. We have shown the polarity on our data and perhaps knowledge about the PPI networks across a variety of species.

Our results confirmed that PPI networks have small diameters and small average degrees. All networks we evaluated display power law degree distribution. However, further statistical analysis indicates an inadequacy of such model in capturing certain features in the data. We strongly believe that further investigation into this issue may shed some new lights on our understanding of PPI networks.

Most of the networks we evaluated also reveal a larger clustering coefficient, indicating the non-random structure of the networks. The values of the clustering coefficient varied significantly across different species, indicating possible specie-specific behavior. However, this may also result from the incompleteness and noise of the data, since we have shown significant differences in the clustering coefficient between networks with different confidence levels. In addition, networks consisting of interactions detected from different experimental systems differed significantly in the values of the clustering coefficient. The spectrum of the average clustering coefficient over the nodes degree k fails to exhibit scale free behavior in most of the networks tested.

One interesting finding from our results is the power law distribution of average node density over the node degree *k*. This may not be total surprise because by computing node density, we introduce a new *k* into the formula. The intriguing part of this finding is coincident to a new definition introduced by [Soffer and Vazquez 2004] to eliminate degree correlation bias. They argue that the dependence of the clustering coefficient with degree *k* is partially due to this bias. The new definition they propose actually eliminates the power law behavior. On the contrary, we did not observe the power law distribution of

$C(k)$ over degree k, but the power law behavior appears when we modify the $C(k)$ to $D(k)$. We expect this information will be helpful because we have already seen its use in the application by [Bader and Hogue 2003].

CHAPTER 4. A NOVEL AND EFFICIENT APPROACH FOR IDENTIFYING A PROTEIN MODULE FROM A SEED

4.1 Introduction

One of the ultimate goals in molecular biology is to determine how genes and their encoding proteins function in the cell. It was an exciting event when gene knock-out technique first emerged, empowering biologists a revolutionary approach to discover gene function by deleting a specific gene and observing its phenotype. A nearly complete collection of single gene knockouts has been performed for *Saccharomyces cerevisiae* [Giaever et al 2002]. However, the function of a large number of genes remains unknown because single knockouts are no longer believed to be very informative due to genetic redundancy and systematic multiple gene deletions for more than two genes quickly become impossible due to the high number of possible gene combinations [Tong et al 2001; Tong et al 2004].

With the advance in high-throughput experimental technologies, more and more large-scale biological networks are being defined. System level understanding of these biological networks becomes a key challenge of the post-genomic era. The results in Chapter 3 and other accumulating evidence indicate that biological networks are composed of interacting modules of individual components [Barabasi and Oltvai 2004; Hartwell et al 1999; Ravasz et al 2002; Rives and Galitski 2003; Wu and Hu 2006a; Wu and Hu 2006b]. Therefore, a promising computational approach to discovery of functions of genes and proteins is to identify functional modules in biological networks. Because modules are sets of genes or proteins that perform biological processes together, it is possible to classify proteins with unknown function by determining what module they belong to [Palla et al 2005]. Correct identification of functional modules also has important biotechnological and pharmaceutical applications.

The work described in this chapter will be focused on the concept of modularity, which will be measured in networks based on protein-protein interaction data. Our intent is to find modules in a systematic way based on the topology of the networks that we have presented a detailed analysis in Chapter 3.

To find modular structure, we aim to choose a method that will isolate the phenomenon of interest, while not being unduly sensitive to the specifics of the approach. Our approach will be constructed to organize biological systems according to their inherent structure, to use this information to contextualize current publicly available data, and to evaluate if this information contributes to our understanding of these data.

Specifically, we begin with the question of "what is the module a given protein (or proteins) belongs to". We attempt to address this question by identifying a module of which a given protein (or proteins) is a member. Such a module will contain a set of proteins that interact with one another more than they interact with other proteins outside the module. We will evaluate whether such a module accurately captures the partitioning of cellular function and can explain the data of interest.

4.2 The *ModuleBuilder* Algorithm

4.2.1 Graph notation

Again, we intuitively model the protein-protein interaction network as an undirected graph, where nodes represent proteins and edges represent interactions between pairs of proteins.

An undirected graph, $G = (V, E)$, is comprised of two sets, nodes V and edges E. An edge e is defined as a pair of nodes (u, v), denoting the direct connection between nodes u and v. The graphs we use are undirected, unweighted, and simple – meaning no self-loops or parallel edges.

In our algorithm, we extend the quantitative definitions of a module defined by [Radicchi et al 2004]. Specifically, in a strong sense, each node in a module connects to other nodes inside the module more than it connects to those outside the module. For a weak module, the sum of edges connecting all nodes inside the module is greater than the sum of edges connecting to outside nodes.

DEFINITION 1

Let $G(V,E)$ be a graph, $G'(V',E')$ be a sub-graph of G ($G' \subset G$). Let i be a node of G'. The ***in-module degree*** for node i, $k_i^{in}(G')$, is defined as the number of edges connecting node i to other nodes inside G'. The ***out-module degree*** of node i, $k_i^{out}(G')$, is defined as the number of edges connecting node i to other nodes that are in G but not in G'. The affinity coefficient of node i to G' is defined as the ratio of the in-module degree of i to the size of G', $\frac{k_i^{in}(G')}{|G'|}$.

DEFINITION 2

Given a graph G and a sub-graph G' ($G' \subset G$), G' is a module in a strong sense if one of the following conditions is satisfied:

1) $k_i^{in}(G') > k_i^{out}(G')$ for each node i in G'.
2) The sum of all degrees within G' is greater than the sum of all degrees from G' to the rest of G.

4.2.2 The *ModuleBuilder* algorithm

The algorithm, called *ModuleBuilder* (MB), accepts the seed protein *s*, gets the neighbors of *s*, finds the core of the module to build, and expands the core to find the eventual module.

The two major components of *ModuleBuilder* are *FindCore* and *ExpandCore*. *FindCore* (line 8 to line 14) actually performs a naïve search for maximum clique in the neighborhood of the seed protein by recursively removing nodes with the lowest in-module degree until either

1) All nodes in the core set have the same in-module degree ($k_{\max} = k_{\min}$, i.e., the resulting sub-graph is a clique) or;

2) All nodes except the seed have the same in-module degree (a star-like structure).

The algorithm performs a breadth first expansion in the core expanding step. It first builds a candidate set containing the core and all nodes adjacent to each node in the core (line 16). A candidate node will then be added to the core if it meets one of the following conditions (line 21):

1) Its in-module degree is greater than its out-module degree, i.e., the quantitative definition of a module in a strong sense ($k_i^{in}(G') > k_i^{out}(G')$) or;

2) Its affinity coefficient is greater than or equals to the affinity threshold *f*.

We define the affinity coefficient of a node to a network as the fraction of its in-module degree over the size of the sub-graph. We introduce the affinity coefficient and the affinity threshold *f* to provide a degree of relaxation when expanding the core because it is too restrictive to require every expanding node to be a strong sense module member. Even though a candidate node may not have an in-module degree larger than out-module degree, it may connect to all (or even most of) other members of the network, indicating a strong tie between the candidate node and the network. When the affinity threshold *f* is set to 1, it means that in order to be eligible to add to the core set, the candidate node has to connect to all other nodes in the core set. However, *f* may be relaxed to be less than 1 if necessary or so desired.

In addition, a distance parameter, *d*, is used to restrict how far away a candidate node to the seed can be considered eligible for expansion. Quite often, a given seed may not always situate in the center of the resulting module. The distance parameter serves as the shortest path threshold to ensure that all members of the obtained sub-network will be within specified proximity to the seed. A large enough value of *d*, such as one that is larger than the longest path from the seed to all other nodes in the network, will virtually lift this distance restriction.

4.2.3 Complexity of the *ModuleBuilder* algorithm

The *FindCore* is a heuristic search for a maximum complete sub-graph in the neighborhood *N* of seed *s*. Let *k* be the size of *N*, then the worst-case running time of *FindCore* is $O(k^2)$. The *ExpandCore* part costs, in the worst case, approximately $|V|$ + $|E|$ + overhead. $|V|$ accounts for the expanding of the core; at most, all nodes in *V*, minus what are already in the core, would be included. $|E|$ accounts for calculating the in- and out-module degrees for the candidate nodes that are not in the core but are in the neighborhood of the core. The overhead is caused by recalculating the in- and out-module degrees of neighboring nodes every time the *FindCore* is recursively called. The number of these nodes is dependent on the size of the module we are building and the connectivity of the module to the rest of the network, but not the overall size of the network. For biological networks, the graphs we deal with are mostly sparse and small world; therefore, the running time of our algorithm is close to linear.

4.3 Experiments and results

To test our algorithm, we downloaded a data set of interactions for *Saccharomyces cerevisae* from the General Repository for Interaction Datasets (GRID) [Breitkreutz et al 2003]. The GRID database contains all published large-scale interaction datasets as well as available curated interactions such as those deposited in BIND [Bader et al 2003] and MIPS [Mewes et al 2002]. The yeast dataset we downloaded has 4,907 proteins and 17,598 interactions.

We applied our algorithm against the network built from the downloaded data set. The average running time for finding a community of about 50 members is about 20 ms.

Algorithm 1 *ModuleBuilder*(*G, s, f, d*)

```
$G(V, E)$ is the input graph with node set $V$ and edge set $E$.
$s$ is the seed node; $f$ is the affinity threshold; $d$ is the distance threshold.
$N \leftarrow$ {Adjacency list of $s$ } $\cup \{s\}$
$C \leftarrow$ FindCore($N$)
$C' \leftarrow$ ExpandCore($C, f, d$)
return $C'$

FindCore($N$)
    for each $v \in N$
        calculate $k_v^{in}(N)$
    end for
    $K_{min} \leftarrow \min\{ k_v^{in}(N), v \in N\}$
    $K_{max} \leftarrow \max\{ k_v^{in}(N), v \in N\}$
    if $K_{min} = K_{max}$ or $(k_i^{in}(N) = k_j^{in}(N), \forall i, j \in N, i, j \neq s, i \neq j)$ then
    return $N$
    else return FindCore($N - \{v\}, k_v^{in}(N) = K_{min}$)

ExpandCore($C, f, d$)
    $D \leftarrow \bigcup_{(v,w)\in E, v\in C, w\notin C} \{v, w\}$
    $C' \leftarrow C$
    for each $t \in D$, $t \notin C$, and distance($t$, $s$) <= $d$
        calculate $k_t^{in}(D)$
        calculate $k_t^{out}(D)$
        if $k_t^{in}(D) > k_t^{out}(D)$ or $k_t^{in}(D)/|D| > f$ then
            $C' \leftarrow C' \cup \{t\}$
    end for
    if $C' = C$ then return $C$
    else return ExpandCore($C', f, d$)
```

Figure 4-1 ***ModuleBuilder*** **algorithm**

Because there is no alternative approach to our method, we decide to compare the performance of our algorithm to the work on predicting protein complex membership by [Asthana et al 2004]. Asthana and colleagues reported results of queries with four complexes using probabilistic network reliability (we will refer their work as PNR method in the following discussion). Four modules are identified by *ModuleBuilder* using one protein as seed from each of the query complexes used by the PNR method. The seed protein is selected randomly from the "core" protein set. The figures for visualizing the identified modules are created using *Pajek* software [Batagelj and Mrvar 1998]. The module figures are extracted from the network we build using the above mentioned data set with out-of-module connections omitted. The proteins in each module are annotated with a brief description obtained from the MIPS complex catalogue database. As a comparison, we use *Complexpander*, an implementation of the PNR method [Asthana et al 2004] and freely available at http://llama.med.harvard.edu/Software.html, to predict co-complex using the core protein set that contains the same seed protein used by *ModuleBuilder*. For all our queries when using *Complexpander*, we select the option to use the MIPS complex catalogue database. We record the ranking of the members in our identified modules that also appear in the co-complex candidate list predicted by *Comlexpander*.

The first module, shown in Figure 4-2, is identified using TAF6 as seed. TAF6 is a component of the SAGA complex which is a multifunctional co-activator that regulates transcription by RNA polymerase II [Wu et al 2004]. The SAGA complex is listed in MIPS complex catalogue as a known cellular complex consisting of 16 proteins. As shown in Table 4-1, the module identified by our algorithm contains 39 members, including 14 of the 16 SAGA complex proteins listed in MIPS (indicated by an asterisk in the Alias column). The module also contains 14 of 21 proteins listed in MIPS as Kornberg's mediator (SRB) complex. The rest of the proteins in the community are either TATA-binding proteins or transcription factor IID (TFIID) subunits or SRB related. TFIID is a complex involved in initiation of RNA polymerase II transcription. SAGA and TFIID are structurally and functionally correlated, make overlapping contributions to the expression of RNA polymerase II transcribed genes [Wu et al 2004]. SRB complex is a mediator that conveys regulatory signals from DNA-binding transcription factors to RNA polymerase II [Guglielmi et al 2004]. In addition, 27 of the top 50 potential co-complex proteins (9 of the top 10), not including the seed proteins, predicted by *Complexpander* are in the identified module.

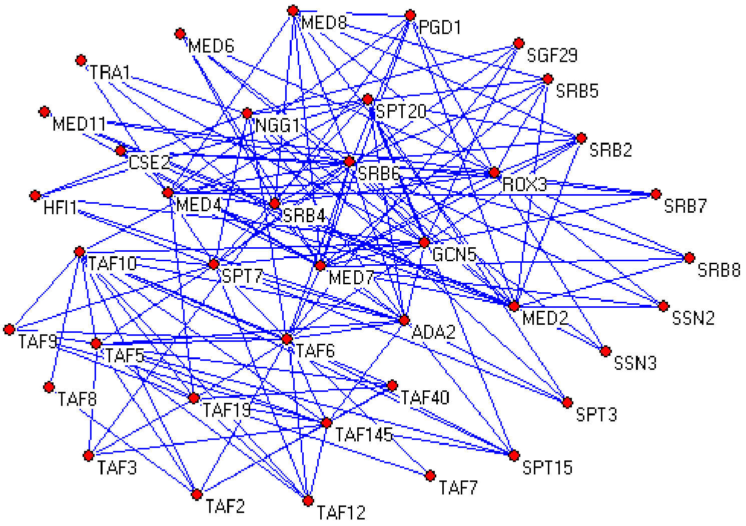

Figure 4-2 The SAGA/SRB module.

Table 4-1 Members of the SAGA/SRB module.

