Systematic Assessment of Protein Interaction Data using Graph Topology Approaches

Jin Chen
B.C.Sc. (Hons)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2006

Systematic Assessment of Protein Interaction Data using Graph Topology Approaches

by

Jin Chen, B.Eng.

Dissertation

Presented to the Faculty of

the School of Computing of

the National University of Singapore

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

National University of Singapore

October 2006

Systematic Assessment of Protein Interaction Data using Graph Topology Approaches

Approved by
Dissertation Committee:

_____

_____

_____

_____

_____

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I would like to express my deep and sincere gratitude to my supervisor, Associate Professor Wynne Hsu, Ph.D., vice dean of the School of Computing, National University of Singapore. Her wide knowledge and her logical way of thinking have been of great value for me. Her understanding, encouraging and guidance have provided a good basis for the present thesis.

I am deeply grateful to my co-supervisor, Associate Professor Mong Li Lee, Ph.D., assistant dean of the School of Computing, National University of Singapore, for her systematic and constructive instructions, and for her important support throughout this work.

I have furthermore to thank my co-supervisor, Dr. See-Kiong Ng, Ph.D, department manager, Knowledge Discovery Department, Institute for Infocomm Research, whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis.

I wish to express my warm and sincere thanks to Professor Limsoon Wang, Ph.D, National University of Singapore, for his constant encouragement and effective comments, which have had a remarkable influence on my entire research in the field of computational biology.

I warmly thank my colleagues, Tiefei Liu, Xin Xu, Zeyar Aung, Hugo Willy and Hon Nian Chua, for their valuable advice, friendly help, and valuable hints.

Their extensive discussions and interesting explorations related to my work have been very helpful for this study. I wish to extend my warmest thanks to all those who have helped me with my work.

Especially, I would like to give my special thanks to my wife, Juan Lang. It is her patient love that enabled me to complete this work. She was of great help in difficult times. Without her encouragement and understanding, it would have been impossible for me to finish my Ph.D study.

Jin Chen

National University of Singapore
October 2006

Systematic Assessment of Protein Interaction Data using Graph Topology Approaches

Publication No. _____

Jin Chen, PhD
National University of Singapore, 2006

Supervisor: Wynne Hsu, Cosupervisor: Mong Li Lee, See-Kiong Ng

Advances in high-throughput protein interaction detection methods enable biologists to experimentally detect protein interactions at the whole genome level for many organisms. However, current protein interaction detection via high-throughput experimental methods such as *yeast-two-hybrid* are reported to be highly erroneous. At the same time, the false negative rate of the interaction networks have also been estimated to be high.

The purpose of this study was to investigate protein interaction networks from the topological aspect, and to develop a series of effective computational methods to automatically purify these networks, *i.e.*, to identify true protein interactions from the existing protein interaction networks and discover unknown protein interactions, by their topological nature.

This thesis introduced three different approaches. First, it presented a novel measure called *IRAP*, and further *IRAP**, to assess the reliability of protein interaction based on the alternative paths in the protein interaction network. A candidate protein interaction is likely to be reliable if it is involved in a closed loop, in which the alternative path of interactions between the two interacting proteins is strong. The algorithm *AlternativePathFinder* was designed to compute the IRAP value for each interaction in a protein interaction network.

Second, the thesis presented a new model to identify true protein interactions with meso-scale (middle size) network motifs in the protein interaction networks. The algorithm *NeMoFinder* was designed to discover such network motifs efficiently. In the algorithm, frequent trees are discovered firstly. Tree is a simper structure than graph and the number of distinct trees is much smaller than the number of graphs with the same size. By finding frequent trees, graph G is naturally divided into a set of graphs GD, in which each graph is an embedding of a frequent tree. Then, the notion of graph cousin was introduced to reduce the computational time of motif candidate generation and frequency counting in GD.

Third, the thesis exploited the currently available biological information that are associated with network motif vertices to capture not only the topological shapes, but also the biological contexts in which they occurred in the PPI networks for network motif applications. We present a method called *LaMoFinder* to label network motifs with Gene Ontology terms in a PPI network. We also show how the resulting labeled network motifs can be used to predict unknown protein functions.

Validation of IRAP and network motifs as measures for assessing the reliability of protein interactions from conventional high-throughput experiments was performed. For *Saccharomyces cerevisiae*, IRAP/motif models discovered 81.5% reliable protein interactions if the cutoff threshold was set to 0.5. If the threshold was increased to 0.85, all the reliable protein interactions could be captured either by the IRAP model or by the network motif model. Experimental results demonstrated that both of the measures are good for assessing the reliability of protein interactions from conventional high-throughput experiments. Furthermore, the performance of IRAP/motif is clearly better than other topology based evaluation methods, such as IG1 and IG2, for identifying true positive and false negative protein interactions. Protein function prediction experiments showed that the labeled network motifs extracted are biologically meaningful and can achieve better performance (both precision and recall) than existing PPI topology based methods for predicting unknown protein functions.

The results suggest that a significant proportion of true protein-protein interactions could be identified by our IRAP/motif models. These two models could

facilitate the rapid construction of protein interaction networks that will help scientists in understanding the biology of living systems. The results also suggest that exploring remote but topologically similar proteins with labeled network motifs could enable a more precise functional prediction of unknown proteins.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Summary

High-throughput protein-protein interaction networks are reported to be highly erroneous, and a large proportion of protein functions are unknown. The purpose of this study was to investigate the protein interaction networks from the topological aspect, and to develop a series of effective computational methods to automatically purify these networks, and to automatically predict protein functions, by their topological nature.

This thesis introduced three distinct approaches. First, it presented a novel measure called IRAP, and further IRAP*, for assessing the reliability of protein interaction based on the alternative paths in the protein interaction network. Second, the thesis presented a new model to identify true protein interactions with large size network motifs in the protein interaction networks. A scalable algorithm *NeMoFinder* was designed to discover meso-scale network motifs. The protein-protein interaction assessment with the resulted meso-scale network motifs showed better performance than small predefined network motifs. Third, this thesis explored not only the topological shapes of the network motifs, but also the biological context in which they occurred. it was also showed the resulting labeled network motifs can be used to precisely predict unknown protein functions.

CHAPTER 1

Introduction

DNA, RNA and proteins are the molecules that participate in life's many vital biological processes. They are unbranched polymer chains, formed by the string together of monomeric building blocks drawn from a standard repertoire that is the same for all living cells. These molecules often interact with each other frequently, and/or conditionally depend on each other to provide higher level functional features, *e.g.,* functions of a protein are usually provided by its interacting with other proteins and genes. This brings the new term, *interactome*, which refers to all the interactions/relations in the cell. The resulted biological networks, such as signal transduction pathways and protein-protein interaction networks, play important roles in many biological processes.

The research work on interactomics is important and necessary. That is because inappropriate protein expression and interactions due to either genetic or environmental factors usually cause diseases. Misunderstanding of these biological networks will cause serious results, especially in new drug design and new medical therapies.

Recent progress in genetics and computer science has offered various solutions to generate vast amounts of data that simultaneously reports on all networks in the cell. These methods include the technological developments in high-

throughput protein interaction detection methods such as *yeast-two-hybrid* [FS89] and *protein chips* [Z+01], which have enabled biologists to experimentally detect protein interactions at the whole genome level for many organisms [ICO+01, UGC+00, MHMF00, DBTM+01, RSDR+01]. In addition, many effective computational protein interaction prediction methods such as gene-fusions[MP+99] and phylogenetic profiles[PMT+99] have been developed to help biologists to predict protein interactions or to narrow down the list of candidates before doing biological experimentations. All these methods can be used to help to reconstruct the biological networks that operate in cells: the collection of interactions can be modelled as a network, with active elements modelled as vertices and interacting nodes connected by edges.

Now that the Human Genome Project and other genome projects have provided us with a partial view of the parts of networks in the cell, scientists' focus has shifted to how those networks operate to make an organism function. This will in turn be easier for genome-based research to generate more data once we can identify and understand existing biological networks. Nevertheless, interactome is much larger than genome and protonome. Consequently, interactome is much more complex and far from fully developed (see Figure 1.1). Current general understanding of these networks still remains rudimentary, even at a qualitative level. For example, most signal transduction pathways are still modelled as a series of uni-directional arrows connecting a linear chain of components. Such diagrams ignore connections to and from other pathways, non-linear structures, and reactions that restore the pathway to its original state when its input disappears, or allow it to adapt to a prolonged stimulus.

Therefore, it will be an appropriate approach to combine classical graph analysis and data mining methods to study the behavior of the biological networks, in the hope of uncovering general principles of network structures, functions, and evolutions that can be used to construct a broad understanding of how cells work.