Protein[a]	Alias	Description	Rank
YDR448w	ADA2[b]	general transcriptional adaptor or co-activator	1
YNR010w	CSE2[c]	subunit of RNA polymerase II mediator complex	
YGR252w	GCN5[b]	histone acetyltransferase	2
YPL254w	HFI1[b]	transcriptional coactivator	3
YMR112c	MED11[c]	mediator complex subunit	
YDL005c	MED2[c]	transcriptional regulation mediator	20
YOR174w	MED4[c]	transcription regulation mediator	23
YHR058c	MED6[c]	RNA polymerase II transcriptional regulation mediator	
YOL135c	MED7[c]	member of RNA Polymerase II transcriptional regulation mediator complex	21
YBR193c	MED8[c]	transcriptional regulation mediator	24
YDR176w	NGG1[b]	general transcriptional adaptor or co-activator	10
YGL025c	PGD1[c]	mediator complex subunit	37
YBL093c	ROX3[c]	transcription factor	
YCL010c	SGF29[b]	SAGA associated factor	43
YER148w	SPT15	the TATA-binding protein TBP	15
YOL148c	SPT20[b]	member of the TBP class of SPT proteins that alter transcription site selection	4
YDR392w	SPT3[b]	general transcriptional adaptor or co-activator	13

YBR081c	SPT7[b]	involved in alteration of transcription start site selection	5
YHR041c	SRB2[c]	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	
YER022w	SRB4[c]	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	27
YGR104c	SRB5[c]	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	
YBR253w	SRB6[c]	DNA-directed RNA polymerase II suppressor protein	19
YDR308c	SRB7[c]	DNA-directed RNA polymerase II holoenzyme and kornberg^s mediator (SRB) subcomplex subunit	46
YCR081w	SRB8	DNA-directed RNA polymerase II holoenzyme and Srb10 CDK subcomplex subunit	
YDR443c	SSN2	DNA-directed RNA polymerase II holoenzyme and Srb10 CDK subcomplex subunit	
YPL042c	SSN3	cyclin-dependent CTD kinase	
YGR274c	TAF1	TFIID subunit (TBP-associated factor), 145 kD	14
YDR167w	TAF10[b]	TFIID and SAGA subunit	7
YML015c	TAF11	TFIID subunit (TBP-associated factor), 40KD	18
YDR145w	TAF12[b]	TFIID and SAGA subunit	8
YML098w	TAF13	TFIID subunit (TBP-associated factor), 19 kD	17
YCR042c	TAF2	component of TFIID complex	22
YPL011c	TAF3	component of the TBP-associated protein complex	50
YBR198c	TAF5[b]	TFIID and SAGA subunit	9
YGL112c	TAF6[b]	TFIID and SAGA subunit	
YMR227c	TAF7	TFIID subunit (TBP-associated factor), 67 kD	
YML114c	TAF8	TBP Associated Factor 65 KDa	
YMR236w	TAF9[b]	TFIID and SAGA subunit	11
YHR099w	TRA1[b]	component of the Ada-Spt transcriptional regulatory complex	12

[a] The open reading frame (ORF) name is used.
[b] Proteins belong to SAGA complex listed in MIPS.
[c] Proteins belong to SRB complex listed in MIPS.

The second module is discovered using NOT3 as seed (Figure 4-3 and Table 4-2). NOT3 is a known component protein of the CCR4/NOT complex which is a global regulator of gene expression and involved in such functions as transcription regulation and DNA damage responses. MIPS complex catalogue lists 5 proteins for NOT complex and 13 proteins (including the 5 NOT complex proteins) for CCR4 complex. The NOT module identified is composed of 40 members. All 5 NOT complex proteins listed in MIPS and 11 of the 13 CCR4 complex proteins are members of the module. POL1, POL2, PRI1, and PRI2 are members of the DNA polymerase alpha (I) – primase complex, as listed in MIPS. RVB1, PIL1, UBR1, and STI1 have been grouped together with CCR4, CDC39, CDC36, and POP2 by systematic analysis [Ho et al 2002]. The module also contains 20 out of 26 proteins of a complex that is probably involved in transcription and DNA/chromatin structure maintenance [Gavin et al 2002].

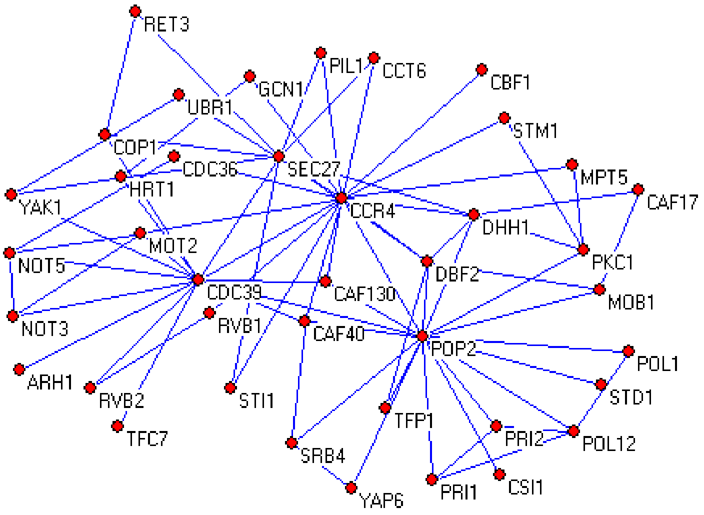

Figure 4-3 The CCR4/NOT module.

Table 4-2 Members of the CCR4/NOT module.

Protein[a]	Alias	Description	Rank
YDR376w	ARH1	mitochondrial protein putative ferredoxin-NADP+ reductase	38
YGR134w	CAF130 [c]	CCR4 Associated Factor 130 kDa	8
YJR122w	CAF17 [b]	CCR4 associated factor	
YNL288w	CAF40 [c]	CCR4 Associated Factor 40 kDa	9
YJR060w	CBF1	centromere binding factor 1	
YAL021c	CCR4 [bc]	transcriptional regulator	3
YDR188w	CCT6 [c]	component of chaperonin-containing T-complex (zeta subunit)	30
YDL165w	CDC36 [bc]	transcription factor	40
YCR093w	CDC39 [bc]	nuclear protein	1
YDL145c	COP1 [c]	coatomer complex alpha chain of secretory pathway vesicles	11
YMR025w	CSI1	Subunit of the Cop9 signalosome, involved in adaptation to pheromone signaling	46
YGR092w	DBF2 [b]	ser/thr protein kinase related to Dbf20p	6
YDL160c	DHH1 [b]	DExD/H-box helicase, stimulates mRNA decapping,	17
YGL195w	GCN1 [c]	translational activator	26
YOL133w	HRT1	Skp1-Cullin-F-box ubiquitin protein ligase (SCF) subunit	
YIL106w	MOB1 [b]	required for completion of mitosis and maintenance of ploidy	10
YER068w	MOT2[bc]	transcriptional repressor	2
YGL178w	MPT5	multicopy suppressor of POP2	
YIL038c	NOT3[bc]	general negative regulator of transcription, subunit 3	
YPR072w	NOT5[bc]	component of the NOT protein complex	5
YGR086c	PIL1	Long chain base-responsive inhibitor of protein kinases Phk1p and Phk2p, acts along with Lsp1p to down-regulate heat stress resistance	
YBL105c	PKC1	ser/thr protein kinase	
YNL102w	POL1[c]	DNA-directed DNA polymerase alpha, 180 KD subunit	32
YBL035c	POL12 [c]	DNA-directed DNA polymerase alpha, 70 KD subunit	28
YNR052c	POP2 [bc]	required for glucose derepression	4
YIR008c	PRI1 [c]	DNA-directed DNA polymerase alpha 48kDa subunit (DNA primase)	34
YKL045w	PRI2 [c]	DNA-directed DNA polymerase alpha , 58 KD subunit (DNA primase)	31
YPL010w	RET3	coatomer complex zeta chain	39
YDR190c	RVB1	RUVB-like protein	29
YPL235w	RVB2 [c]	RUVB-like protein	21
YGL137w	SEC27 [c]	coatomer complex beta^ chain (beta^-cop) of secretory pathway vesicles	7
YER022w	SRB4	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	44
YOR047c	STD1	dosage-dependent modulator of glucose repression	

YOR027w	STI1	stress-induced protein	
YLR150w	STM1	specific affinity for guanine-rich quadruplex nucleic acids	
YOR110w	TFC7[c]	TFIIIC (transcription initiation factor) subunit, 55 kDa	25
YDL185w	TFP1[c]	encodes 3 region protein which is self-spliced into TFP1p and PI-SceI	27
YGR184c	UBR1	ubiquitin-protein ligase	
YJL141c	YAK1	ser/thr protein kinase	
YDR259c	YAP6	transcription factor, of a fungal-specific family of bzip proteins	

[a] The open reading frame (ORF) name is used.
[b] Proteins belong to CCR4/NOT complex listed in MIPS.
[c] Proteins considered part of a complex involved in transcription and DNA/chromatin structure maintenance.

The third module is identified by using RFC2 as the seed (Figure 4-4 and Table 4-3). RFC2 is a component of the RFC (replication factor C) complex, the “clamp loader”, which plays an essential role in DNA replication and DNA repair. The module identified by our algorithm has 17 members. All five proteins of RFC complex listed in MIPS complex catalogue database are members of this module, as shown in Table 4-3. All but one member in this module are in the functional category of DNA recombination and DNA repair or cell cycle checkpoints according to MIPS. This module also includes the top 8 ranked proteins predicted by *Complexpander*.

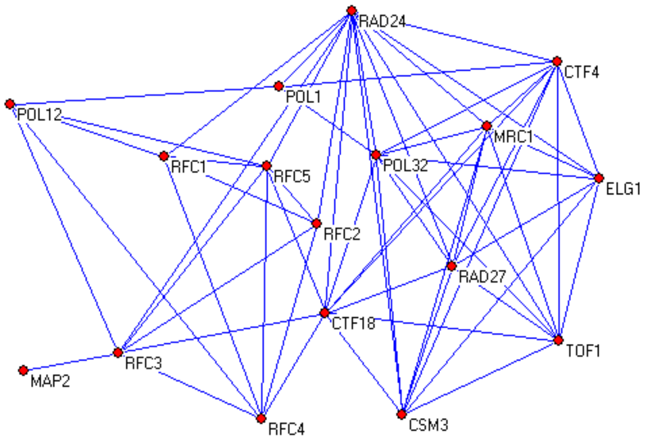

Figure 4-4 The RFC module.

Table 4-3 Members of the RFC module.

Protein[a]	**Alias**	**Description**	**Rank**
YMR048w	CSM3[c]	Protein required for accurate chromosome segregation during meiosis	
YMR078c	CTF18[c]	required for accurate chromosome transmission in mitosis and maintenance of normal telomere length	6
YPR135w	CTF4[c]	DNA-directed DNA polymerase alpha-binding protein	
YOR144c	ELG1[c]	Protein required for S phase progression and telomere homeostasis, forms an alternative replication factor C complex important for DNA replication and genome integrity	7
YBL091c	MAP2	methionine aminopeptidase, isoform 2	
YCL061c	MRC1[c]	Mediator of the Replication Checkpoint	
YNL102w	POL1[c]	DNA-directed DNA polymerase alpha, 180 KD subunit	19
YBL035c	POL12[c]	DNA-directed DNA polymerase alpha, 70 KD subunit	5
YJR043c	POL32[c]	polymerase-associated gene, third (55 kDa) subunit of DNA polymerase delta	
YER173w	RAD24[c]	cell cycle checkpoint protein	1
YKL113c	RAD27[c]	ssDNA endonuclease and 5^-3^exonuclease	
YOR217w	RFC1[bc]	DNA replication factor C, 95 KD subunit	8
YJR068w	RFC2[bc]	DNA replication factor C, 41 KD subunit	
YNL290w	RFC3[bc]	DNA replication factor C, 40 kDa subunit	2
YOL094c	RFC4[bc]	DNA replication factor C, 37 kDa subunit	4

YBR087w	RFC5[bc]	DNA replication factor C, 40 KD subunit	3
YNL273w	TOF1[c]	topoisomerase I interacting factor 1	

[a] The open reading frame (ORF) name is used.
[b] Proteins belong to RFC complex listed in MIPS.
[c] Proteins listed in the functional category of DNA recombination and DNA repair or cell cycle checkpoints in MIPS.

We use ARP3 as seed to identify the last module (Figure 4-5). ARP2/ARP3 complex acts as multi-functional organizer of actin filaments. The assembly and maintenance of many actin-based cellular structures likely depend on functioning ARP2/ARP3 complex [Machesky and Gould 1999]. The identified module contains all 7 proteins of the ARP2/ARP3 complex listed in MIPS (Table 4-4). Not including the seed (ARP3), these proteins represent the top 6 ranked proteins predicted by *Complexpander*. As indicated in Table 4-4, there are 14 members belonging to the same functional category of budding, cell polarity, and filament formation, according to MIPS.

By using the MIPS complex data as "gold standard", we also calculate the recall and precision of the four modules (or co-complexes as called by *Complexpander*) obtained by *ModuleBuilder* and *Complexpander*. The results are presented in Table 4-5 and Figure 4-6.

Comparing the recall values, *ModuleBuilder* performs better than *Complexpander* in two cases (81.1% vs 73% for SAGA/SRB and 84.6% vs 38.5% for

CCR4/NOT) and performs as good as *Complexpander* in the other two cases (both are 100%).

Comparing the precision scores, *ModuleBuilder* performs much better than *Complexpander* in all four cases.

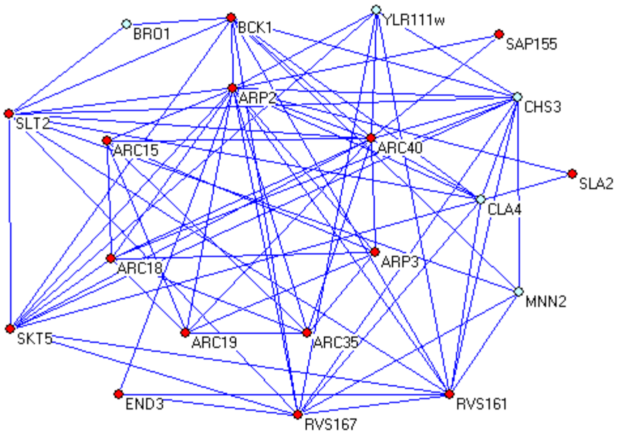

Figure 4-5 The ARP2/ARP3 module.

Table 4-4 Members of the ARP2/ARP3 module.

Protein[a]	Alias	Description	Rank
YLR111w	YLR111w	hypothetical protein	
YIL062c	ARC15[bc]	subunit of the Arp2/3 complex	1
YLR370c	ARC18[b]	subunit of the Arp2/3 complex	4
YKL013c	ARC19[bc]	subunit of the Arp2/3 complex	3
YNR035c	ARC35[b]	subunit of the Arp2/3 complex	5
YBR234c	ARC40[bc]	Arp2/3 protein complex subunit, 40 kilodalton	6
YDL029w	ARP2[bc]	actin-like protein	2
YJR065c	ARP3[b]	actin related protein	
YJL095w	BCK1[c]	ser/thr protein kinase of the MEKK family	
YPL084w	BRO1	required for normal response to nutrient limitation	
YBR023c	CHS3[c]	chitin synthase III	

YNL298w	CLA4[c]	ser/thr protein kinase
YNL084c	END3[c]	required for endocytosis and cytoskeletal organization
YBR015c	MNN2	type II membrane protein
YCR009c	RVS161[c]	protein involved in cell polarity development
YDR388w	RVS167[c]	reduced viability upon starvation protein
YFR040w	SAP155[c]	Sit4p-associated protein
YBL061c	SKT5[c]	protoplast regeneration and killer toxin resistance protein
YNL243w	SLA2[c]	cytoskeleton assembly control protein
YHR030c	SLT2[c]	ser/thr protein kinase of MAP kinase family

[a] The open reading frame (ORF) name is used.