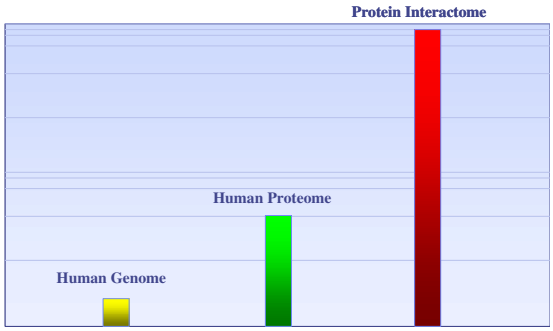

Figure 1.1: Information Complexity.

1.1 Background

The function of a cell is based on complex networks of interacting chemical reactions carefully organized in space and time. The cell can be viewed as an overlay of at least three types of networks, which describes protein-protein, protein-DNA, and protein-metabolite interactions. Interaction networks provide a convenient framework for understanding complex biological systems and the study of their inherent properties has proven extremely useful. However, understanding the structure of these intracellular networks is a complex task, which is complicated by the presence of and interactions between networks of different kinds of elements.

**To make the problem simple, this thesis focuses only on protein-protein interaction (PPI) networks, to interpret the activity of proteins as well as how these proteins interact from the graph topological prospect. It would be easy to append the application to other real networks.

With the development of recent high-throughput techniques, a large amount of PPI data are available. Unfortunately, a significant proportion of the PPIs obtained from these high-throughput biological experiments has been found to contain false positives. Recent surveys have revealed that the reliability of popular high-throughput yeast-two-hybrid assay can be as low as 50% [LWG01, MKS+02, SSM03]. These errors in the experimental protein interaction data will lead to spurious discoveries that can be potentially costly, *e.g.*, wrong drug targets for diseases. It is therefore important to develop systematic methods to detect reliable PPIs from high

throughput experimental data.

Meanwhile, valuable information, such as the function and localization of uncharacterized proteins, and the existence of novel protein complexes and signal-transduction pathways are still not clear to us. People realize that the interaction networks may provide a convenient framework for exploring and understanding the complex biological systems. Even current network analysis is sometimes too abstract to be readily applicable to biology and the networks lack structural details, knowledge could still be learned even from the currently very incomplete networks, for example, unknown protein function predictions based on existing PPI networks.

1.2 Aims

The purpose of this study was to investigate the PPI networks from the topological aspect, and to develop a series of effective computational methods for reconstructing portions of the networks so as to (1) automatically purify interactions for various genomes. *i.e.*, to identify true protein-protein interactions and discover hidden interactions by their topological nature, and (2) predict unknown protein functions based on existing PPIs. To do this art, the three following approaches were taken:

- **Identifying the most promising alternative path for each protein interacting pairs**

  The alternative interaction paths in PPI networks were used as a measure to indicate the functional linkage between two proteins. The existence of strong alternative path is likely to indicate a true-positive interaction. For example, the presence of alternative paths in the PPI networks form circular contigs, and proteins that are found together within a circular contig in *yeast-two-hybrid* screens have been detected for known proteins in macromolecular complexes as well as signal transduction pathways [WSL$^+$00, WBV00]. These closed loops (the alternative path plus the direct linkage) indicate an increased likelihood of biological relevance for the corresponding potential interactions [WSL$^+$00, WBV00, ICO$^+$01].

- **Finding unique and frequent network motifs in a protein-protein interaction network**

  The conserved property of network motifs has been adopted as a measure to validate interaction candidates. Network motifs, such as triad or tetrad, usually represent particular topological patterns which appear only in one kind of networks rather than in any other networks [MSOI+02]. The over-represented property of the network motifs has been confirmed in a wide variety of protein complexes [MSOI+02, SOMMA02]. Network motif can be used as a measure for PPI validation as an interaction appearing frequently in curtain network motifs is knowing to be reliable [SSH02a].

- **Labelling network motifs in protein interactomes for protein function prediction**

  Current network motif finding algorithms models the PPI network as a uni-labeled graph, discovering only unlabeled and thus relatively uninformative network motifs as a result. To exploit the currently available biological information that are associated with the vertices (the proteins), a method called LaMoFinder is presented to label network motifs with Gene Ontology terms in a PPI network. The resulting labeled network motifs are then used to predict unknown protein functions.

  Current protein function prediction methods are based on the functional information of nearby proteins in the network. The missing interactions in an incomplete PPI network usually cause a false prediction. By labeling network motifs, we are able to exploit the currently available biological information that are associated with the vertices (the proteins), and associate remote proteins that are topologically and functionally correlated. The use of labeled network motifs will enable, for the first time, the exploitation of remote but topologically similar proteins for the functional prediction of unknown proteins.

This research may provide a precise and efficient way to automatically verify protein interactions and predict protein functions in the existing protein-protein

interaction networks of many organisms. It could help biologists in identifying true protein interactions and predict unknown protein functions. It also may guide researchers to discover unknown protein links or narrow down the list of candidates before biological experiments. The tools presented in the study could be used to generate highly reliable protein interaction networks, which are helpful for discovering structures and functions of key proteins for new drug design. The set of labelled network motifs generated may be of importance in explaining the functional and physical linages among proteins inside or cross these network motifs.

1.3 Scope

These three approaches only focus on the topological properties of the protein-protein interaction networks. Other properties, such as functional similarity or subcellular co-localization, are mainly used as criteria to validate these three approaches.

The target of this study is to identify "true physical" links. Hence, only the physical interaction networks are adopted in the experiments to validate the three approaches. Functional links, which size are much larger, are not used.

1.4 Organization

The rest of this thesis is organized as follows. First, the topological properties of the protein interaction network and its existing PPI evaluation methods will be reviewed in detail in chapter 2. Chapter 3 introduces a quantitative measure with alternative path approach for the reliability of protein interactions detected in high-throughput genome-wide experiments. Chapter 4 describes a novel method as a computational complement for repurification of the highly erroneous protein interactomes, involving an iterative process of removing false positive interactions and adding interactions detected as false negatives. Chapter 5 presents another strategy by using network motifs to access the reliability of interaction pairs. The network motif strategy can evaluate protein interacting pairs which have no alternative path. Chapter 6 exploits

the currently available biological information that are associated with the proteins to capture not only the topological shapes of the network motifs, but also the biological context in which they occurred in the PPI networks for network motif applications. Finally, we conclude in Chapter 7 with discussions about further work.

Networks have been used to model real-world relationships to better understand them and to guide experiments to predict their behavior. Since incorrect models will lead to incorrect predictions, it is vital to find a good model to fit the protein-protein interaction networks that networks turns to be scale-free. In chapter 2, we will first introduce the PPI network and then explore its topological properties, both on global and local scale, which may reveal design principles of the network.

CHAPTER 2

Literature Review

In this chapter we first introduce existing protein-protein interaction (PPI) networks. Then we review the global and local topological properties of the PPI networks. In the end, we review recent protein interaction evaluation methods and protein function prediction methods. Most of these methods are based on graph topologies.

2.1 Terminology

This section introduces the graph theoretic terminology and biological terminology which will be used in the rest of the thesis.

2.1.1 Graph Theoretic Terminology

Biomolecular interaction data, generally referred to as biological or cellular networks, are frequently abstracted using graph models. Biological networks are abstract representations of biological systems, which capture many of their essential characteristics. In a biological network, molecules are represented by vertices, and their interactions are represented by edges. We present here basic graph theoretic terminology used in this thesis. We also give definitions of basic biological terms used in this thesis. We assume that the definitions of DNA, RNA, protein, genome,

proteome and interactome are commonly known and do not include them here.

A graph is a collection of points and lines connecting a subset of them; the points are called vertices or vertices, and the lines are called edges. A graph is usually denoted by $G = (V, E)$, where V is the set of vertices and $E \subseteq V \times V$ is the set of edges of G. We also use $V(G)$ to represent the set of vertices of a graph G, and $E(G)$ to represent the set of edges of a graph G. A graph is undirected if its edges are undirected, and otherwise it is directed. Vertices joined by an edge are said to be adjacent. A neighbor of a vertex v is a vertex adjacent to v. We denote by $N(v)$ the set of neighbors of vertex v (called the neighborhood of v). The degree of a vertex is the number of edges incident with the vertex. In directed graphs, an in-degree of a vertex is the number of edges ending at the vertex, and the out-degree is the number of edges originating at the vertex. A graph is complete if it has an edge between every pair of vertices. Such a graph is also called a clique. A complete graph on vertices is commonly denoted by K_n. A path in a graph is a sequence of vertices and edges such that a vertex belongs to the edges before and after it and no vertices are repeated; a path with k vertices is commonly denoted by P_k. The path length is the number of edges in the path. The shortest path length between vertices u and v is commonly denoted by $d(u, v)$. The diameter of a graph is the maximum of $d(u, v)$ over all vertices u and v; if a graph is disconnected, we assume that its diameter is equal to the maximum of the diameters of its connected components. A subgraph of G is a graph whose vertices and edges all belong to G. A subgraph with k vertices is said to be a size-k subgraph; a subgraph with n vertices and m edges is represented as g_m^n.