[b] Proteins belong to ARP2/ARP3 complex listed in MIPS.

[c] Proteins listed in the functional category of budding, cell polarity, and filament formation in MIPS.

Table 4-5 Comparison of the results from ModuleBuilder (MB) and Complexpander (CE). In MIPS complex category, there are 37 proteins in SAGA/SRB complexes, 13 proteins in CCR4/NOT complex, 5 proteins in RFC complex, and 7 proteins in ARP2/ARP3 complex.

Complex /Module	***SAGA/SRB***		***CCR4/NOT***		***RFC***		***ARP2/ARP3***	
	MB	***CE***	***MB***	***CE***	***MB***	***CE***	***MB***	***CE***
# of Proteins	39	60	40	50	17	72	20	64
# in MIPS	30	27	11	5	5	5	7	7
Recall	81.1%	73.0%	84.6%	38.5%	100.0%	100.0%	100.0%	100.0%
Precision	76.9%	45.0%	27.5%	10.0%	29.4%	6.9%	35.0%	10.9%

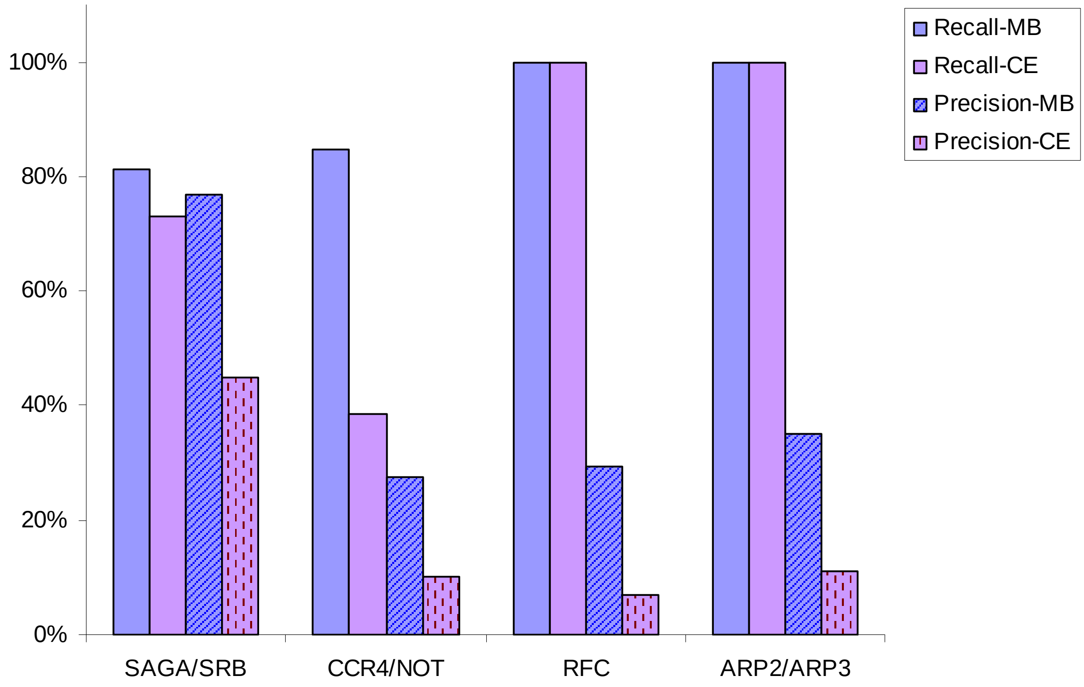

Figure 4-6 Recall and precision of *ModuleBuilder* (MB) and *Complexpander* (CE).

4.4 Discussion

We present in this chapter an efficient approach to building a module from a given seed protein. It uses topological property of modular structure of a network and takes advantage of local optimization in searching for the module comprising of the seed protein. Due to the complexity and modularity of biological networks, it is more desirable and computationally feasible to model and simulate a network of smaller size. Our approach builds a module of manageable size and scales well to large networks. Its usefulness is demonstrated by the experimental results that all the four modules identified reveal strong structural and functional relationships among member proteins. It provides a fast and accurate way to find a module comprising a protein or proteins with known functions or of interest. For those module members that are not known to be part of a protein complex or a functional category, their relationship to other members in the same module may deserve further investigation which in turn may provide new insights.

Although we do not explicitly use our approach to the prediction of co-complexed proteins, the results of comparing with the PNR method developed by [Asthana et al 2004] have shown that the modules identified by our approach do include the top ranked candidates of co-complexed proteins. Both the recall and precision scores depict better performance of *ModuleBuilder* over *Complexpander* in retrieving the complex proteins (Table 4-5 and Figure 4-6).

Compared to the methods in predicting co-complexed proteins, our approach can discover a module rather than a single complex. In the context of this discussion, the notion of a module can be a complex, but it can also be a functional group consisting of several complexes, such as the SAGA/SRB module (Figure 4-2). This does provide benefits of delineating the structure-function relationships beyond a single complex. In this spirit, one part of our future work is to further explore the relaxation threshold (*f*) aiming to identify either a more tightly connected module under a more strict expanding condition or a more loosely connected module under a relaxed condition so that we could study interactions of different strengths within a module.

Our approach does not consider the quality of data in our downloaded data set. By using the strong sense definition of a module [Radicchi et al 2004], we could reduce the noises to some degree. However, to improve the quality of an identified module, we have to take into account the quality of data and that is another part of our future work. One possible way is to use the probabilities assigned to individual protein pairs as used by [Jansen et al 2002; Radicchi et al 2004; Bader 2003; Bader et al 2004].

CHAPTER 5. GLOBAL ANALYSIS OF PROTEIN TRANSLATION REGULATORY NETWORKS IN YEAST

5.1 Introduction

The central dogma of molecular biology describes that the genetic information is transferred from DNA to mRNA through transcription and from mRNA to protein via translation. Transcriptional regulation and translational regulation are two critical control points in any biological systems.

As we have discussed earlier in Chapter 1 and Chapter 2, the complexity and the large sizes of the transcriptional and translational machineries make computational approaches attractive and necessary in facilitating our understanding the design principles and functional properties of the cell. Transcriptional regulation, used by cells to control gene expression, has been a focus in a variety of computational methods to infer the structure of genetic regulatory networks or to study their high level properties. However, research on translational regulatory networks has caught little attention in the bioinformatics and computational biology community, either being underestimated or neglected.

In every living organism, translation is a vital cellular process in which the information contained in the mRNA sequence is translated into the corresponding protein by the complex translation machinery. Therefore, study of protein translational regulation plays an important role in our understanding of the molecular mechanisms of living cells. Traditionally this study has been carried out by using reductionistic approaches, i.e. through experiments that are individually designed to identify specifically targeted proteins and/or interactions. However, with the completion of genome sequence projects and development of a wide range of functional genomics tools, it has become possible to apply system-level approaches to understand the function of biological systems and in particular protein translational control. Whereas the final objective of systems biology is to enable quantitative prediction of the dynamics of cellular processes, an important first step is to reconstruct the network structure of these processes

Toward the goal of understanding how translation machinery functions from a system's perspective, it is imperative that we have a better understanding of the global properties of protein translation networks, especially integrated with functional perspectives. In this chapter, we take this first step in pursuing such a goal. We use a graph theoretic approach to reconstruct and investigate protein translation regulatory networks (PTRN) in yeast by integrating the protein-protein interaction data, the functional annotations documented in MIPS and GO databases, and some of the recent research results on cellular localization and protein phosphorylation in regard to PTRN.

5.2 Methods

5.2.1 Graph notation

We will use the same graph notations as described in previous chapters.

5.2.2 Data sets

The yeast protein-protein interactions data were downloaded from the General Repository for Interaction Datasets (GRID) [Breitkreutz et al 2003]. We select GRID because it contains arguably the most comprehensive data. The GRID database includes all published large-scale interaction datasets as well as available curated interactions such as those deposited in BIND [Bader et al 2003] and MIPS [Mewes et al 2002]. The yeast dataset we downloaded has 4,948 distinct proteins and 18,817 unique interactions. From this network, we derive the protein translation networks which contain all proteins with MIPS functional categories related to protein translation as described in Section 3.

We also compiled yeast functional annotations and essentiality of proteins from MIPS and GO. Protein phosphorylation data were obtained from [Ptacek et al 2005] and protein localization data from [Huh et al 2003].

5.2.3 Analysis of network topology

We measure the following basic properties of a PTRN: 1) the number of proteins, measured by the number of nodes; 2) the number of interactions, measured by the number of edges; 3) the size of the largest (or giant) component, measured by the size of the largest connected sub-graph.

We also measure the following topological metrics:

- The average degree (<*k*>), defined as

  $$\langle k \rangle = \frac{2|E|}{|V|},$$

  where $|E|$ is the total number of edges and $|V|$ is the total number of nodes.
- The degree distribution, $P(k)$, which measures the frequency of a node having degree of *k*.
- The diameter of a network $\langle l \rangle$, as defined in Chapter 3.
- The clustering coefficient C_i of a node *i*, as defined in Chapter 3.
- The average clustering coefficient of a network $\langle C \rangle$ and its equivalent random network $\langle C_{rand} \rangle$, as defined in Chapter 3.

All statistical analyses are performed by using SPSS software package.

5.3 Results

5.3.1 Global properties of PTRNs within the full yeast interactome

We extract two sets of proteins that are involved in protein biosynthesis from MIPS functional category database. Table 5-1 shows the functional categories used.

The first set, we name it N1247, contains 136 unique proteins and belongs to the following categories:

- 12.04 (translation)
- 12.04.01 (translation initiation)
- 12.04.02 (translation elongation)
- 12.04.03 (translation termination)
- 12.07 (translational control)

The second set, referred as N12, contains 479 unique proteins in all categories listed in Table 5-1. Therefore, the first set of proteins is actually a subset of the second.

We first study the PTRN by using these two sets of proteins in the context of the full yeast interaction network, constructed from yeast protein-protein interaction data from GRID. The basic properties of the full yeast interaction network and the PTRN are shown in Table 5-2.

One interesting observation of the PTRN is the existence of proteins that do not have any interacting partners in the full network. We call them the loner proteins. This reflects the low coverage of the current interaction database rather than the actual lack of interactions.

Two fundamental network metrics, node degree and clustering coefficient, are employed to evaluate the global network characteristics. The node degree describes the number of interacting partners for each node in the network, whereas the clustering coefficient quantifies how well connected are the neighbors of a node in the network. These metrics provide useful insights into the architecture of the underlying network.

Table 5-1 MIPS functional categories related to protein translation.

Category	Description	# of Proteins
12	protein synthesis	22
12.01	ribosome biogenesis	63
12.01.01	ribosomal proteins	245
12.04	translation	18
12.04.01	translation initiation	40
12.04.02	translation elongation	21
12.04.03	translation termination	9
12.07	translational control	55
12.10	aminoacyl-tRNA-synthetases	39

Table 5-2 Properties of the full yeast interaction network and the protein translation regulatory networks.

Network	N1247	N12	Full
# of interacting proteins	785	1471	4948
# of loner proteins	13	76	0
# of unique interactions	1100	2715	18817
Average degree <*k*>	10.18	8.08	7.61
Average clustering coefficient $<C_k>$	0.1096	0.1241	0.1118
C_{rand}	0.0555	0.0304	0.0243

For the first PTRN (N1247), the core set of proteins extracted from MIPS, contains 136 unique proteins. There are 13 of them that do not have interacting partners in the full network. The remaining 123 proteins form a network containing 1100 unique interactions that involve additional 662 proteins. As shown in Table 5-2, the average degree of N1247 is significantly higher than the other larger PTRN and the full network.

For the second PTRN (N12), the number of extracted unique proteins from MIPS is 479. Again, 76 of them are not shown interacting partners in the data we used. There are additional 1068 proteins interacting with the remaining 403 proteins through 2715 distinct interactions.

All three networks show a larger average clustering coefficient than the corresponding $< C_{rand} >$, indicating a non-random structure of the underlying networks.

The degree distribution is a function describing the probability of a node having a specified degree. It is used regularly to classify networks, such as random networks (Poisson distribution) or scale-free networks (power law distribution). As shown in Figure 5-1, the degree distribution for the full interaction network displays an approximate power law. However, the two translation networks show only weak power law degree distributions.

Regression analysis is performed and the power values with *R* squares are shown in Figure 5-1. When we use an alternative approach to evaluate the degree distributions [Newman 2003a], in which the probability $P(k)$ is for all k values greater than or equal to k, we find that the degree distributions for the two translation networks fit better to an exponential regression instead of a power law. It has been proved [Newman 2003a] that the cumulative distribution follows a power law if the original distribution does so, but with a different exponent that is one less than the original exponent. Therefore, considering the nature of the data we have (incomplete and may contain noise), the result in Figure 5-2 indicates that the scale-free or non-scale-free topology of the PTRN remains an issue to be solved.

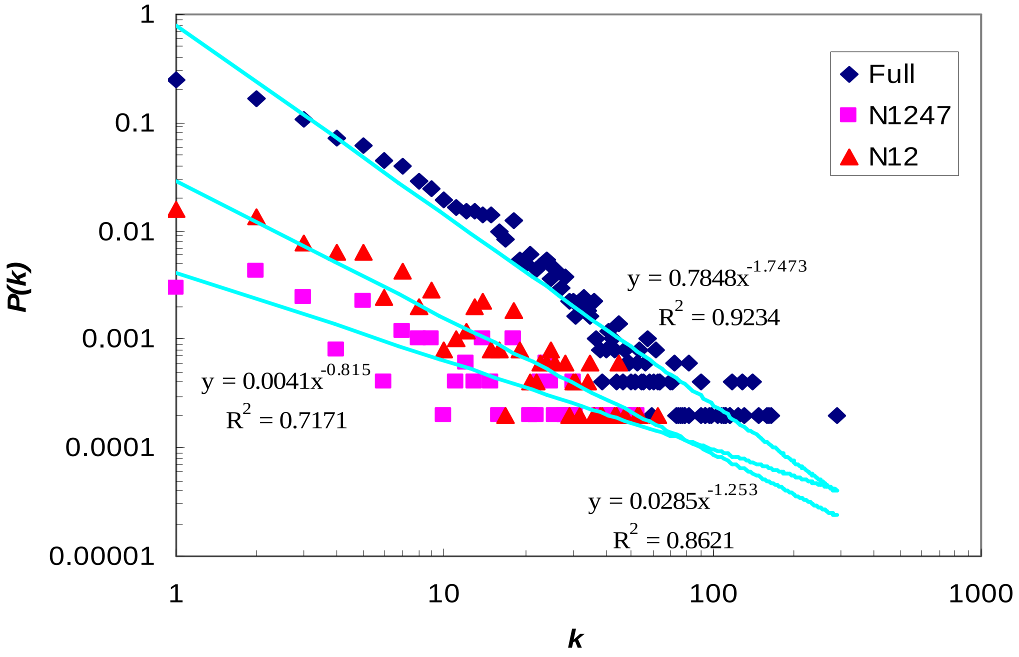

Figure 5-1 Degree distributions. The cyan lines show the power law regression.

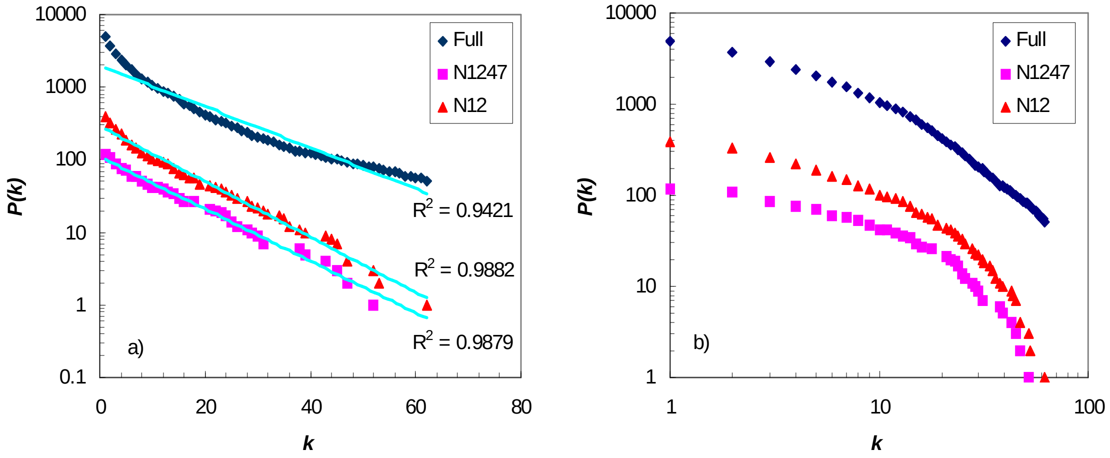

Figure 5-2 Cumulative degree distributions. a) Semi-logarithmic plot with exponential regression. b) Log-log plot.

5.3.2 Reconstruct PTRN

To further investigate and gain more insights into the organization of PTRN, we construct the following networks:

1. The first network (named N1247S hereafter, Figure 5-3) contains only the first set of 136 proteins described in Section 3.1. All edges are extracted from the full interaction network. This network only contains interactions between proteins that are both in N1247.
2. The second network (N1247SA, Figure 5-4) is extended from N1247S. It is in fact the isolated N1247 network mentioned in last section, containing other proteins that interact with those in N1247. Each interaction in the network has at least one of the interacting partners in N1247.
3. The third network (N12S, Figure 5-5a) contains only proteins in N12. In addition to all of the proteins in N1247, network N12S also contains ribosomal proteins, proteins involved in ribosome synthesis, and proteins with aminoacyl-tRNA transferase activities.
4. The fourth network (N12SA, Figure 5-5b) is extended from N12S containing proteins either in N12 or interacting with proteins in N12.

The basic properties of networks are listed in Table 5-3. It is noted that all these constructed networks has significantly lowered average degree due to the existence of the "loner proteins" that do not interact with others. Again, no interactions here do not necessarily mean no interactions in reality. Rather, the lack of interactions is either due to the lack of data in our data sets (false negatives) or the lack of qualified interactions (such as restraints posted on N1247S and N12S).

As demonstrated in Figure 3, the interactions between proteins within this group are surprisingly low. Most of the interactions exist in clusters corresponding to the three stages of protein translation. The highest connected cluster belongs to the group of proteins (in green) that are involved in translation initiation. This no double reflects the belief that translation initiation is one of the most important points in translational control and the active research leading to the high coverage in this area. Figure 3 also indicates the same logic for proteins involved in translation elongation (in yellow). Probably the most surprising finding is that there are very few direct interactions of translational control proteins (in blue) either between themselves or with proteins from other clusters (especially those in translation initiation as one might expect). One possible reason behind this finding may be that the interactions between the control proteins and others are transient such that current technologies such as high-throughput ones are unable to capture them (false negatives). Another reason might be that the control may exert through indirect (intermediates) interactions.

Nonetheless, Figure 3 demonstrates the usefulness of network and cluster analyses in helping to delineate the translation process.

By extending the network N1247S to include all interacting partners, N1247SA becomes much more connected, as indicated by the increased size of the giant component (from 41% of total proteins to 89%), and the decreased number of loner proteins. There are 662 proteins outside these defined categories that interact with proteins in N1247. A subsequent and natural question to ask is: what are those proteins? The answer is quite intriguing. Table 3 lists the top 10 functional categories to which these 662 proteins belong. It should be noted that one protein may be listed in multiple categories. We examine these functional categories at different levels (where available) in the function hierarchy used by MIPS. At the top level, more than 46% of these proteins are in the functional category of “CELL CYCLE AND DNA PROCESSING”, 42% in “TRANSCRIPTION”, and more than 37% in “METABOLISM”. Subsequent levels further detail the distributions of these proteins in child categories of the top level parents. This result clearly demonstrates the close relationship between translation and other cellular processes especially transcription and metabolism.

Table 5-3 Functional categories of proteins interacting with translation networks.

Functional Category	Description	% of Proteins
Top level		
10	CELL CYCLE AND DNA PROCESSING	46.2%
11	TRANSCRIPTION	42.0%
1	METABOLISM	37.5%
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	31.0%
20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	26.7%
14	PROTEIN FATE (folding, modification, destination)	26.1%
42	BIOGENESIS OF CELLULAR COMPONENTS	20.7%
43	CELL TYPE DIFFERENTIATION	11.6%
32	CELL RESCUE, DEFENSE AND VIRULENCE	11.3%
34	INTERACTION WITH THE CELLULAR ENVIRONMENT	10.0%
Second level		
10.03	cell cycle	24.2%
10.01	DNA processing	21.9%
11.04	RNA processing	20.1%
11.02	RNA synthesis	19.6%
20.09	transport routes	16.0%
16.03	nucleic acid binding	13.3%
43.01	fungal/microorganismic cell type differentiation	11.6%
1.05	C-compound and carbohydrate metabolism	11.2%
14.07	protein modification	10.9%
32.01	stress response	10.3%
Third level		
11.02.03	mRNA synthesis	17.7%
10.03.01	mitotic cell cycle and cell cycle control	14.5%
11.04.03	mRNA processing (splicing, 5'-, 3'-end processing)	11.8%
43.01.03	fungal and other eukaryotic cell type differentiation	11.6%
10.01.05	DNA recombination and DNA repair	10.0%
16.01	protein binding	9.8%
16.03.03	RNA binding	9.2%
34.11.03	chemoperception and response	7.6%
01.05.01	C-compound and carbohydrate utilization	7.4%
01.04.01	phosphate utilization	7.3%
Fourth level		
11.02.03.04	transcriptional control	10.4%
16.01	protein binding	9.8%
16.03.03	RNA binding	9.2%
43.01.03.05	budding, cell polarity and filament formation	9.2%
11.04.03.01	splicing	7.6%
10.03.01	mitotic cell cycle and cell cycle control	7.3%
01.04.01	phosphate utilization	7.3%
99	UNCLASSIFIED PROTEINS	7.1%
11.04.01	rRNA processing	6.3%
10.01.05.01	DNA repair	6.2%

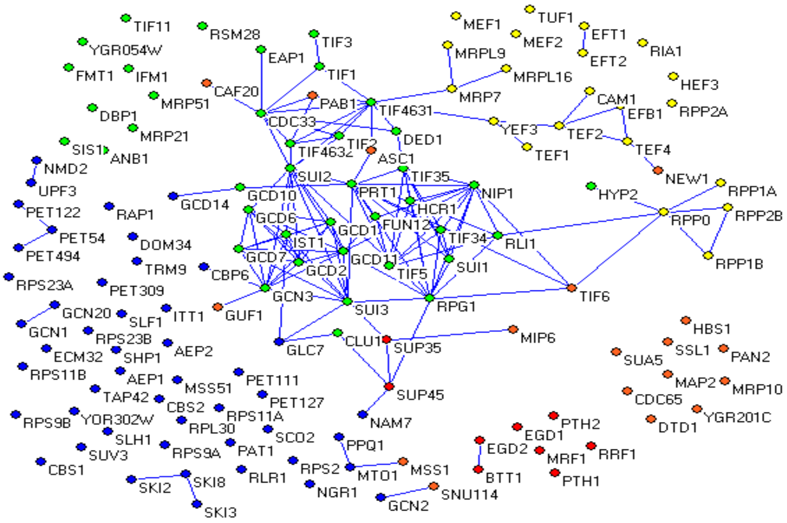

Figure 5-3 Network N1247S. All proteins are in MIPS functional categories of 12.04 (translation, orange), 12.04.01 (translation initiation, green), 12.04.02 (translation elongation, yellow), 12.04.03 (translation termination, red), and 12.07 (translational control, blue).

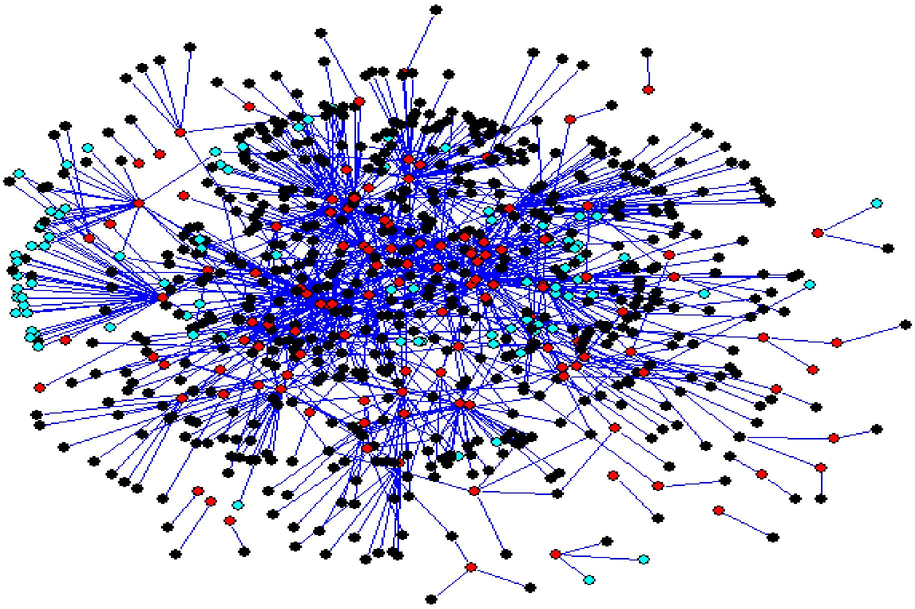

Figure 5-4 Network N1247SA. At least one of the interacting proteins is in N1247. Proteins in N1247 are indicated by red. Proteins in N12 but not in N1247 are indicated by cyan. All other proteins are black.

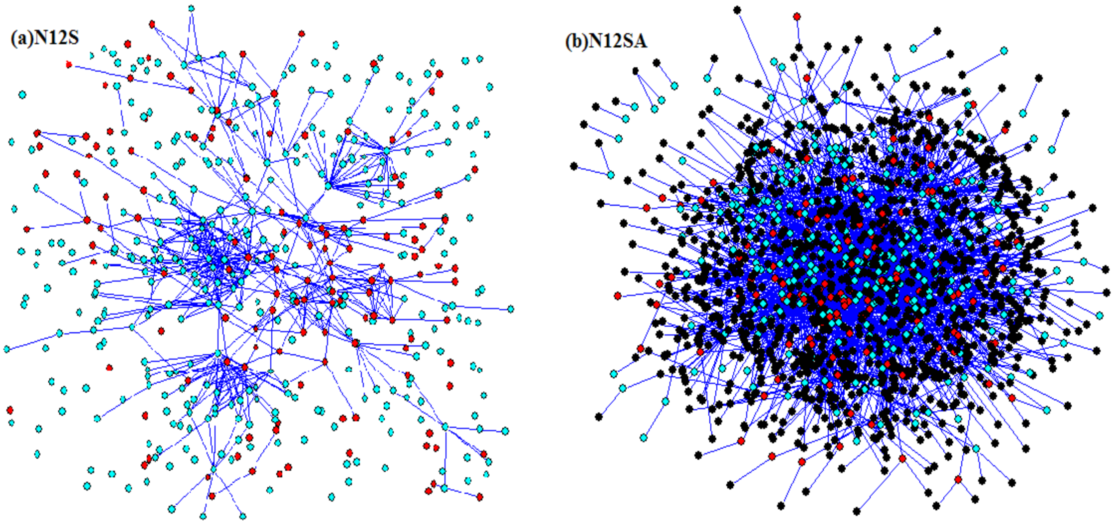

Figure 5-5 Networks N12S and N12SA. (a) Network N12S represents proteins in N12. (b) Network N12SA contains all proteins that are either in N12 or interacting with proteins in N12. For both networks, proteins in N1247 are in red, remaining N12 proteins are in cyan, all other proteins are in black.

Table 5-4 Properties of PTRN.

Network	N1247S	N1247SA	N12S	N12SA	Full
# of proteins	136	798	479	1547	4948
# of unique interactions	152	1100	543	2715	18817
# of proteins in giant component	56	714	218	1394	4857
# of loner proteins	60	13	230	76	0
Average degree <*k*>	2.24	2.76	2.27	3.51	7.61
Diameter <*l*>	3.57	4.67	4.36	4.79	4.07

5.3.3 Essentiality of proteins in PTRN

Network degree (or connectivity) has long been related to protein essentiality [Jeong et al 2001]. Therefore, we examine here the essentiality of proteins in the translation networks using a gene disruption data set downloaded from MIPS. As shown in Figure 5-6, about 28% of the proteins in the translation networks are lethal to disruption and 70% of them are non-essential.

We also examine the essentiality of the loner proteins. As one can expect, the percentage of loner proteins that are essential decreases significantly. Only 15% of the loner proteins are essential. In three networks studied, the average degrees of essential proteins are significantly higher than those of non-essential proteins (Figure 5-7), demonstrating that more connected proteins (with higher degrees) are more likely to be essential.

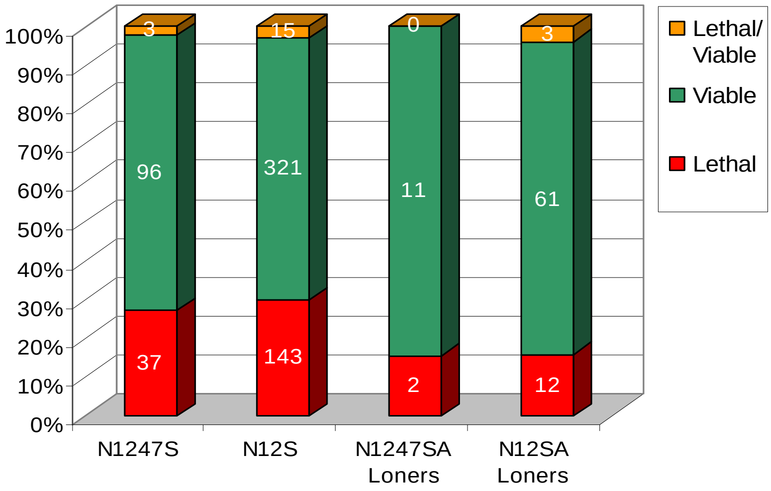

Figure 5-6 Essentiality of proteins in PTRN.