2.1.2 Biological Terminology

Proteins are important components of a cell. They are able of transferring signals, controlling the function of enzymes, regulating production and activities in the cell etc. To do this, they interact with other proteins, DNA, and other molecules. Some of the PPIs are permanent, while others happen only during certain cellular processes. Groups of proteins that together perform a certain cellular task are called protein complexes. There is evidence that protein complexes correspond to complete

or "nearly complete" subgraphs of PPI networks.

A molecular pathway is a chain of cascading molecular reactions involved in cellular processes. Thus, they are naturally directed.

Homology is a relationship between two biological features which have a common ancestor. The two subclasses of homology are orthology and paralogy. Two genes are orthologous if they have evolved from a common ancestor by speciation; they often have the same function, taken over from the precursor gene in the species of origin. Orthologous gene products are believed to be responsible for essential cellular activities. In contrast, paralogous proteins have evolved by gene duplication; they either diverge functionally, or all but one of the versions is lost.

2.2 Protein-protein interaction network

Proteins are the molecules that actually participate in life's many biological processes. They are often described as the "workers" in living cells. Similar to social animals, proteins often interact with each other frequently. Functions of a protein are usually provided by its interacting with other proteins and genes. The interactions results in a large, and consequently complex, interaction network.

PPI networks are commonly represented in a graph format, with vertices corresponding to proteins and edges corresponding to protein-protein interactions. An example of a PPI network constructed in this way is presented in Figure 2.1 [PWJ04]. The network consists of many small subnets (groups of proteins that interact with each other but not interact with any other protein) and one large connected subnet comprising more than half of all interacting proteins. The volume of experimental data on protein-protein interactions is rapidly increasing thanks to high-throughput techniques which are able to produce large batches of PPIs. For example, yeast contains over 5000 proteins, and currently about 18000 PPIs have been identified between the yeast proteins, with hundreds of labs around the world adding to this list constantly [XRS+00]. The analogous networks for mammals are expected to be much larger. For example, humans are expected to have around 12000 proteins and about 10^6 interactions.

The relationships between the structure of a PPI network and a cellular function are just starting to be explored. Many recent research works have been done on interactome, including protein interaction network construction, topological analysis, network purification, functional prediction, *etc.,*

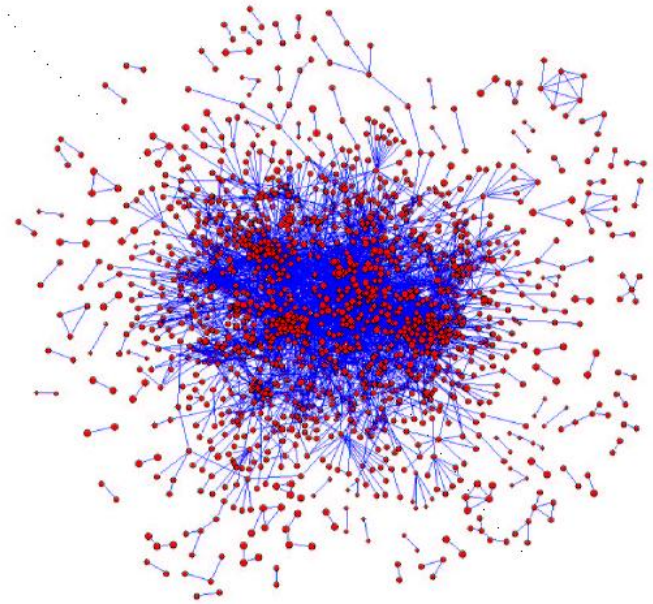

Figure 2.1: The PPI network constructed on 11000 yeast interactions involving 2401 proteins from [PWJ04]. The network consists of many small subnets (groups of proteins that interact with each other but not interact with any other protein) and one large connected subnet comprising more than half of all interacting proteins.

2.2.1 Yeast PPI Network

Yeast, perhaps the best understood eukaryotic organism at the molecular and cellular levels, is a tiny form of fungi or plant-like microorganism that exist in or on all living matter, *i.e.,* water, soil, plants, air, *etc.,*. A common example of a yeast is the bloom we can observe on grapes. There are hundreds of different species of yeast identified in nature, but the genus and species most commonly used for baking is *Saccharomyces cerevisiae*. The scientific name *Saccharomyces cerevisiae* means "a mold which ferments the sugar in cereal (saccharo-mucus cerevisiae) to produce alcohol and carbon dioxide". The ultimate reaction of importance in this process is

the conversion of simple sugars to ethyl alcohol and carbon dioxide.

$$C_6H_{12}O_6 \rightarrow 2\ CH_3CH_2OH + 2\ CO_2$$

A yeast cell is about 0.001 millimeter in diameter, which weighs about 0.008 to 0.010 milligram. Inside each cell are the following: a liquid solution of protoplasm, protein, fat and mineral matter; one or more dark patches called vacuoles; and a darker spot which is the nucleus. Nucleus is where the cell's genetic information is stored which controls all the operations of the cell.

A yeast cell has about 6000 different proteins. Like any living thing, yeast is made up of chromosomes; there are 16 different chromosomes in yeast compared with 23 in humans. In present, about 18000 protein-protein interactions have been discovered and stored in databases. Protein interaction network databases such as DIP [XSD+02], BIND [BDH03] and MIPS [MFG+02] documents these experimentally determined protein-protein interactions. They also present protein interaction from the molecular level to the pathway level for various organisms. The abundant number of protein interactions allows us to analyze organisms at the genome level.

Recent studies on the reliability of high-throughout detection of protein interactions using Y2H have revealed high error rates [EIKO99, MKS+02], some reporting as high as 50% false positive rates [SSM03]. And as pointed out frequently, there is very little overlap of observed interactions among yeast proteins when more than one method is used [MKS+02]. The low coverage and small overlap suggest that high false negatives coexist with high positives exhibited by current experimental detection methods. Accordingly, methods for assessing the reliability of each candidate protein-protein interaction are urgently needed.

2.2.2 PPI networks of other genomes

Protein interactions of other genomes, such as E. Coli, C. elegans, D. melanogaster (fruit fly), Mus musculus (house mouse) and even Homo sapiens (Human), are actively studied as well as S. cerevisiae. For example, C. elegans is an ideal model for studying how protein networks relate to multicellularity [SCN+04]. Table 2.1 lists

the current PPI networks for various genomes generated with high-throughput tools [XRS+00, P+03].

ORGANISM	PROTEINS	INTERACTIONS
D. melanogaster	7052	20988
S. cerevisiae	4920	18228
E. coli	1852	7430
D. melanogaster	7052	20988
C. elegans	2638	4030
H. pylori	710	1425
H. sapiens	18284	33710
M. musculus	202	293
R. norvegicus	87	109

Table 2.1: PPI networks for various genomes. Data collected from DIP [XRS+00] and HPRD [P+03]

In the rest of the chapter, global and local natural network topology properties will be studied. Unlike random distributed networks, natural networks turns to be scale-free. Then, the state-of-art computational protein interaction evaluation methods and protein function prediction methods will be discussed. Computational methods using natural network topology properties are able to show better performance than previous methods.

2.3 Network Topological Properties

Biological networks are usually modeled using various graph theoretic formalisms. Metabolic pathways, for instance, are naturally modeled using directed hyper-graphs, with vertices representing compounds (substrates and products), and hyper-edges representing enzymes (reactions). It is possible to reduce such a model into a general directed graph with vertices representing enzymes, and a directed edge from an enzyme to another implying that the product of the first enzyme is consumed by a reaction catalyzed by the other. Similarly, protein interaction networks are modeled by simple graphs with edges corresponding to an observed interaction between pairs of proteins.

The PPI networks tend to be scale-free from their global topology aspect, *i.e.*, the number of connections per protein is not distributed randomly. Instead, they follow a power-law distribution such that most vertices have only a few connections, and a small number of 'hubs' are highly connected. The deletion of such hubs is often lethal, which is logical because something so centrally connected probably affects many crucial cellular processes. Locally, PPI networks are seen to share structural principles with engineered networks [Alo03]. Three of the most important shared principles are modularity, robustness to component tolerances, and use of recurring circuit elements. It is a complex task to understand and properly use these global and local topologies of the PPI networks to evaluate protein interactions and predict protein functions.

2.3.1 Global Properties

Recent works in network analysis [MSOI+02, GBBK02, YLSea04] have revealed that the topology of complex natural networks such as protein-protein interaction (PPI) networks are far from random. Many of these networks have been shown to exhibit common global topological features such as the "small-world" and "scale-free" properties [WS98, BR99].