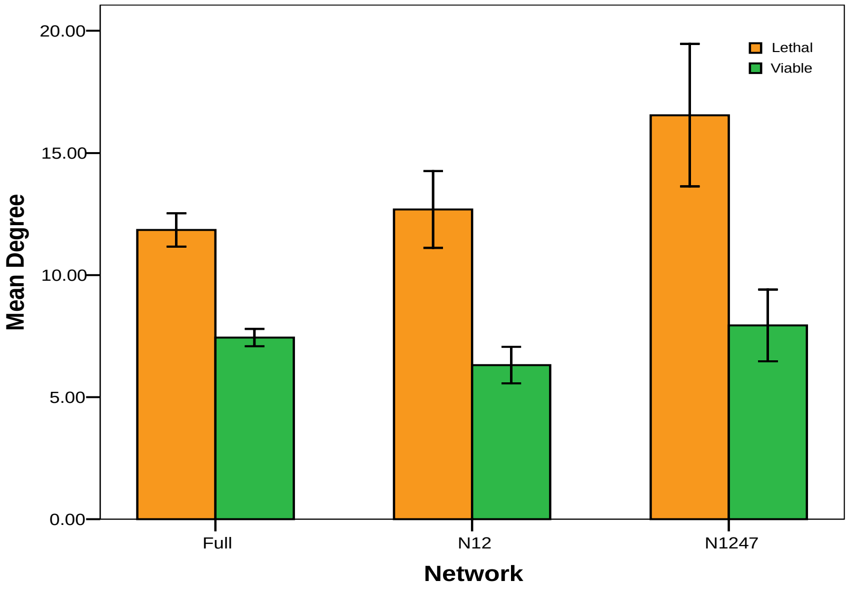

Figure 5-7 Essentiality of proteins in PTRN. Error bar at 95% confidence intervals, $p < 0.05$, between lethal and viable proteins in all networks (ANOVA test).

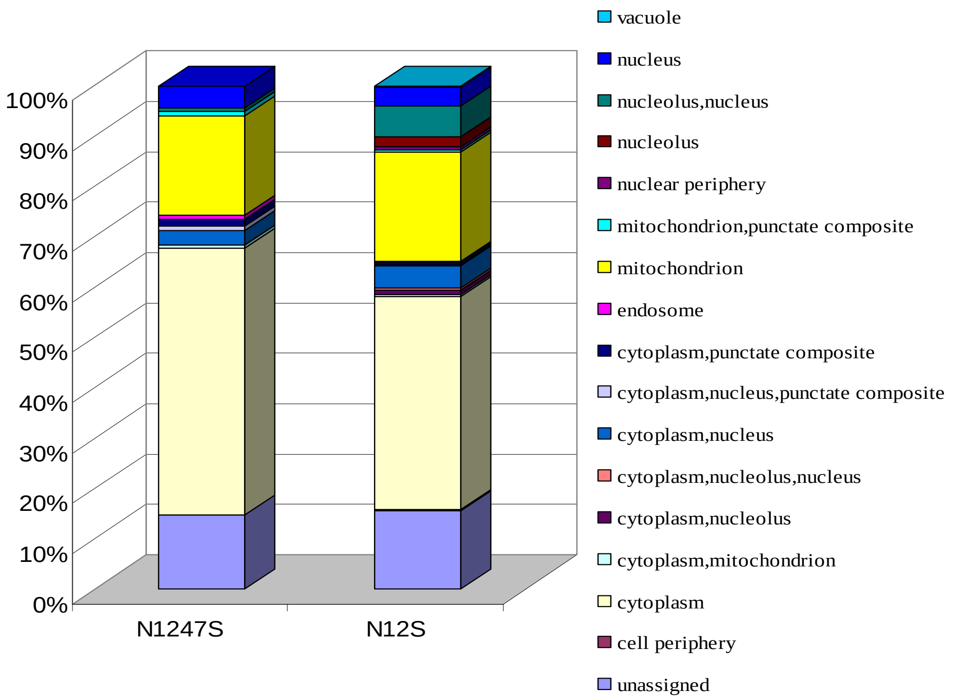

Figure 5-8 Cellular localization of proteins in translation regulatory network.

5.3.4 Cellular localization of proteins in PTRN

As one may expect, most of the proteins in the translation networks are located in cytoplasm and mitochondria (Figure 5-8). However, since the translation machinery in cells is highly complex and translational control may involve many different mechanisms, we see a variety of distributions of proteins in such locales as nucleolus and nucleus.

5.3.5 Protein phosphorylation and PTRN

Protein phosphorylation is a major regulatory mechanism that controls many basic cellular processes. A phosphorylation map for yeast is recently generated by [Ptacek et al 2005]. By using their data, we map the proteins in translation networks to either kinases or substrates for kinases. About 22% of the proteins in translation networks are identified substrates for protein kinases. Even though neither N1247S nor N12S contains any of the 87 kinases testing the map we used, there are 12 proteins in N1247SA and 22 proteins in N12SA are in deed protein kinases. In addition, 31 proteins in N1247S are substrates for 30 different kinases; 69 kinases can phosphorylate 190 proteins in N1247SA; 56 kinases can phosphorylate 105 proteins in N12S and 361 proteins are substrates for 78 kinases (that is almost 90% of the 87 kinases in the yeast kinase-substrate map).

5.4 Discussion

In this chapter, we present a systematic global analysis of protein translation networks in yeast. As far as we know, this is the first report of this kind of study.

We first construct the full protein-protein interaction network and examine the translation proteome in the context of this full network. We define the translation proteome by using the MIPS functional category. The average degree of the protein interaction network containing the major translation-related proteins is significantly higher than an expanded translation network and the full network. While the full network is scale-free, the degree distributions of the translation networks do not display clear power law behavior. The clustering coefficients of the translation networks indicate non-random or hierarchical structures of underlying networks. Reconstruction and analysis of the translation networks clearly demonstrate: 1) the existence of clusters corresponding to different stages of the translation process; 2) the close relationship between translation machinery and other cellular processes especially transcription and metabolism; and 3) the relationship between the translation networks and protein phosphorylation.

This work is the first step in our effort to elucidate the structure and properties of the protein translation networks. Such effort may facilitate the computational dissection of translation networks and provide new insights into mechanisms of translational control from a system's perspective.

CHAPTER 6. ANALYZING PROTEIN TRANSLATION REGULATORY NETWORKS USING HIERARCHICAL RANDOM GRAPHS

6.1 Introduction

In Chapter 5, we present a global network analysis of Protein Translation Regulatory Networks (PTRN) in yeast. In this chapter, we extend our efforts to study another important network feature of the PTRN: hierarchy.

Complex networks are believed to be organized through multiple scales. On the smallest scale is the collection of individual nodes. Some basic properties such as node degree can provide information about these single nodes. The next scale arises when nodes are interacting with each other in pairs. With three or more nodes coming into play, we may see the next scale such as network motifs. Larger groups of nodes form modules. Hierarchy describes such multi-scale organization by explaining how single nodes connect to each other to form motifs, how motifs in turn are organized into modules, and how modules are combined to shape the entire network.

Biological processes are hierarchically organized, evident from interactions between molecules within a cell to relationships among members of an ecological system, and hierarchical structure plays an important role in the dynamics of these processes.

Active research has been done to assess whether a network is actually organized in a hierarchical manner and to identify the different levels in the hierarchy. The majority of the work has been focusing on identifying ‘‘global signatures’’ of a hierarchical network architecture. For example, [Ravasz et al 2002] have shown that the metabolic network of several organisms can be organized into highly connected modules that hierarchically combine into larger units. They have demonstrated that the uncovered hierarchical modularity closely overlaps with known metabolic functions in *E. coli*.

Out of many methods proposed to investigate the hierarchical organization in a network [Guimerà et al 2005; Soffer et al 2005; Sales-Pardo et al 2007], an especially appealing one is the algorithm based on hierarchical random graph model introduced by [Clauset et al 2007; Clauset et al 2008].

In a nutshell, the basis of the hierarchical random graph model is how two nodes in the network are connected. Each pair of nodes in a non-hierarchical random graph is connected with the same probability. However, in a hierarchical random graph, the probability of any two nodes connecting to each other is no longer a constant. Instead, they are connected with a hierarchy of probabilities.

One advantage of hierarchical random graph model is its high flexibility. With suitable inner probabilities selected, this model has demonstrated to be able to capture most of the currently known network characteristics: degree distributions, degree–degree correlations, undirected motifs, and modules.

What makes this algorithm especially appealing and valuable is its ability to detect false positives and missing links in relation to biological networks.

Despite the vast efforts and progress in high throughput technologies, available cellular information is still sparse. Even in such well-studied organism as *Saccharomyces cerevisiae*, the gene regulatory or the protein–protein interaction network is far from complete, evident also from our analysis about protein translation regulatory network in Chapter 5. Besides the missing links (false negatives) in data, every experimental method may also induce unavoidable biases. For example, low throughput experiments for protein–protein interaction measurements tend to focus on well-known proteins due to constraints in cost and time, whereas high throughput experiments without quality control are notorious for producing false positives [Vidal 2001]. The false negatives or false positives in the network naturally will impact the outcome of network analysis.

Using the hierarchical random graph model and the associated hierarchy of link probabilities, the algorithm by [Clauset et al 2008] allows for the discovery of false positive and false negative links. False positives are identified by the links that exist but with low link probability found by the method. False negatives are identified by the links that do not exist in the network but have high link probabilities.

In the following, we define a PTRN that contains proteins involved in translational regulation and controls. We then describe the hierarchical random graph model and the adapted approach we use based on this model to infer the hierarchical structure of the constructed network and further to predict missing links within the network.

6.2 Methods

6.2.1 Data sets

We use the same data sets as used in Chapter 5. Again, the yeast protein-protein interactions data were downloaded from the General Repository for Interaction Datasets (GRID). We select GRID because it contains arguably the most comprehensive data. The GRID database includes all published large-scale interaction datasets as well as available curated interactions such as those deposited in BIND and MIPS. The yeast dataset we downloaded has 4,948 distinct proteins and 18,817 unique interactions. From this network, we derive the protein translation networks which contain proteins with MIPS functional categories related to protein translation as described next.

6.2.2 Construction of PTRN

We extract proteins that are involved in protein biosynthesis from MIPS functional category database as shown in Table 6-1. The extracted proteins belong to the following categories: 12.04 (translation), 12.04.01 (translation initiation), 12.04.02 (translation elongation), 12.04.03 (translation termination), and 12.07 (translational control). There are totally 133 unique proteins in this dataset. We then build the network by using protein-protein interaction data, including interactions among the selected proteins only and ignoring all other interactions. With the exclusion of the isolated proteins – those without any edges connecting to them – and self-looping interactions, the resulted network contains 108 nodes and 342 edges.

There are several reasons for such a construction. First of all, our interest in this research has been focused on protein translation regulatory networks. Secondly, protein-protein interaction data are notorious noisy and incomplete. The approach we take allows us not only to study the hierarchy but also to predict missing links even with the noise and incompleteness in the background. At current stage, it is also more feasible computationally with networks of smaller sizes. In addition, we want to examine if hierarchical structure exists even in such isolated sub-networks.

Table 6-1 MIPS functional categories related to protein translation. A protein may belong to more than one category. The number of proteins is the number of entries stored in each category.

Category	Description	# of Proteins
12.04	translation	88
12.04.01	translation initiation	40
12.04.02	translation elongation	21
12.04.03	translation termination	9
12.07	translational control	55

6.2.3 Hierarchical random graphs

Our approach is based on a hierarchical random graph proposed by [Clauset et al 2008], incorporating with work by [Sales-Pardo et al 2007]. There are two important assumptions in this approach. Firstly, if a network has sub-networks with an equal probability connecting them, then the network can be represented by splitting off the sub-network one at a time until the last one. Secondly, there may be more than one hierarchical random graph that best fits the observed network data.

In hierarchical random graphs, the probabilities of connecting any two nodes and sub-networks are independent of the presence or absence of other connections. This is similar to the classical Erdös-Rényi random graph. However, in the hierarchical random graph, the probabilities are dependent on the topological structure of the graph.

6.2.3.1 Graph notation

We user the same graph notations defined in previous chapters.

6.2.3.2 Definition of a hierarchical random graph

Let $G=(V,E)$ be a graph. Let n be the size of node set, n = $|V|$. Let D be the dendrogram with n leaves representing nodes of G. Let r be an internal node of D with a probability P_r which denotes the probability that an edge exists between two nodes sharing *r* as their lowest common ancestor in *D*. A hierarchical random graph is thus defined by $(D,\{P_r\})$.

6.2.3.3 Inferring the hierarchical structure

As stated earlier, one assumption is that the likelihood of all hierarchical random graphs is a priori equal. By Bayes' theorem, the probability that a model $(D,\{P_r\})$ explains the observed data is proportional to the posterior probability or likelihood L.

Let E_r be the number of edges in *G* with *r* as their lowest common ancestor, L_r and R_r be the numbers of leaves in the left and right sub-trees rooted at *r* in *D*. We have

$$L(D,\{p_r\}) = \prod_{r \in D} p_r^{\,E_r} (1 - p_r)^{L_r R_r - E_r}$$

For each internal node *r* in *D*, the probability P_r is defined as

$$P_r = \frac{E_r}{L_r R_r}.$$

Thus, the likelihood of *D* is

$$L(D,\{p_r\}) = \prod_{r \in D} [\overline{p_r}^{\,\overline{p_r}} (1 - \overline{p_r})^{1-\overline{p_r}}]^{L_r R_r}\ .$$

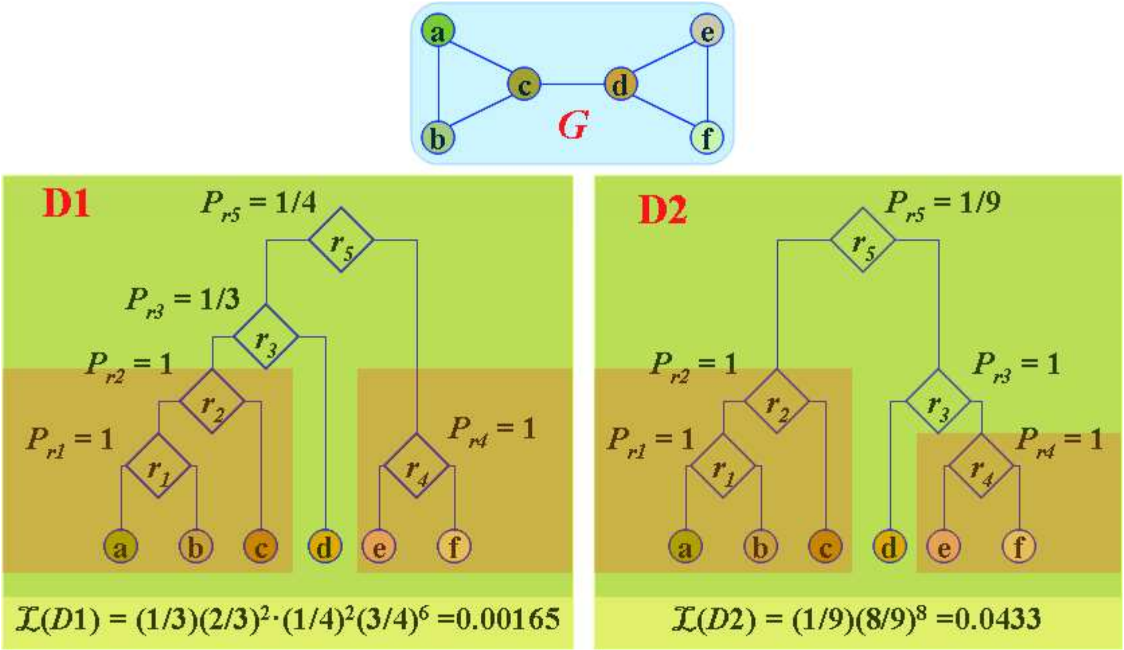

Figure 6-1 A sample network *G* and the likelihood of two possible dendrograms D1 and D2 (Modified from [Clauset et al 2007]).

Conveniently, instead of using the above equation directly, we use its logarithm form:

$$\log L(D) = -\sum_{r \in D} L_r R_r h(\overline{P_r}) \,.$$

6.2.3.4 Markov chain Monte Carlo method

Since it is an NP hard problem to maximize $\mathcal{L}(D, \{P_r\})$, the estimation is done by using a Markov chain Monte Carlo method by sampling D with probability proportional to their likelihood $\mathcal{L}(D)$.

To create the Markov chain we need to pick a set of transitions between possible dendrograms, as illustrated by Figure 6-1. To accept or reject a newly generated dendrogram, we evaluate the likelihood change, as done by [Clauset et al 2008],

$$\Delta\log\mathcal{L} = \log\mathcal{L}(D') - \log\mathcal{L}(D).$$

If $\Delta\log\mathcal{L}$ is nonnegative, meaning that the newly generated D' is at least as likely as D; otherwise, the transition is not accepted.

There are typically many dendrograms with roughly equal likelihoods. In order to increase the effectiveness of the acceptance and rejection rate, we incorporate the idea used in [Sales-Pardo et al 2007] by introducing the modularity change, ΔQ. Modularity of a network with m modules is defined [Newman and Girvan 2004, Newman 2006] as

$$Q = \sum_{i=1}^{m}\left[\frac{l_i}{L} - \left(\frac{d_i}{2L}\right)^2\right],$$

where L is the total number of edges in the network, l_i is the number of edges within module i, d_i is the sum of degrees of all of the nodes inside module i.

With networks of relative small sizes, the Markov chain converges fairly quickly. Therefore, it is suitable for our constructed PTRNs.

6.2.4 The hierarchical random graph algorithm

We now describe the algorithm that we developed to explore the hierarchical structure of PTRN networks and to predict missing links within these networks in Figure 6-2.

6.3 Results

6.3.1 Fitting the hierarchical random graph to data

We construct our protein translation network using protein-protein interactions among extracted proteins and then fit the hierarchical random graph model to the constructed network. Figure 6-3 shows an example of maximum likelihood dendrogram with *logL* = -539. The dendrogram clearly divides the majority of proteins into groups coherent to their MIPS function categories.

```
1     Construct PTRN.
2     Fit PTRN to Markov chain Monte Carlo (MCMC) algorithm and obtain a fitted
      hierarchical random graph dendrogram $D_f$.
3     $D_f$ is fed back into MCMC.
4     Sample dendrograms at regular intervals thereafter from those generated by MCMC.
5     A consensus dendrogram $D_c$ is obtained by outputting only the dendrogram features in
      the majority of the sampled models.
6     To predict missing links, do the following:
6.1   Find all pairs of nodes that do not have edges between them in the original network.
6.2   For each pair of nodes $i, j$ found in 6.1, calculate the average probability $<P_{ij}>$ over the
      corresponding probabilities, $P_{ij}$, in each of the sampled dendrograms.
6.3   Output the pairs $i, j$ in decreasing order of $<P_{ij}>$.
```

Figure 6-2 The hierarchical random graph algorithm.

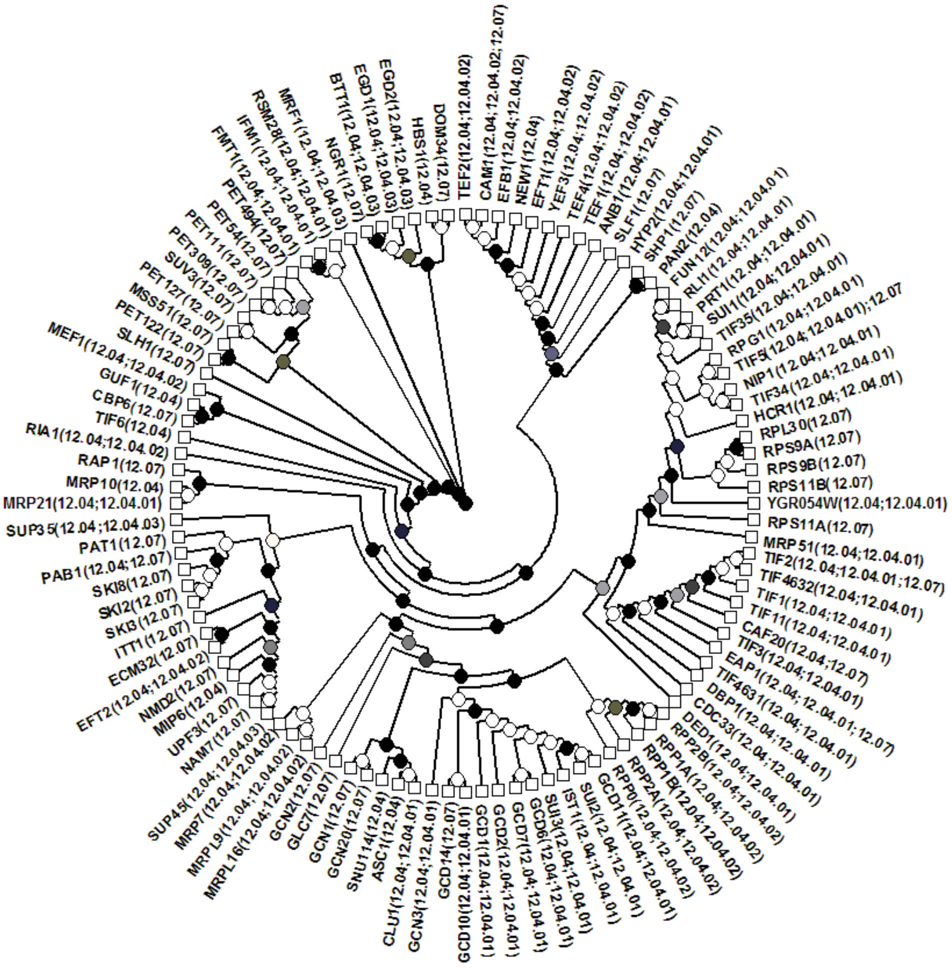

Figure 6-3 An example of maximum likelihood dendrogram with logL= -539. The leaves are labeled with protein names with corresponding MIPS function categories in parentheses. The probabilities are shown as gray-scale values.

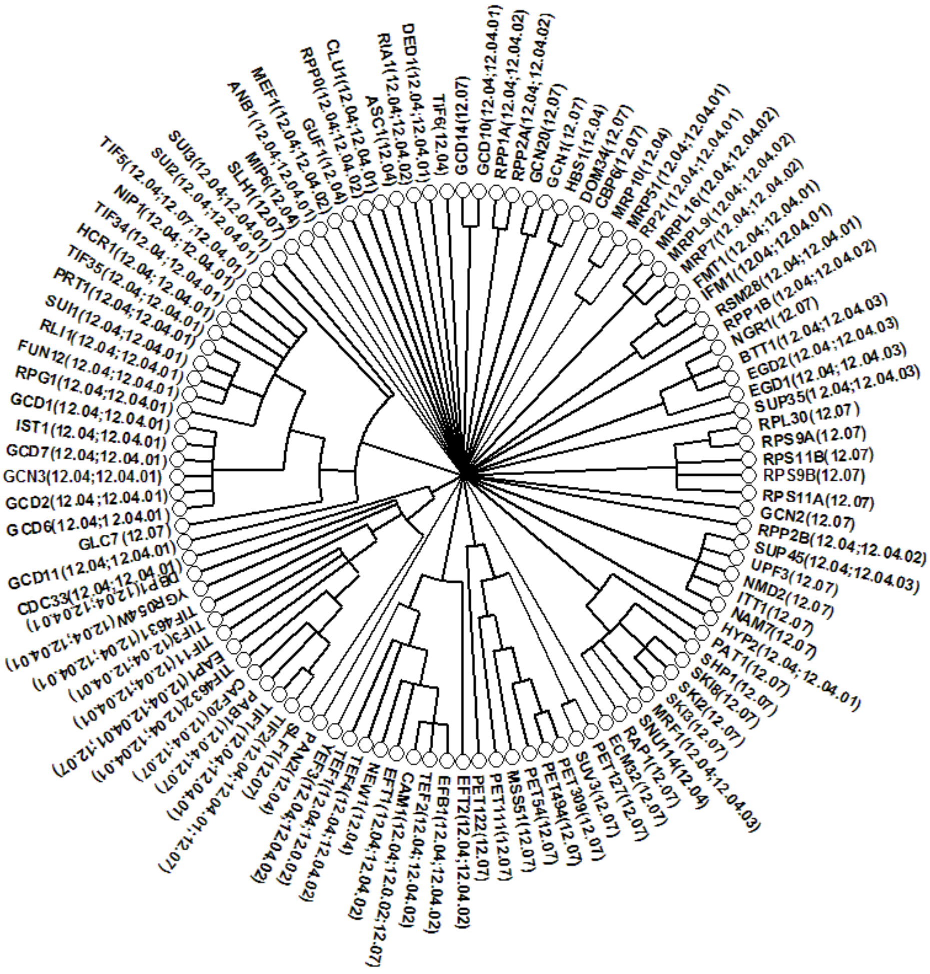

Figure 6-4 The consensus dendrogram.

6.3.2 Consensus dendrogram

Figure 6-4 shows an example of a consensus dendrogram constructed from the sampled hierarchical random graphs. A consensus dendrogram is a summary of a set of dendrograms that fit the observed data. We may expect it to capture the topological features consistent across the majority of the dendrograms and can better characterize the structure of the network than any individual dendrogram.

6.3.3 Prediction of missing links

The most interesting and possibly the most useful application of hierarchical random graphs is the prediction of missing interactions in networks in which the available information is incomplete as in the case of protein-protein interaction data, especially in our case of studying protein translation regulatory networks. Table 6-2 is the compiled result of top 15 possible missing links with the highest probabilities from 10 runs of the predicting algorithms.

On top of the list is the interaction between SUP35 and PAT1. SUP35 is translation termination factor eRF3, involved in the termination of protein translation. PAT1 is topoisomerase II-associated deadenylation-dependent mRNA-decapping factor. It is required for faithful chromosome transmission, maintenance of rDNA locus stability, and protection of mRNA 3'-UTRs from trimming. There is no interaction between these two proteins in our downloaded data sets. However, this interaction has been reported rather recently [Wilmes et al 2008].

An intriguing finding of the prediction results is that a few proteins have multiple highly probable missing links, such as GCD11, SUI3, SUI2, RLI1, IST1, and HCR1. GCD11 is the gamma subunit of the translation initiation factor eIF2, involving in the identification of the start codon. Its interaction with HCR1 has been reported recently [Wilmes et al 2008]. RLI1 is an essential iron-sulfur protein required for ribosome biogenesis and translation initiation. Its interaction with SUI3 is also reported [Wilmes et al 2008]. SUI3 is the beta subunit of the translation initiation factor eIF2, involved in the identification of the start codon and possibly in mRNA binding as well. HCR1 is a dual function protein involved in translation initiation as a substoichiometric component (eIF3j) of translation initiation factor 3 (eIF3) and is required for processing of 20S pre-rRNA. The interaction between SUI3 and HCR1 has also been reported [Wilmes et al 2008].

6.4 Discussion

In this chapter, we present the exploratory analysis of a protein translation regulatory network using hierarchical random graphs.

Table 6-2 Prediction of missing links.

Protein 1	Protein 2	Probability	Reference
SUP35	PAT1	0.8895	[Wilmes et al 2008]
RLI1	PRT1	0.7343	
GCD11	IST1	0.7163	
GCD11	SUI1	0.7161	
GCD11	RLI1	0.7159	
GCD11	HCR1	0.7157	[Wilmes et al 2008]
SUI3	HCR1	0.6977	
SUI3	TIF35	0.6976	
SUI3	FUN12	0.6976	
SUI3	RLI1	0.6976	[Wilmes et al 2008]
SUI3	TIF34	0.6973	
SUI2	FUN12	0.6683	
SUI2	HCR1	0.6682	[Wilmes et al 2008]
SUI2	IST1	0.6681	
SUI2	SUI1	0.6681	

We constructed a protein translation network by extracting proteins categorized in MIPS function database [Mewes et al 2002] and protein-protein interaction data curated in BioGRID [Bader et al 2003]. One important feature of such reconstructed networks is its incompleteness. Our current knowledge about the links may only be a fraction of all interactions among these proteins that may exist in reality. It thus is an enormous challenge to study such partial networks.

As shown in Figure 6-3, by using the hierarchical random graphs, the reconstructed dendrogram divided the majority of proteins into groups corresponding to their MIPS function categories. Our results clearly demonstrated 1) the existence of the hierarchical structure in the constructed protein translation network; and 2) the usefulness of the hierarchical random graph model in exploring the network structure.

Our results also show the ability of predicting missing links in networks by using the hierarchical random graph. At least four of the top 15 predicted missing links has been reported recently [Wilmes et al 2008]. It is very beneficial for experimental biologists to use such drastically narrowed list to formulate and validate hypotheses. One of our future work will be to collaborate with biologists to validate the predicted missing links and eventually help build up a much more complete translation regulatory network.

A limitation of current approach using Markov chain Monte Carlo is its high computational cost. Improving the computation efficiency in the future will allow us to apply this approach to larger networks.

CHAPTER 7. CONCLUSIONS AND FUTURE WORK

7.1 The summary of topological analysis of biological networks

In order to discover and validate the topological structure of biological networks, we have started this thesis with a comprehensive evaluation of the topological structure of protein-protein interaction (PPI) networks by mining and analyzing graphs constructed from the popular data sets publicly available to the bioinformatics research community. We compare the topology of these networks across different species, different confidence levels, and different experimental systems used to obtain the interaction data. In regard to the characterized scale-free signature, our results confirm that the degree distribution follows a power law in some of the networks whereas not in others. Furthermore, further statistical analysis shows that residues are not independent on the fit values, indicating that the power law model may be inadequate. Our results also show that the dependence of the average clustering coefficient on the node degree is far from a power law, contradicting many published results. For the first time, we show that the average node density exhibits a strong powder law dependence on the node degree for all the networks studied, regardless of species, confidence levels, and experimental systems.