Small-world

In 1998, small-world networks were identified as a class of random graphs by Duncan Watts and Steven Strogatz [WS98] by noting that graphs could be classified according to their clustering coefficient and their mean-shortest path length. Small-world networks, as compared to other random graphs with the same number of vertices and edges, are characterized by clustering coefficients significantly higher than expected and mean shortest-path length lower than expected.

Small-world networks mean that it does not take many hops to get from one vertex to another - the science behind the notion that there are only six degrees of separation between any two people in the world. Many empirical graphs are well modeled by small-world networks [Wat03], including social networks, the Internet,

and gene networks. By definition, small-world networks have high representation of cliques and subgraphs that are a few edges shy of being cliques, i.e. small-world networks have sub-networks that are characterized by the presence of connections between almost any two vertices within them. This follows from the requirement of a high cluster coefficient. Secondly, most pairs of vertices will be connected by at least one short path. This follows from the requirement that the mean-shortest path length be small.

It is hypothesized that the prevalence of small-world networks in biological systems may reflect an evolutionary advantage of such an architecture [BR99]. One possibility is that small-world networks are more robust to perturbations than other network architectures. If this were the case, it would provide an advantage to biological systems that are subject to damage by mutation or viral infection.

Scale-free

In 1999, Albert-Laszlo Barabasi and his colleagues at the University of Notre Dame mapped the connectedness of the Web with a web crawler. They were surprised to find that the structure of the web didn't conform to the then-accepted model of random connectivity. Instead, their experiment yielded a network that they christened "scale-free": the ratio of very connected vertices to the number of vertices in the rest of the network remains constant as the network changes in size [BR99].

The follow-up discoveries about networks have been found to have implications well beyond the Internet, including some social and biological networks [BJR$^+$02, FFF99, Wuc01]. The notion of scale-free networks has turned the study of a number of fields upside down. Scale-free networks have been used to explain behaviors as diverse as those of airline traffic routes, power grids, the stock market and cancerous cells, the dispersal of sexually transmitted diseases, as well as the biological network functions and behaviors.

From the topological view, the vertices of a scale-free network aren't randomly or evenly connected. Scale-free networks include many "very connected" vertices, hubs of connectivity that shape the way the network operates, while the rest of vertices have limited number of neighbors. In contrast, random connectivity dis-

tributions predicted that there would be no well-connected vertices, or that there would be so few that they would be statistically insignificant. Although not all vertices in that kind of network would be connected to the same degree, most would have a number of connections hovering around a small, average value. Also, as a randomly distributed network grows, the relative number of very connected vertices decreases.

Mathematically, a scale-free network is defined by the presence of a power-law tail in the degree distribution $P(k)$ (probability distribution of the number of links per vertex over the network), see Equation 2.1. The power-law behavior emerges by a non-zero probability to find vertices with high number of links (hence high number of neighbors). While in random networks, all the vertices are likely to have the same degree $k \sim (k)$, as a consequence the system defines a "scale" $(k \sim (k))$.

$$P(k) \sim k^{-\gamma} \tag{2.1}$$

The ramifications of this difference between the two types of networks (scale-free and randomly distributed) are significant, but it's worth pointing out that both scale-free and randomly distributed networks can be what are called "small-world" networks. So, in both scale-free and randomly distributed networks, with or without very connected vertices, it may not take many hops for a vertex to make a connection with another vertex. There's a good chance, though, that in a scale-free network, many transactions would be funneled through one of the well-connected hub vertices.

Because of these differences, the two types of networks behave differently as they break down. The connectedness of a randomly distributed network decays steadily as vertices fail, slowly breaking into smaller, separate domains that are unable to communicate. Scale-free networks, on the other hand, may show almost no degradation as random vertices fail. With their very connected vertices, which are statistically unlikely to fail under random conditions, connectivity in the network is maintained. It takes quite a lot of random failure before the hubs are wiped out, and only then does the network stop working. In a targeted attack, however, in which failures aren't random but are the result of mischief, or worse, directed at hubs, the

scale-free network fails catastrophically. Take out the very connected vertices, and the whole network stops functioning. For example, viruses have evolved to interfere with the activity of hub proteins such as p53 in a protein-protein interaction network, thereby bringing about the massive changes in cellular behavior which are conducive to viral replication.

2.3.2 Local Topological Properties

Complex biological networks have been classified by global characteristics such as scale-free [BR99, BJR+02, FFF99, Wuc01] and small-world network connection topologies [WS98, Wat03]. In order to investigate networks further beyond their global features requires an understanding of the potential basic structural elements which make up complex networks. Here are 2 of the most important principles of local topology: modularity and network motif.

Modularity

Apart from these global topological characteristics, the complex networks are very different from each other, but they all share the property that their structures are like the result of dynamic non-Markovian processes of individual decisions [Alo03].

A closer observation found that these networks share striking local properties: the presence of many small dense subnetworks/clusters, namely, modules. For example, proteins are known to work in slightly overlapping, co-regulated groups such as pathways and complexes. An understanding of this principle will enable us to model and search these networks effectively.

Modules, usually called clusters, in PPI networks of different size have been found using the Highly Connected Subgraphs (HCS) algorithm [HS00] for cluster analysis. By definition, a module is a set of vertices that have strong interactions and a common function. A module has boundary vertices that control the input/output interactions with the rest of the network. A module also has internal vertices that do not significantly interact with vertices outside the module. Modules may have special features that make them easily embedded in almost any system.

For example, output vertices usually have "low impedance" [Alo03], so that adding on additional downstream clients should not drain the output to existing clients. Modules convey an advantage in situations where the environments change from time to time. Therefore, modular biological networks may have an advantage over non-modular networks in real-life ecologies, which change over time, *i.e.*, modular networks can be readily reconfigured to adapt to new conditions.

The modules in complex networks also make the networks robust to perturbation. This makes sense in biology, because biological networks must work under all plausible interferences that come with the inherent properties of the components and the environment. Thus, for example, E. coli needs to be robust with respect to temperature changes over a few tens of degrees, and no circuit in the cell should depend on having precisely copies of a curtain protein.

Recent analysis on experimentally derived PPI networks observed that with increasing size of the PPI network, the number of vertices in individual modules increases, while the number of identified modules decreases [PWJ04]. This result may be due to increasing noise in the data, or to an aggregation of transient complexes in the overall network.

Network Motif

It turns out that many local topological patterns can be detected in the large complex natural networks. For example, Milo *et al.*[MSOI$^+$02] discovered various significant patterns of local connections occurring more frequently in complex networks than in random networks. They called these recurring local topological substructures as "network motifs".

While relatively less widely studied than the global topological features, such network motifs can lead to better understanding about various classes of complex networks, as some network motifs may be particular to specific classes of networks, such as filtering out spurious input fluctuation, generating temporal programs of expression or accelerating the throughput of the network. Whereas, a curtain part of network motifs are discovered to be conserved in one class of networks. For example, curtain triad and tetrad motifs are found to appear commonly in gene

transcription networks of *S. cerevisiae* and *E. coli* but rather than in any other kinds of networks [MSOI$^+$02]. In addition, the presence of such network motifs also indicates the basic structural elements that underlie the hierarchical and modular architecture of such complex natural networks as PPI networks.

It is important to stress that the similarity in network motif topology does not necessarily stem from duplication. Evolution, by constant tinkering, appears to converge on these network motifs in different non-homologous systems, presumably because they are optimally suited to carry out key functions [Wag03].

Network motifs can be detected by algorithms that compare the patterns found in the target network to those found in suitably randomized networks. Once a dictionary of network motifs and their functions is established, one could envision researchers detecting network motifs in new networks just as protein domains are currently detected in the sequences of new genes. Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it.

2.4 Protein Interaction Evaluation Methods

With the development of recent screening techniques, a large amount of protein-protein interaction data are available, from which biologically important information such as the function and localization of uncharacterized proteins and the existence of novel protein complexes and signal-transduction pathways can be recognized. However, existing data on protein interactions contain many false positives, which may lead to spurious discoveries that can be potentially costly, *e.g.,* wrong drug targets for diseases. Consequently, computational methods of assessing the reliability of each candidate protein-protein interaction are urgently needed.

The evaluation methods based on the topological properties of the protein-protein interaction networks could be divided into three types: experimental results combination, interaction generalities and network motifs.

2.4.1 Experimental Results Combination

The initial approach proposed by Mering *et al.* is to consider combining the results from multiple independent detection methods[MKS+02]. The multi-occurring interactions are thought to be highly reliable because of its reproducible property.