7.2 The summary of module detection

Modular structure is a topological property common to many networks. We have developed in this thesis an efficient and accurate approach to detect a module in a protein-protein interaction network from a given seed protein. Our experimental results show strong structural and functional relationships among member proteins within each of the modules identified by our approach, as verified by MIPS complex catalogue database and annotations.

In addition, the experimental results show that the performance of our approach is superior in terms of recall and precision to a published method, *Complexpander*. The usefulness of our approach is also demonstrated by its successful applications in mining, dynamic fuzzy simulation [Hu et al 2007], and predictive modeling of biological networks from biomedical literature databases [Hu and Wu 2007].

7.3 The summary of analyzing protein translation regulatory networks

7.3.1 Global analysis of protein translation regulatory networks in yeast

Protein translation is a vital cellular process for any living organism. The maturation of high-throughput technologies and the success of genome projects make it possible to apply computational approaches to the study of biological systems. The availability of interaction databases in particular provides an opportunity for researchers to exploit the immense amount of data *in silico* such as studying biological networks using network analysis. There has been an extensive effort using computational methods in deciphering the transcriptional regulatory networks. However, research on translation regulatory networks has caught little attention in the bioinformatics and computational biology community probably due to the nature of available data and the bias of the conventional wisdom. In this paper, we present a global network analysis of protein translation networks in yeast, a first step in attempting to facilitate the elucidation of the structures and properties of translation networks. We extract the translation proteome using MIPS functional category and analyze it in the context of the full protein-protein interaction network. We further derive the individual translation networks from the full interaction network using the extracted proteome. We show that the protein translation networks do not exhibit power law degree distributions in contrast to the full network. In addition, we demonstrate the close relationship between the translation networks and other cellular processes especially transcription and metabolism. We also examine the essentiality and its correlation to connectivity of proteins in the translation networks, the cellular localization of these proteins, and the mapping of these proteins to the kinase-substrate system. These results have potential implications for understanding mechanisms of translational control from a system's perspective.

7.3.2 Hierarchical structure in protein translation regulatory networks

In this thesis, we present an exploratory analysis of yeast protein translation regulatory networks using hierarchical random graphs. We derive a protein translation regulatory network from a protein-protein interaction dataset. Using a hierarchical random graph model, we show that the network exhibits well organized hierarchical structure. In addition, we apply this technique to predict missing links in the network.

The hierarchical random graph mode can be a potentially useful technique for inferring hierarchical structure from network data and predicting missing links in partly known networks. The results from the reconstructed protein translation regulatory networks have potential implications for better understanding mechanisms of translational control from a system's perspective.

7.4 The contribution of the thesis

The contribution of this thesis work is many-fold.

First, we develop a novel approach to detect a protein module from a given seed. The proposed method is very efficient and effective in terms of finding a structural and functional coherent protein module. It is also very scalable, performing very well for large networks as well as smaller ones.

Second, by performing a comprehensive evaluation of protein-protein interaction networks, we report a couple of new discoveries about the topological behaviors, naming the scale-free node density distribution over degree and the inadequacy of the well-received scale-free feature in describing the degree distribution.

Third, as far as we know, we are the first one in attempting to identify, characterize, and reconstruct the protein translation regulatory networks through network analysis and mining. We have not noticed other published research on this subject. This is indeed a task proved to be more challenging than we initially anticipated. It demands quite in-depth domain knowledge both in biology and computational fields. Nonetheless, we have been able to analyze and mine the protein translation regulatory networks. In addition, we have explored in identifying the hierarchy structure as well as missing and false positive links in these reconstructed networks.

7.5 Future work

In this thesis, we have developed an efficient approach to detect a protein module from a given seed. We apply this approach on finding several protein modules whose members demonstrate structural and functional coherency. However, there are certain aspects about this approach needs additional work. First is to adapt it so that it may be used to discover all possible modules in the entire network. Second is to modify it so that when it is used to identify all modules, it will allow identifying modules with overlapping members.

As to the protein translation regulatory networks, a lot of work remains to be done. First, efficient computational methods are needed to extract and reconstruct PTRN to a reasonably complete state. These include finding the component proteins and the missing links, especially for those “lower” proteins. Second, in regard to the hierarchical random graph model, develop a more efficient algorithm to improve the performance in the probability fitting step. Third, develop an efficient computational framework so that it may be used to study the hierarchy architecture on all the scales, from the individual component scale to the medium scales (such as motifs, modules) and all the way to the entire network scale. Ideally, details about members at each lower scale will be encapsulated and hidden from a higher scale.

List of References

[Albert and Barabási 2002]Albert R and Barabási A-L: Statistical mechanics of complex networks. *Rev. Mod. Phys*. 2002, 74: 67–97.

[Alm and Arkin 2003] Alm E, Arkin AP: Biological networks. *Curr Opin Struct Biol*. 2003, 13(2):193-202.

[Aloy and Russell 2002] Aloy P and Russell RB: Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA* 2002, 99:5896-5901.

[Aloy and Russell 2003] Aloy P and Russell RB: InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 2003, 19:161-162.

[Asthana et al 2004] Asthana S, King OD, Gibbons FD, Roth FP: Predicting Protein Complex Membership Using Probabilistic Network Reliability. *Genome Res*. 2004, 14: 1170-1175

[Bader 2003] Bader JS: Greedily building protein networks with confidence. *Bioinformatics* 2003, 19(15): 1869-1874.

[Bader et al 2003] Bader GD, Betel D, and Hogue CW: BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2003, 31(1): 248-250.

[Bader and Hogue 2003] Bader GD and Hogue CW: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4:2

[Bader et al 2004] Bader JS, Chaudhuri A, Rothberg JM, Chant J: Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 2004, 22(1):78-85.

[Barabási and Albert 1999] Barabási A-L and Albert R: Emergence of scaling in random networks. *Science* 1999, 286:509-12.

[Batagelj and Mrvar 1998] Batagelj V and Mrvar A: Pajek: Program for large network analysis. *Connections* 1998, 21: 47–57.

[Barabási and Oltvai 2004] Barabási A-L and Oltvai ZN: Network biology: understanding the cell's functional organization. *Nature Rev Genet* 2004, 5:101-114.

[Bock and Gough 2001] Bock JR and, Gough DA: Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001, 17: 455-460.

[Bork et al 2004] Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, and Marcotte EM: Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 2004, 14: 292–299.

[Breitkreutz et al 2003] Breitkreutz B-J, Stark C, Tyers M: The GRID: The General Repository for Interaction Datasets. *Genome Biol* 2003, 4: R23.

[Brohee and van Helden 2006] Brohee S and van Helden J: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7: 488.

[Bu et al 2003] Bu D, et al: Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.* 2003, 31: 2443–2450.

[Cabusora et al 2005] Cabusora L, et al: Differential network expression during drug and stress response. *Bioinformatics* 2005, 21: 2898–2905.

[Clauset et al 2007] Clauset A, Moore C, Newman MEJ: Structural inference of hierarchies in networks. In *Lecture Notes in Computer Science*. Edited by E.M. Airoldi et al: Springer-Verlag, Berlin Heidelberg; 2007, 4503:1–13

[Clauset et al 2008] Clauset A, Moore C, Newman MEJ: Hierarchical structure and the prediction of missing links in networks. *Nature* 2008, 453: 98-101.

[Colizza et al 2005] Colizza V, Flammini A, Maritan A, Vespignani A: Characterization and modeling of protein-protein interaction networks. *Physica A* 2005, 352: 1-27.

[Danon et al 2005] Danon L, Díaz-Guilera A, Duch J, Arenas A: Comparing community structure identification. *J. Stat Mech* 2005, P09008 doi: 10.1088/1742-5468/2005/09/P09008.

[Deane et al 2002] Deane CM, Salwinski L, Xenarios I, and Eisenberg D: Protein-protein interactions – two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* 2002, 1: 349.

[Dittrich et al 2008] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T: Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 2008, 24(13): i223-i231.

[Daudin et al 2008] Daudin JJ, Picard F, Robin S: A mixture model for random graphs. *Statistics and Computing* 2008, 18: 173-183.

[de Jong 2002] de Jong H: Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002, 9: 67-103.

[Donetti and Munoz 2004] Donetti L, Munoz MA: Detecting Network Communities: a new systematic and efficient algorithm. *J. Stat. Mech.* P10012. 50.

[Dorogovtsev and Mendes 2002] Dorogovtsev SN and Mendes JFF: Evolution of networks. *Adv. Phys.* 2002, 51: 1079–1187.

[Erdös and Rényi 1959] Erdös P and Rényi A: On random graphs I. *Publ Math*, 1959. 6: 290-297.

[Espadaler et al 2005] Espadaler J, Romero-Isart O, Jackson RM, Oliva B: Prediction of protein-protein interactions using distant conservation ofsequence patterns and structure relationships. *Bioinformatics* 2005, 21: 3360-3368.

[Fell and Wagner 2000] Fell DA and Wagner A: The small world of metabolism. *Nat. Biotechnol*. 2000, 189: 1121–1122.

[Flake et al 2002] Flake GW, Lawrence SR, Giles CL, Coetzee FM: Self-organization and identification of Web communities, *IEEE Computer* 2002, 35: 66-71.

[Fraser et al 2002] Fraser HB, et al: Evolutionary rate in the protein interaction network. *Science* 2002, 296: 750.

[Fraser et al 2003] Fraser HB, et al: A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.* 2003, 3: 11.

[Gandhi et al 2006] Gandhi TK, et al: Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006, 38: 285-293.

[Gavin et al 2002] Gavin A-C, et al: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, 415: 141–147.

[Giaever et al 2002] Giaever G, Chu AM, Ni L, Connelly C, Riles L: Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002, 418: 387–391.

[Giot et al 2003] Giot L, et al: A protein interaction map of *Drosophila melanogaster*. *Science* 2003, 302: 1727-36.

[Girvan and Newman 2002] Girvan M and Newman MEJ: Community structure in social and biological networks. *Proc Natl Acad Sci USA* 2002, 99: 7821-7826.

[Guglielmi et al 2004] Guglielmi B, van Berkum NL, Klapholz B, Bijma T, Boube M, Boschiero C, Bourbon HM, Holstege FCP, and Werner M: A high resolution protein interaction map of the yeast Mediator complex. *Nucleic Acids Res.* 2004, 32: 5379–5391.

[Guimerà et al 2005] Guimerà R, Sales-Pardo M, Amaral LAN: Modularity from fluctuations in random graphs and complex networks. *Phys Rev E* 2005, 70:025101.

[Guo et al 2007] Guo Z, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, and Wang J: Edge-based scoring and searching method for identifying condition -responsive protein-protein interaction sub-network. *Bioinformatics* 2007, 23: 2121–2128.

[Han et al 2004] Han DS, Kim HS, Jang WH, Lee SD, Suh JK: PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res* 2004, 32:6312-6320.

[Handcock et al 2007] Handcock MS, Raftery AE, and Tantrum JM: Model-based clustering for social networks. *J. R. Statist. Soc. A*, 170: 301-354.

[Holland 2004] Holland EC: Regulation of translation and cancer. *Cell Cycle* 2004, 3: 452-455.

[Holme et al 2003] Holme P, Huss M, and Jeong H: Subnetwork hierarchies in biochemical pathways. *Bioinformatics* 2003, 19: 532–538.

[Hartwell et al 1999] Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell biology. *Nature* 1999, 402: C47–C52.

[Hashimoto et al 2004] Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER: Growing genetic regulatory networks from seed genes. *Bioinformatics* 2004, 20: 1241–1247.

[Ho et al 2002] Ho Y, et al: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002, 415: 180–183.

[Holme et al 2003] Holme P, Huss M, Jeong H: Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 2003, 19(4): 532-538.

[Hu et al 2004] Hu X, Yoo I, Song I-Y, Song M, Han J, Lechner M: Extracting and Mining Protein-Protein Interaction Network from Biomedical Literature, in the *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in*

Bioinformatics and Computational Biology (IEEE CIBCB 2004), Oct. 7-8, 2004, San Diego, USA.

[Hu 2005] Hu X: Mining and Analyzing Scale-free Protein-Protein Interaction Network, *International Journal of Bioinformatics Research and Application* 2005, 1: 81-101.

[Hu and Wu 2007] Hu X, Wu D: Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2007, 4: 251-263.

[Hu et al 2007] Hu X, Sokhansanj B, Wu D, Tang Y: A Novel Approach for Mining and Dynamic Fuzzy Simulation of Biomolecular Network. *IEEE Transactions on Fuzzy Systems* 2007, 15: 1219-1229

[Huang et al 2004] Huang TW, Tien AC, Huang WS, Lee YC, Peng CL, Tseng HH, Kao CY, Huang CY: POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics* 2004, 20: 3273-3276.

[Huh et al 2003] Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: Global analysis of protein localization in budding yeast. *Nature* 2003, 425: 686-691.

[Ideker et al 2002] Ideker T, Ozier O, Schwikowski B, Siegel AF: Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002, 18(Suppl. 1), S: 33–S240.

[Ideker and Sharan 2008] Ideker T and Sharan R: Protein networks in disease. *Genome Res* 2008, 18: 644–652.

[Ito et al 2001] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, and Sakaki Y: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 2001, 98: 4569–4574.

[Jansen et al 2002] Jansen R, Lan N, Qian J, Gerstein M: Integration of genomic datasets to predict protein complexes in yeast. *J Struct Functional Genomics* 2002, 2: 71–81.

[Jeong et al 2000] Jeong H, Tombor B, Albert R, Oltvai ZN, and Barabási A-L: The large-scale organization of metabolic networks. *Nature* 2000, 407: 651–654.

[Jeong et al 2001] Jeong H, Mason SP, Barabási A-L, and Oltvai ZN: Lethality and centrality in protein networks. *Nature* 2001, 411: 41-42.

[Kanehisa and Goto 2000] Kanehisa M and Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, 28: 27–30.

[Kim et al 2002] Kim WK, Park J, Suh JK: Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform Ser Workshop Genome Inform* 2002, 13: 42.

[Kitano 2002] Kitano H: Systems biology: A brief overview. *Science* 2002, 295: 1662–1664.

[Krishnamurthy et al 2003] Krishnamurthy L, Nadeau J, Ozsoyoglu G, Ozsoyoglu M, Schaeffer G, Tasan M, Xu W: Pathways database system: an integrated system for biological pathways. *Bioinformatics* 2003, 19: 930–937.

[Lancichinetti et al 2009] Lancichinetti A, Fortunato S, Kertész J: Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* 2009, 11: 033015.

[Li et al 2004] Li S, et al: A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, 303: 540-543.

[Lockhart and Winzeler 2000] Lockhart DJ, Winzeler EA: Genomics, gene expression and DNA arrays. *Nature* 2000, 405: 827-836.

[Lu et al 2006] Lu H, et al: Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochem Biophys Res Commun* 2006, 345: 302-309.

[Ma et al 2004] Ma H-W, Buer J, Zeng A-P: Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* 2004, 5: 199.

[Ma'ayan et al 2005] Ma'ayan A, et al: Formation of regulatory patterns during signal propagation in a ammalian cellular network. *Science* 2005, 309: 1078-1083.

[Machesky and Gould 1999] Machesky LM and Gould KL: The Arp2/3 complex: a multifunctional actin organizer. *Curr. Opin. Cell Biol.* 1999, 11: 117–121.

[Maraziotis et al 2007] Maraziotis IA, Dimitrakopoulou K, Bezerianos A: Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics* 2007, 8: 408.

[Martin et al 2005] Martin S, Roe D, Faulon JL: Predicting protein-protein interactions using signature products. *Bioinformatics* 2005, 21: 218-226.

[McGraith et al 2000] McGraith S, Holtzman T, Moss B, and Fields S: Genome-wide analysis of *vaccinia* virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 2000, 97: 4879–4884.

[Mehra et al 2003] Mehra A, Lee KH, Hatzimanikatis V: Insights into the relation between mRNA and protein expression patterns: I. Theoretical consideration. *Biotechnol Bioeng* 2003, 84: 822-833.

[Mehra and Hatzimanikatis 2006] Mehra A, Hatzimanikatis V: An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys J* 2006, 90: 1136-1146.

[Merrick 2004] Merrick WC: Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* 2004, 332: 1-11.

[Mewes et al 2002] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, and Weil B: MIPS: A database for genomes and protein sequences. *Nucleic Acids Res* 2002, 30: 31–34.

[Mucha et al 2010] Mucha PJ, Richardsony T, Macony K, Porter MA, Onnela J-P: Community structure in time-dependent, multiscale, and multiplex networks. *Science* 2010, 328: 876-878.

[Nacu et al 2007] Nacu S, Critchley-thorne R, Lee P, and Holmes S: Gene expression network analysis and applications to immunology. *Bioinformatics* 2007, 23: 850–858.

[Najafabadi and Salavati 2008] Najafabadi HS, Salavati R: Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology* 2008, 9: R87.

[Newman 2000] Newman MEJ: Models of the small world. *J. Stat. Phys.* 2000, 101: 819–841.

[Newman 2001] Newman MEJ: The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* 2001, 98: 404–409.

[Newman 2003a] Newman MEJ: The Structure and Function of Complex Networks. *SIAM Review* 2003, 45: 167-256.

[Newman 2003b] Newman MEJ: Random graphs as models of networks, in: S. Bornholdt, H.G. Schuster (Eds.), *Handbookof Graphs and Networks: from the Genome to the Internet*, Wiley-VCH, Berlin, 2003, pp. 35–68.

[Newman 2004a] Newman MEJ: Detecting community structure in networks. *Eur Phys J B* 2004, 38: 321-330.

[Newman 2004b] Newman MEJ: Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69: 066133.

[Newman and Girvan 2004] Newman MEJ, Girvan M: Finding and evaluating community structure in networks. *Phys. Rev. E* 69: 026113.

[Newman 2006] M. E. J. Newman MEJ: Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* 2006, 103: 8577–8582.

[Newman and Leicht 2007] Newman MEJ and Leicht EA: Mixture models and exploratory analysis in networks. *Proc Natl Acad Sci USA* 2007, 104: 9564-9569.

[Ng and Huang 2004] Ng KL, Huang CH: A cross-species study of the protein-protein interaction networks via the random graph approach. *BIBE04*, 2004.

[Ogmen et al 2005] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: PRISM: protein interactions by structural matching. *Nucleic Acids Res* 2005, 33:W331-336.

[Oltvai and Barabási 2002] Oltvai ZN and Barabási AL: Life's complexity pyramid. *Science* 2002, 298: 763–764.

[Pain 1996] Pain VM: Initiation of protein synthesis in eukaryotic cells. *Eur. J. Biochem.* 1996, 236: 747–767

[Palla et al 2005] Palla G, Derenyi I, Farkas I, Vicsek T: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005, 435: 814.

[Pastor-Satorras and Vespignani 2001]Pastor-Satorras R and Vespignani A: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 2001, 86: 3200–3203.

[Pei and Zhang 2005] Pei P and Zhang A: A topological measurement for weighted protein interaction network. *CSB2005*, August 8-11, 2005, Stanford, CA, USA.

[Pitre et al 2006] Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Ashkan Golshani: PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics* 2006, 7: 365.

[Ptacek et al 2005] Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M: Global analysis of protein phosphorylation in yeast. *Nature* 2005, 438: 679-84.

[Przulj et al 2004] Przulj N, Corneil DG, Jurisica I: Modeling interactome: scale-free or geometric? *Bioinformatics* 2004, 20: 3508-3515.

[Radicchi et al 2004] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D: Defining and identifying communities in networks. *Proc Nat Acad Sci USA* 2004, 101: 2658–2663.

[Rajagopalan and Agarwal 2005] Rajagopalan D, Agarwal P: Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* 2005, 21: 788–793.

[Rain et al 2001] Rain JC, Selig L, de Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, and Schachter V: The protein-protein interaction map of *Helicobacter pylori*. *Nature* 2001, 409: 211–215.

[Ravasz et al 2002] Ravazs E, Somera AL, Mongru DA, Oltvai ZN, Balabasi AL: Hierarchical Organization of Modularity in Metabolic Networks. *Science* 2002, 297:1551-1555.

[Ravasz and Barabási 2003] Ravasz E, Barabási A-L: Hierarchical organization in complex networks. *Phys. Rev. E* 2003, 67: 026112.

[Rives and Galitski 2003] Rives AW and Galitski T: Modular organization of cellular networks. *Proc Natl Acad Sci USA* 2003, 100: 1128–1133.

[Sales-Pardo et al 2007] Sales-Pardo M, Guimera R, Moreira AA, Amaral LAN: Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci USA* 2007 104: 15224-15229.

[Shen et al 2009] Shen H, Cheng X, Cai K, Hu M: Detect overlapping and hierarchical community structure in networks. *Physica A* 2009, 388: 1706-1712.

[Soffer and Vazquez 2004] Soffer S, Vazquez A: Clustering coefficient without degree correlations biases. *Phys Rev E* 2005, 71: 057101.

[Sohler et al 2004] Sohler F, et al: New methods for joint analysis of biological networks and expression data. *Bioinformatics* 2004, 20: 1517–1521.

[Spirin and Mirny 2003] Spirin V, Mirny LA: Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, 100: 12123–12128.

[Strogatz 2001] Strogatz AH: Exploring complex networks. *Nature* 2001, 410: 268-276.

[Tanaka et al 2005] Tanaka R, Yi T, and Doyle J: Some protein interaction data do not exhibit power law statistics. *FEBS Lett.* 2005, 579: 5140-5144.

[Thomas et al 2003] Thomas A, Cannings R, Monk NAM, and Cannings C: On the structure of protein-protein interaction networks. *Biochem Soc Trans* 2003, 31: 1491-1496.

[Tong et al 2001] Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD: Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science* 2001, 294: 2364–2368.

[Tong et al 2004] Tong AHY, et al: Global Mapping of the Yeast Genetic Interaction network. *Science* 2004, 303: 808–813.

[Uetz et al 2000] Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshorn D, Narayan V, Srinivasan M, and Pochart P: A comprehensive analysis of protein-protein interactions of *Saccharomyces cerevisiae*. *Nature* 2000, 403: 623–627.

[Valente and Cusick 2006] Valente AXCN and Cusick M: Yeast protein interactome topology provides framework for coordinated-functionality. *Nucleic Acids Res* 2006, 34: 2812–2819.

[Vidal 2001] Vidal M: A biological atlas of functional maps. *Cell* 2001, 104: 333-339.

[Wagner and Fell 2001]Wagner A and Fell DA: The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B*, 2001, 268: 1803–1810.

[Walhout et al 2000] Walhout A, Sordella R, Lu X, Hartley J, Temple G, Brasch M, Thierry-Mieg N, and Vidal M: Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 2000, 287: 116-122.

[Wang and Zhang 2007]Wang Z and Zhang J: In Search of the Biological Significance of Modular Structures in Protein Networks. *PLoS Comput Biol* 2007, 3: e107. doi:10.1371/journal.pcbi.0030107

[Watts and Strogatz 1998] Watts DJ, Strogatz SH: Collective dynamics of "small world" networks. Nature 1998, 393: 440-42.

[Weng et al 1999] Weng G, Bhalla US, Iyengar R: Complexity in biological signaling systems. *Science* 1999, 284: 92–96.

[White and Smyth 2005] White S and Smyth P: A Spectral Clustering Approach to Finding Communities in Graphs. *SIAM International Conference on Data Mining 2005*, Newport Beach, CA, USA.

[Wilkinson and Huberman 2004] Wilkinson D and Huberman BA: A Method for Finding Communities of Related Genes. *Proc Natl Acad Sci USA* 2004. 101(Suppl 1): 5241-5248.

[Wilmes et al 2008] Wilmes GM, Bergkessel M, Bandyopadhyay S, Shales M, Braberg H, Cagney G, Collins SR, Whitworth GB, Kress TL, Weissman JS, Ideker T, Guthrie C, Krogan NJ: A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. *Mol Cell* 2008, 32: 735-746.

[Wu et al 2004] Wu PY, Ruhlmann C, Winston F, Schultz P: Molecular architecture of the S. cerevisiae SAGA complex. *Mol. Cell* 2004, 15: 199–208.

[Wu and Hu 2005] Wu D and Hu X: An Efficient Approach to Detect a Protein Community from a Seed, in the *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (*CIBCB 2005*), pp 135-141.

[Wu and Hu 2006a] Wu D and Hu X: Mining and Analyzing the Topological Structure of Protein-Protein Interaction Networks, in *the Proceedings of the 2006 ACM Symposium on Applied Computing* (*SAC 2006*), April 23-27, Dijon, Bourgogne, France, pp185-189.

[Wu and Hu 2006b] Wu D and Hu X: Topological Analysis and Sub-Network Mining of Protein-Protein Interactions, in *Advances in Data Warehousing and Mining*, D. Taniar (Ed), Idea Group Publisher, Dec, 2006.

[Wu and Hu 2006c] Wu D and Hu X: Global Analysis of Protein Translation Networks in Yeast. *CSB 2006* (poster). August 14-18, 2006, Stanford University, California.

[Wu and Hu 2007] Wu D and Hu X: Integrative Analysis of Yeast Protein Translation Networks, in *Knowledge Discovery in Bioinformatics*, X Hu and Y Pan (Eds), Wiley & Son, May 2007.

[Wuchty 2004] Wuchty S: Evolution and topology in the yeast protein interaction network. *Genome Res* 2004, 14: 1310-1314.

[Xenarios et al 2000] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: DIP: The Database of Interacting Proteins. *Nucleic Acids Res* 2000, 28: 289-291.

[Yang et al 2008] Yang L, Vondriska TM, Han Z, Maclellan WR, Weiss JN, Qu Z: Deducing topology of protein-protein interaction networks from experimentally measured sub-networks. *BMC Bioinformatics* 2008, 9:301.

[Yook et al 2004] Yook SH, et al: Functional and topological characterization of protein interaction networks. *Proteomics* 2004, 4: 928-42.

VITA

DANIEL DUANQING WU

EDUCATION

Ph.D. Information Science & Technology (Expected Sept 2010), Drexel University, PA

M.S. Computer Science (2001), Physiology (1996), Pennsylvania State University, PA

M.S. Biochemistry (1990), Peking Union Medical College, China

B.S. Biochemistry (1987), Xiamen University, Xiamen, Fujian, China

RESEARCH INTERESTS

Data mining, database management systems, graph theory, and bioinformatics.

SELECTED PUBLICATIONS

1. **D Wu**, X Hu, EK Park, X Wang, J Feng, X Wu (2010): Exploratory analysis of protein translation regulatory networks using hierarchical random graphs. *BMC Bioinformatics* 11(Suppl 3): S2.
2. **D Wu**, X Hu, T He (2009): Exploratory Analysis of Protein Translation Regulatory Networks Using Hierarchical Random Graphs. *BIBM 2009*: 118-123.
3. X Zhou, X Hu, X Zhang, **D Wu**, T He, A Luo (2008): A Mixture Language Model for Class-Attribute Mining from Biomedical Literature Digital Library. *BIBM 2008*: 17-22.
4. X Hu, BA Sokhansanj, **D Wu**, Y Tang (2007): A Novel Approach for Mining and Fuzzy Simulation of Subnetworks From Large Biomolecular Networks. *IEEE T. Fuzzy Systems* 15(6): 1219-1229.
5. X Hu, **D Wu** (2007): Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases. *IEEE/ACM Trans. Comput. Biology Bioinform.* 4(2): 251-263.
6. X Zhang, **D Wu**, X Zhou, X Hu (2006): A Language Modeling Text Mining Approach to the Annotation of Protein Community. *BIBE 2006*: 12-19.
7. X Hu, X Zhang, **D Wu**, X Zhou, P Rumm (2006): Integration of Instance-Based Learning and Text Mining for Identification of Potential Virus/Bacterium as Bio-terrorism Weapons. *ISI 2006*: 548-553.
8. **D Wu**, X Hu (2006). Mining and Analyzing the Topological Structure of Protein-Protein Interaction Networks, *SAC 2006*: 185-189, April 23-27, 2006, Dijon, Bourgogne, France.
9. **D Wu**, X Hu (2006). Topological Analysis and Sub-Network Mining of Protein-Protein Interactions, in *Advances in Data Warehousing and Mining*, David Taniar (Ed), Idea Group Publisher.
10. X Hu, X Zhang, I Yoo, X Zhou, **D Wu** (2006): A Comprehensive Comparison Study of 7 Methods of Mining Hidden Links from Biomedical Literatures, in *Knowledge Discovery in Bioinformatics: Techniques, Methods and Applications*, X. Hu & Y. Pan (Eds). Wiley & Son.
11. X Hu, X Zhang, **D Wu**, Z Zhou, P Rumm (2006): Text Mining the Biomedical Literature for Identification of Potential Virus / Bacterium as Bioterrorism Weapons, in *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*, H. Chen, E. Reid, J. Sinai, A. Silke, B. Ganor (Eds), Springer.
12. **D Wu**, X Hu (2005). An efficient approach to detect a protein community from a seed, *CIBCB 2005*: 135-141, Nov. 14-15, 2005, San Diego, USA.
13. R Weber, **D Wu** (2004): Knowledge Management for Computational Intelligence Systems. *HASE 2004*: 116 – 125, Los Alamitos, CA.
14. **D Wu**, R Weber, F Abramson (2004): A Case-Based Framework for Leveraging NutriGenomics Knowledge and Personal Nutrition Counseling. Case-Based Reasoning in Health Sciences Workshop. 7th European Conference in Case-Based Reasoning. Madrid, Spain. August 30 - September 2, 2004.