In the abstract, it is easy to demonstrate that combining independent data sets results in a lower error rate overall. For instance, combining three independent binary-type data sets with error rates of 10% reduces the overall error rate to 2.8% (for both false positives and negatives) [HN8] (7). Moreover, interrelating two different types of whole-genome data also enables one to discover potentially important but not obvious relationships–for example, between gene expression and the position of genes on chromosomes, or between gene expression and the subcellular localization of proteins (8, 9). (Enhanced: Integrating Interactomes. Science) However, this is a limited approach because of the low overlap between the different detection methods[HF01, MKS+02]. In Mering's analysis, out of the 80,000 available interactions between yeast proteins from the different high-throughput methods, only a surprisingly small number (2,400) is supported by more than one method[MKS+02]. That is mainly because the interactions generated from these methods do not reach saturation, and also because a significant fraction of protein interactions detected are false positives. Therefore, co-existing interactions in more than one experiment are usually treated as a good validation instead of a stand alone evaluation method.

2.4.2 Logistic Regression Model

Bader *et al* [BCC04] developed a quantitative method recently to compute confidence values for protein interacting pairs with a logistic regression approach, in which statistical and topological descriptors are used to predict the biological relevance of protein-protein interactions. The training set is generated by comparing networks from two major biological protein interaction detection methods, *yeast-two-hybrid*[FS89] and *co immunoprecipitation*[G+02]. Pairs of proteins close together in both networks were selected as positive examples, and proteins connected in one network and far apart in the second network were selected as negative examples.

After that, a logistic regression model is built in the training set to shift the dividing surface between low and high confidence. Explanatory variables are based on the data source, the topological properties of the interaction partners, *etc.,*. The model is then used to predict confidence scores for pair-wise interactions in the full data set.

Although the high-confidence interactions in Bader's experiments show high agreement with similar database annotations, it is abnormal that the *co immunoprecipitation* interactions have a **negative** correlation with mRNA co-expression, while the *yeast-two-hybrid* interactions have a **positive** correlation with mRNA co-expression in their experiments [BCC04]. The contravention could be explained by the fact that both of the biological methods have specific strengths and weaknesses [MKS+02]. For example, interactions detected by the *yeast-two-hybrid* technology largely fail to cover certain categories, such as proteins involved in translation.

2.4.3 Interaction Generalities

Besides the various works on the results from different biological experiments, another approach is to model the expected topological characteristics of true protein interaction networks, and then devise mathematical measures to assess the reliability of the candidate interactions. Saito *et al.* developed a series of computational measures called *interaction generalities* (IG) [SSH02b, SSH02a] to assess the reliability of protein-protein interactions.

Interaction Generality 1 (IG1). The IG1 measure was based on the idea that interacting proteins that appear to have many interacting partners that have no further interactions were likely to be false positives. IG1 was defined as the number of proteins that directly interact with the target protein pair but do not interact with any other proteins. The higher the IG1 value for an interaction, the more likely it was a false positive.

This is a reasonable model for *yeast-two-hybrid* data, as some 'sticky' proteins in *yeast-two-hybrid* assays do have a tendency to turn on the positive signals of the assay by themselves. In yeast two-hybrid assays, candidate proteins carry different parts of the biological mechanism necessary for the transcription of a re-

porter gene; the interaction of two proteins brings about the complete assembly for the transcription of the reporter gene, turning on a positive signal that can be detected for the interaction. A sticky protein, however, can activate transcription of the reporter gene without actually interacting with their partners, which leads to an excess number of candidate partners for the protein. These proteins would be observed to interact with a large number of random proteins in the experimental data. They often have *in silico* high IG1 values.

Interaction Generality 2 (IG2). IG1 is a local measure which does not consider the topological properties of the protein interaction network beyond the candidate protein pair. As such, it has limited coverage for the different types of experimental data errors. Saito *et al.* developed the IG2 measure [SSH02a] to incorporate topological properties of interactions beyond the candidate interacting pairs by considering the five possible topological relationships of a third protein C with a candidate interacting pair (A, B). IG2 was the weighted sum of the five topological components with respect to C. The weights were assigned *a priori* by performing a principal component analysis on the entire interaction network. Experimental results demonstrated that IG2 performed better than IG1 [SSH02a].

But IG2 remains a fixed local measure since the topological context that it considers involved only five topological components of a neighbor C. As such, both the IG1 and IG2 measures do not consider the underlying system-wide topological structure of the entire interaction network to determine the reliability of the discovered protein interactions.

2.4.4 Network Motifs

Either IG1 or IG2 uses the alternative links which connect the target protein pairs separately. A more advanced method is to group these links into network motifs and use these network motifs to access the reliability of the interaction pairs.

To understand the complex wiring diagrams of real networks–including the protein interaction network–with unknown design principles, researchers thought of breaking down such networks to identify the simplest units of commonly used network architecture. In 2002, Milo introduced the concept of "network motif"

[MSOI+02] as small patterns of interconnections that occur in the network at numbers that are significantly higher than those in randomized networks.

Because different types of networks are constructed by different network motifs, network motifs can be used to uncover the structural design principles of complex networks. For example, the motifs in ecological food webs were distinct from the motifs shared by the genetic networks of *Escherichia coli* and *Saccharomyces cerevisiae* [MSOI+02]. These network motifs provide specific regulatory capacities based on network topology. The star network motif was discovered frequently in computer networks, because a central hub represented a switch or router which connects a number of computers in the network[MFCG03]. The eukaryotic network motifs were found in regulator gene interaction networks. Their topological structures can be explained by the functional modules composed by these motifs[LRR+02]. Therefore, these significative motifs can be assembled into network structures that help the researchers evaluate the reliability of protein interaction pairs. An interaction that appears frequently in multiple network motifs is usually thought to be reliable.

However, existing algorithms for detecting network motifs are not scalable enough to find large network motifs. These algorithms mainly act by exhaustively enumerating all subgraphs with a given number of vertices in the network. The runtime of such algorithms increases significantly with network size. The subproblem, maximal independent set, is NP-hard and even has no heuristic algorithm which could be accomplished in polynomial time. Moreover, the number of possible network motifs increases exponentially with the motif size and the subgraph isomorphism problem, an essential technique to identify different network motifs, has already been proved to be at least NP-complete. It greatly limits the number of network motifs researchers could scale to. Hence, to find meaningful network motifs, which are sufficiently large, is difficult and very expensive.

2.4.5 Methods for Performance Study

All the evaluation methods based on topological properties of the protein-protein interaction network will give a weight to each interaction. Normally the higher the weight, the more reliable the interaction. Consequently, a series of experiments were

carried out to evaluate the effectiveness of using these weights to detect reliable protein-protein interactions.

1. *Experimentally reproducible interactions.* Protein interactions that have been detected by multiple independent experiments are able to be used as the desired "gold standards". The proportion of reproducible interactions should increase in filtered protein interaction data with the increasing of the weight;

2. *Annotated functional associations.* By the 'guilt-by-association' principle [Oli00], true interacting proteins should share at least a common functional role. The proportion of interacting proteins with a common functional role should increase in filtered interaction data with the increasing of the weight;

3. *Gene expression correlations.* Genes that are co-expressed indicate that their gene products (the proteins) partake in the same pathway—the corresponding proteins are thus highly likely to be interacting. Here, we need to check whether the filtered interactions can be confirmed by co-expression at the mRNA level;

4. *Cellular localization cross-talks.* For two proteins to be interacting *in vivo*, they should at least be at a common cellular localization. Hence, the rate of cellular localization cross-talk should be decreased in filtered interactions with the increasing of the weight, indicating a reduced degree of biologically irrelevant interactions in the rest data.

5. *Biologically interacting cross-talkers.* Biologically genuine cross-talkers, such as the proteins involved in signal transduction pathways, share same functions but are not co-localized. Therefore, there should be a proportion of the cross-talking interactions with high weights having functional matches.

6. *Many-Few interaction trend in protein networks.* Maslov *et al.*[MS02] found that there is a "many-few" interaction pattern in protein interaction networks. As the weight thresholds increased, the proportion of "many-few" interactions also should increase in the filtered interaction data.

Another way to do performance study is to use the Gene Ontology (GO) [1]. The Gene Ontology (GO) is one of the most important otologies within the bioinformatics community. 3554 out of the 4141 Saccharomyces cerevisiae proteins, 2175 out of the 2911 Caenorhabditis elegans proteins and 6132 out of the 7621 Drosophila melanogaster proteins were annotated under GO for our evaluation experiments.

To compute the degree of biological homogeneity for any two proteins according to their GO molecular function annotations, biological processing and subcellular localizations, an enriched GO similarity measure is introduced by Lord et al [LSBG03], which gives different weights to GO terms based on the GO term frequency in the target species, and the similarity values of different GO terms are determined by their shared parents, as follows:

$$Similarity(t_a, t_b) = \frac{2 \times \ln p(t_{ab})}{\ln p(t_a) + \ln p(t_b)} \tag{2.2}$$

Here t_a and t_b are any two GO terms, and t_{ab} is their common parent. The probabilities $p(t_a)$, $p(t_b)$, and $p(t_{ab})$ refer to the probabilities of the respective GO terms occurring in the target species based on their term frequencies.

[1] Available at http://www.geneontology.org/

CHAPTER 3

IRAP: Interaction Reliability by Alternative Path

Current protein interaction detection via high-throughput experimental methods such as yeast-two-hybrid has been reported to be highly erroneous, leading to potentially costly spurious discoveries. This work introduces a novel measure called IRAP, *i.e.*, "Interaction Reliability by Alternative Path", for assessing the reliability of protein interactions based on the underlying topology of the protein-protein interaction(PPI) network.

A candidate protein interaction is considered to be reliable if it is involved in a closed loop in which the alternative path of interactions between the two interacting proteins is strong. We devise an algorithm called *AlternativePathFinder* to compute the IRAP value for each interaction in a complex PPI network. Validation of the IRAP as a measure for assessing the reliability of PPIs is performed with extensive experiments on yeast PPI data. Results show consistently that IRAP measure is an effective way for discovering reliable PPIs in large datasets of error-prone experimentally-derived PPIs. Results also indicate that IRAP is better than IG2, and markedly better than the more simplistic IG1 measure.

Experimental results demonstrate that a global, system-wide approach—such

as our IRAP measure that considers the entire interaction network instead of merely local neighbors—is a much more promising approach for assessing the reliability of PPIs.

3.1 Introduction

Technological developments in high-throughput protein-protein interaction (PPI) detection methods such as yeast-two-hybrid [FS89] and protein chips [Z+01] have enabled biologists to experimentally detect protein interactions at the whole genome level for many organisms [ICO+01, UGC+00, MHMF00, DBTM+01, RSDR+01]. Unfortunately, a significant proportion of the PPIs obtained from these high throughput biological experiments has been found to contain false positives. Recent surveys have revealed that the reliability of popular high-throughput yeast-two-hybrid assay can be as low as 50% [LWG01, MKS+02, SSM03]. These errors in the experimental protein interaction data will lead to spurious discoveries that can be potentially costly, e.g., wrong drug targets for diseases. It is therefore important to develop systematic methods to detect reliable PPIs from high throughput experimental data.

Biological studies have shown that the interaction clusters obtained from contiguous connections that form closed loops in PPI networks indicate an increased likelihood of biological relevance for the corresponding potential interactions [WSL+00, WBV00, ICO+01]. Proteins that are found together within a circular contig in yeast-two-hybrid screens have been detected for known proteins in macromolecular complexes as well as signal transduction pathways [WSL+00, WBV00]. For example, the configuration A/B/C/D/A indicates that protein A binds to B, B binds to C, C binds to D and D binds to A. We observe that such circular contigs are formed by the presence of alternative paths in the interaction networks. This has led to the use of alternative interaction paths in PPI networks as a measure to indicate the functional linkage between two proteins [ICO+01].

In this chapter, we propose to use the length and strength of the alternative paths between pairs of interacting proteins as a basis for detecting reliable PPIs from high-throughput experimental data. We introduce a quantitative measure called

"Interaction Reliability by Alternative Path" (IRAP) [CHLN04, CHLN05b] for assessing the reliability of a detected PPI with respect to the presence of alternative reliable interaction paths in the underlying topology of the experimentally derived interaction network. We devise an *AlternativePathFinder* algorithm to compute the IRAP values of the interactions in large complex PPI networks. Using the yeast PPI data with annotated functional information as well as other experimental data, we show positive experimental results that validate IRAP as a good system-wide measure for discovering reliable PPIs in error-prone high-throughput experimental data.

The rest of this chapter is organized as follows. Section 3.2 gives the related work and the motivation for this work. Section 3.3 introduces IRAP as a quantitative measure for the reliability of PPIs detected in high-throughput genome-wide experiments. In Section 3.5, we describe the *AlternativePathFinder* algorithm for computing IRAP values in complex PPI networks. Section 3.7 presents the various comparative results of using the computed IRAP values for discovering reliable PPIs for yeast. Finally, we conclude in Section 3.8 with discussions about further work.

3.2 Background

The reported high false positive rates associated with high-throughput experimental PPI data [MKS+02, SSM03] have led researchers to develop methods to assess the reliability of PPIs generated by large-scale biological experiments.

One approach is to combine the results from multiple independent detection methods to derive highly reliable data [MKS+02]. However, this approach has limited applicability because of the low overlap [HF01, MKS+02] between the different detection methods.

Another approach is to model the expected topological characteristics of true PPI networks, and then devise mathematical measures to assess the reliability of the candidate interactions. Saito *et al.* develop a series of computational measures called *interaction generalities* (IG) [SSH02b, SSH02a] to assess the reliability of PPIs.

Interaction Generality 1 (IG1). The IG1 measure is based on the idea

that interacting proteins that appear to have many interacting partners that have no further interactions are likely to be false positives. IG1 is defined as the number of proteins that directly interact with the target protein pair, subtracted by the number of proteins interacting with more than one protein. The higher the IG1 value for an interaction, the more likely it is a false positive.

This is a reasonable model for yeast two-hybrid data, as some 'sticky' proteins in yeast two-hybrid assays do have a tendency to turn on the positive signals of the assay by themselves. In yeast two-hybrid assays, candidate proteins carry different parts of the biological mechanism necessary for the transcription of a reporter gene; the interaction of two proteins brings about the complete assembly for the transcription of the reporter gene, turning on a positive signal that can be detected for the interaction. A sticky protein, however, can activate transcription of the reporter gene without actually interacting with their partners, which leads to an excess number of candidate partners for the protein. These proteins will be observed to interact with a large number of random proteins in the experimental data. They can be detected *in silico* with high IG1 values.

Interaction Generality 2 (IG2). IG1 is a local measure which does not consider the topological properties of the PPI network beyond the candidate protein pair. As such, it has limited coverage for the different types of experimental data errors. Saito *et al.* develop the IG2 measure [SSH02a] to incorporate topological properties of interactions beyond the candidate interacting pairs. By considering the five possible topological relationships of a third protein C with a candidate interacting pair (A, B), IG2 is the weighted sum of the five topological components with respect to C. The weights are assigned *a priori* by performing a principal component analysis on the entire interaction network. Experimental results demonstrate that IG2 performs better than IG1.

We observe that IG2 remains a local measure since the topological context that it considers involved only five topological components of a neighbor C. As such, both the IG1 and IG2 measures do not consider the underlying system-wide topological structure of the entire interaction network to determine the reliability of the discovered PPIs. In contrast, the proposed alternative path approach aims to

provide a comprehensive interaction reliability measure that does not impose any restriction on the number of intervening proteins.

Evolution studies in the conservation of PPI networks [PB01] have suggested association of PPIs with alternative paths, as the global interaction networks evolve by augmenting existing interactions with new interactions in order to yield PPI networks that are more efficient and robust to changes. Therefore, we introduce a quantifiable measure called IRAP to evaluate the reliability of a detected PPI with respect to the presence of a reliable alternative interaction path between the two proteins in the global interaction network. IRAP takes into consideration both the *strength* and the *length* of the alternative paths connecting the two proteins. Extensive experimental results on yeast experimental data (see Section 3.7) will show that IRAP is able to detect the reliable PPIs from error-prone high-throughput experimental interactions better than existing assessment measures.

3.3 IRAP: Interaction Reliability by Alternative Path

In this section, we define the proposed interaction reliability measure—*Interaction Reliability by Alternative Path (IRAP)*—that assigns a reliability value to each candidate interacting protein pair in genome-wide interaction data. IRAP takes into consideration both the *strength* and the *length* of the alternative paths connecting the two proteins. The reliability of a candidate PPI is indicated by the collective reliability of the strongest alternative path of interactions connecting the two proteins in the underlying interaction network. A reliable PPI is accompanied by at least one reliable alternative interaction path in the underlying interaction network.

3.3.1 Network Construction

An experimentally detected PPI network can be modelled using an undirected network $G = (V, E)$. Each node in the network represents a unique protein. An edge exists between two nodes v_A and v_B if there is an interaction between the corre-

sponding proteins A and B. The weight for this edge is initialized as the normalized value of reversed IG1 [SSH02b]:

$$weight(v_A, v_B) = 1 - \left(\frac{IG1^G(A,B)}{IG1^G_{max}} \right) \tag{3.1}$$

$$IG1^G(A,B) = 1 + |\{(A',B') \in E | A' \in \{A,B\} \& deg^G(B') = 1\}| \tag{3.2}$$

As defined by Saito *et al.*, $IG1^G(A,B)$ is the number of proteins that directly interact with the candidate protein pair, subtracted by the number of proteins interacting with more than one protein [SSH02b], while $IG1^G_{max}$ is the maximum IG1 value in the interaction network G.

We use reversed and normalized IG1 as the initial edge weights to reflect the local reliability of each interaction in the PPI network. Since IG1 is an reverse index (*i.e.* the lower the better), we first reverse it to make it more natural (*i.e.* the higher the better). Then, we normalize the the reversed IG1 values to fall between 0 and 1 so that it can be treated as a proper weight in our algorithm. The distribution of the modified weights remains the same as that of IG1.

The task is to find the strongest alternative path that connects a candidate pair of interacting proteins A and B. We initialize the weight value for node v_A to 1 and the rest of the nodes in the network G to 0. To compute $IRAP(A,B)$, we calculate the weight product through a path from v_A to v_B in the network that excludes the direct connection between the two nodes.

3.3.2 Path Selection

True PPI networks are known to be real world networks that have short average distances between vertices [GR03]. This suggests that we should use path *length* as a path selection criterion. However, (short) path length should not be used as the sole selection criterion in PPI networks constructed from high-throughput experiments—we should also take into consideration inherent but path length-independent experimental errors such as the presence of sticky proteins which are measured by such local topological values as IG1. In other words, we should consider both the path

lengths and *strengths* when selecting a path in an interaction network constructed from high throughout experimental data. If we choose the shortest path regardless of the local strengths of the connections (in terms of IG1, say), it is likely that we may select a path with spurious connections involving sticky proteins. On the other hand, if we choose the strongest path regardless of the path lengths, we could end up with a lengthy path which is highly likely to be formed by some spurious link(s). For example, a path consisting of 30 locally strong interactions of weight 0.9 each is less reliable compared to a path with a single but weaker interaction of weight 0.1. This is because of the highly erroneous nature of such PPI networks constructed from high-throughput experiments that have been shown to contain approximately 50% false positives [LWG01, MKS$^+$02, SSM03].

Our IRAP algorithm takes into consideration both the path length and path weight with the following path selection strategy. Whenever there is sharing of nodes, we use the shortest path to approximate the (biologically) strongest alternative path that connects the candidate interacting pair of proteins A and B in the interaction network. This is done in IRAP by considering only non-reducible paths (Definition 1) as candidate alternative paths. Then, given all the candidate non-reducible paths connecting nodes v_A and v_B that do not have any common nodes with each other, we select the (experimentally) strongest path that has the largest weight product as indicated by local experimental weights.

Definition 1. Non-reducible Path. *A path $\phi = v_1, \ldots, v_n$ is a non-reducible path of edge (v_A, v_B) if we have $v_1 = v_A, v_n = v_B$ (or vice versa); and there is no shorter path ϕ' connecting node v_A and v_B that shares some common intermediate nodes with the path ϕ. That is, $\nexists$ path $\phi' = u_1, \ldots, u_m$ such that $(u_i, u_{i+1}) \in E, u_1 v_A, u_m = v_B, u_r = v_s$ for some $r \in [2..m-1]$, $s \in [2..n-1]$, $m < n$.*

Figure 3.1 shows 3 alternative paths between the nodes A and B. Two of the paths <A-D-E-B> and <A-F-G-D-E-B> have nodes D and E in common. The shorter path is selected as a non-reducible path.

Formally, we define IRAP as follows:

Definition 2. IRAP. *The reliability of a candidate PPI (A, B), $IRAP(A, B)$, is*

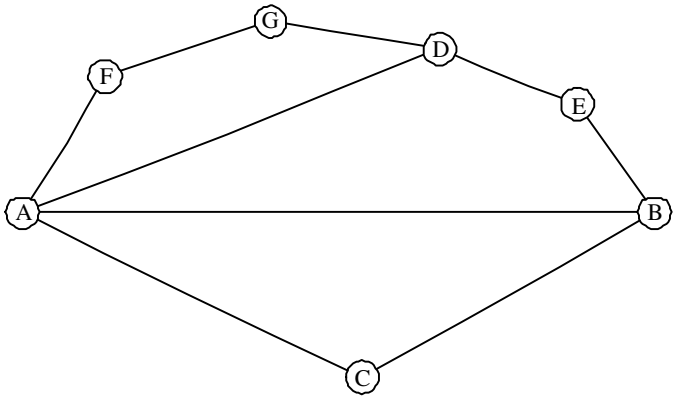

Figure 3.1: An example of alternate paths.

indicated by the collective reliability of the strongest alternative path of interactions connecting the two proteins in the underlying interaction network.

$$IRAP(A,B) = \max_{\phi \in \Phi(A,B)} \prod_{(u,v)\in\phi} weight(u,v) \tag{3.3}$$

where $weight(u,v)$ *denotes the weight value for edge* (u,v) *in the PPI network* G*;* $\Phi(A,B)$ *denotes the set of non-reducible paths.*

IRAP uses IG1-derived values as the local edge weights to identify interactions that are more likely to be "experimentally-correct". At the same time, by considering only non-reducible paths as candidate alternative paths and by globally taking the products of the normalized individual local weights as the path weights for these candidate paths, IRAP also favors for shorter paths[1] that are more likely to be "biologically-correct". The empirical results in Section 5 will show that such combined strategy in IRAP's path selection is indeed robust and effective for identifying reliable interactions in networks constructed from interaction data that contain a high percentage of false positives.

[1]As the local edge weights are normalized between 0 and 1, their product tends to become smaller as more weights are multiplied together, resulting in a tendency for IRAP to favor shorter paths.

3.4 Statistics of Alternative Paths in PPI networks

First, we analyzed protein-protein interaction (PPI) datasets from three different species (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) to investigate the extent to which alternative paths are present in PPI datasets. We focus here only on interactomes that are derived by the popular high-throughput assays such as Y2H. Then, we provide some actual examples in which the presence or absence of alternative paths can be used to increase or decrease the confidence of protein-protein interactions.

3.4.1 PPI Statistics

The *Saccharomyces cerevisiae* PPI dataset has a total of 7,903 interactions and 4,141 proteins. After removing redundant and self- links, the dataset has 7,686 interactions between the 4,141 proteins. 5,802 (75.5%) of these interactions have at least one alternative path. The average length of the alternative path detected by IRAP is 4.98. Note that the alternative path determined by IRAP is not necessarily the shortest path according to its definition (see [CHLN04], [CHLN05b], and the main manuscript for the technical details).

The *Drosophila melanogaster* PPI dataset is a much larger one — it has 24,477 interactions between 7,621 proteins. After removing redundancy and self- links, the dataset is left with 22,437 interactions between the 7,621 proteins. 19,732 (87.9%) of these interactions have at least one alternative path with an average length of 4.64.

The *Caenorhabditis elegans* PPI dataset has 5,123 interactions between 2,911 proteins. After removing redundancy and self-links, the dataset has 5,025 interactions between the 2,911 proteins. 3,312 (65.9%) of these interactions have at least one alternative path with an average length of 3.93.

Table 3.1 summarizes these PPI statistics of the three experimental data sets.

Table 3.1: PPI statistics of the various interactomes.

Species	Interactome size	PPI's w/ alternative paths	Avg path len
Saccharomyces cerevisiae	7,686 interactions btw 4,141 proteins	5,802 (75.5%)	4.98
Drosophila melanogaster	22,437 interactions btw 7,621 proteins	19,732 (87.9%)	4.64
Caenorhabditis elegans	5,025 interactions btw 2,911 proteins	3,312 (65.9%)	3.93

3.4.2 Example Alternative Paths

In our works[CHLN04, CHLN05b], we noted biological studies have showed that the interaction clusters obtained from contiguous connections forming closed loops in PPI networks have indicated an increased likelihood of biological relevance for the corresponding potential interactions [WSL$^+$00, WBV00, ICO$^+$01]. Proteins that are found together within a circular contig in yeast-two-hybrid screens have been detected for known proteins in macromolecular complexes as well as signal transduction pathways [WSL$^+$00, WBV00]. These observations have led to the use of alternative interaction paths in protein interaction networks as a measure to indicate the functional linkage between two proteins [ICO$^+$01].

We illustrate here several actual examples from our yeast PPI dataset in which the presence or absence of alternative paths can be used to increase or decrease the confidence of protein-protein interactions.

Example 1: Absence of or weak alternative path indicating a false positive PPI.

An interaction has been detected between the protein pair ⟨Snf4, Yjl114w⟩ (BIND ID 6321323 and 6322348) with Y2H assays. However, the degree of functional homogeneity between the pair of proteins, as measure by enriched GO term similarity [LSBG03], is as low as 0.062224. This biological observation indicates that the detected interaction between ⟨Snf4, Yjl114w⟩ is highly likely to be a false positive.

We verify whether we can come to the same conclusion using only network topological measures. The reversed IG1 value for ⟨Snf4, Yjl114w⟩ is a high 0.977012, which supports the possibly wrong suggestion that this is a true interaction. In contrast, the IRAP value is a low 0.02108, which means that even the strongest

alternative path between the two proteins ⟨Snf4, Yjl114w⟩ (in this case, "Snf4-Yjr083c-Hsp82-Yjl114w") has been deemed unreliable in our IRAP model. This concurs well with the biological observation of a low degree of functional homogeneity between the proteins.

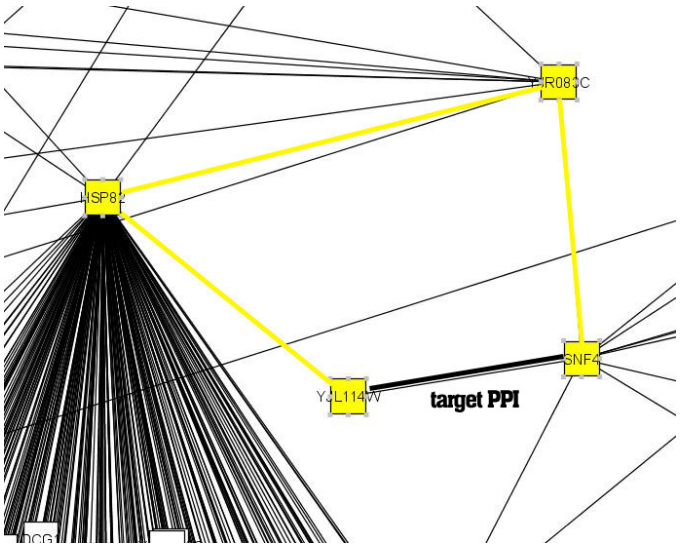

Figure 3.2: Example: absence of or weak alternative path indicating a false positive PPI. $GOSimilarity(Snf4, Yjl114w) = 0.062224$. $IG1(Snf4, Yjl114w) = 0.977012$. $IRAP(Snf4, Yjl114w) = 0.02108$. $Path = Snf4 - Yjr083c - Hsp82 - Yjl114w$

Example 2: Strong alternative path indicating a true positive PPI.

In the previous example, a low IRAP value indicates a false positive PPI. IRAP can thus be used to detect and eliminate possible false positives in an interactome. Meanwhile, a strong alternative path can be used to identify true positives. We give two examples below: the first example illustrates that IRAP can detect the same true positive as IG1, while the second example shows a PPI that was missed by IG1 but was detected with our IRAP model.

- **IRAP is high, IG1 is high**

  The protein pair ⟨Ste5, Fus3⟩ (BIND ID 6320308 and 6319455, labeled as 5 and 32 in the following figure) has a high degree of functional homogeneity - in fact, its enriched GO term similarity is 1.000000. This interaction has a

high probability to be true, because the two proteins have the same functions (MAP-kinase scaffold activity, signal transduction during conjugation with cellular fusion, etc.,) and are located at the same place (e.g. cytoplasm) in the cell.

We verify whether using only network topological information can also help us identify this interaction. Indeed, both its IRAP and reversed IG1 values are 1.0000. The alternative path selected by IRAP was "Ste5-Ste11-Fus3". In this case, both IRAP and IG1 correctly identified the true positive PPI.

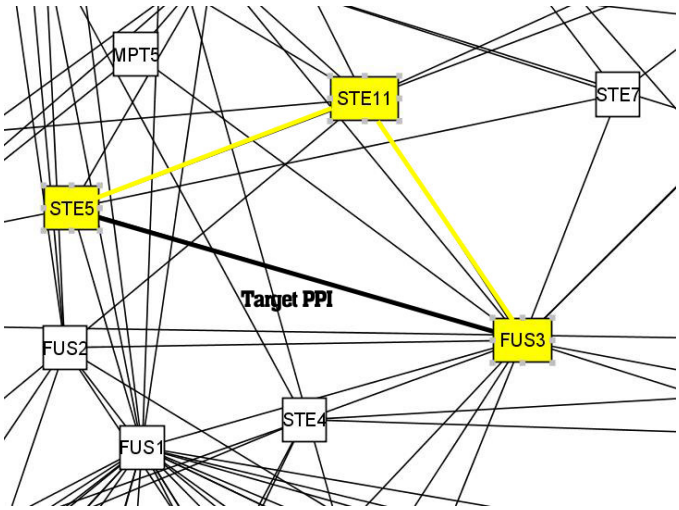

Figure 3.3: Example: a strong alternative path indicating a strong positive PPI. GO Similarity(Ste5, Fus3)=1.0000, Function=MAP-kinase scaffold activity. IG1(Ste5, Fus3)=1.0000. IRAP(Ste5, Fus3)=1.0000. Path=Ste5-Ste11-Fus3

- **IRAP is high, IG1 is low**

  Another protein pair of interest is ⟨Spc34, Jsn1⟩ (BIND ID 6322890 and 6322550). This interaction was supported by the high degree of functional homogeneity between the two proteins, which have a relatively high enriched GO term similarity of 0.886994.

  For this interaction, IG1 failed to detect it. The reversed IG1 value is a low 0.103448. On the other hand, it has a relatively high IRAP value of 0.504180, with an alternative path of "Spc34-Spc19-Ykr083c-Ask1-Vps20-Taf40-Jsn1".

Note that in this case, the corresponding alternative path detected by IRAP is fairly long, illustrating that the shortest path need not be the strongest one.

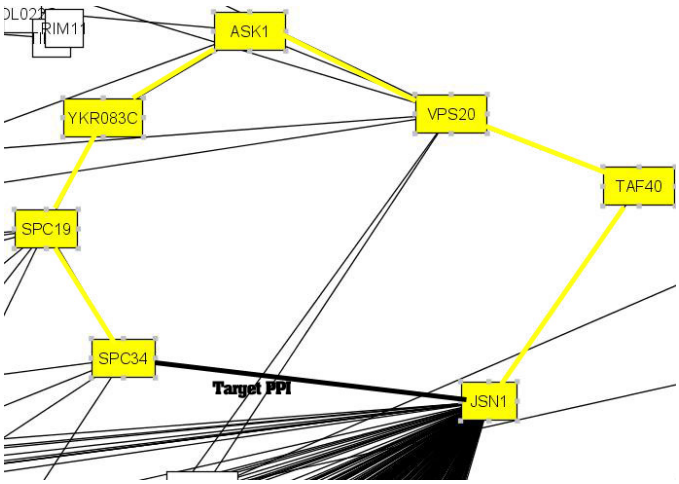

Figure 3.4: Example: strong alternative path indicating a strong positive PPI. GO Similarity(Spc34, Jsn1)=0.886994. IG1(Spc34, Jsn1)=0.103448. IRAP(Spc34, Jsn1)=0.504180. Path=Spc34-Spc19-Ykr083c-Ask1-Vps20-Taf40-Jsn1

3.5 AlternativePathFinder Algorithm

The yeast PPI network is very large in size and highly loopy. The network constructed for the yeast PPIs in our experiments has more than 4,000 nodes and 8,000 edges with many loopy components. Hence, it is necessary to develop an efficient method to find the strongest alternative path and compute the IRAP value for each candidate interacting pair (v_A, v_B) in G where G is a PPI network as described in Section 3.3.1.

Based on the definition of IRAP, the strongest alternative path is not necessarily the shortest path. For example, for 2 vertices u and v in V, if path P_a shares an edge with another path P_b but does not share any edge with path P_c, $|P_c| < |P_a| < |P_b|$, $Weight(P_c) < Weight(P_a) < Weight(P_b)$, then P_a is stronger than P_b and P_c according to the definition of IRAP, but P_a is neither the shortest weighted path nor the shortest unweighted path. Thus, standard shortest path algorithms, such as Dijkstra [Dij59], cannot be directly used here to find the strongest

alternative path. We develop a method called *alternative path finder* that utilizes a breadth first search to compute the IRAP values in a large undirected network. Algorithm 1 shows the details of the procedure.

The algorithm AlternativePathFinder first removes the edge (v_A, v_B) from the network, and initializes the weight W of node v_A to 1 and the rest of the nodes in the network to 0. In each iteration t, the algorithm computes $W(v) = \max(weight(v, v') * W(v'))$ for each node v in the current level, where v' is a node connected to v and $W(v') * weight(v, v') > W(v)$. The edge (v, v') is then removed from the network. The process stops when no more edge can be removed or when all the edges connected to v_B have been removed. Note that the function $append(p, v)$ appends the node v to the end of path p and returns the new path. The function $overlap(p, P)$ returns true if the path p overlaps with any path in the path set P.

Our algorithm is based on breadth first search (BFS). It terminates when all the edges of target pair have been removed. In the worst case, it traverses the whole graph. The computational time for each interaction pair is therefore linear to the number of edges, m. Since there are altogether m candidate interaction pairs, the total computational time is $O(m^2)$.

Consider again Figure 3.1 which shows 3 alternative paths between the nodes A and B. Two of the paths <*A-D-E-B*> and <*A-F-G-D-E-B*> have nodes D and E in common. Let us illustrate how the algorithm computes the IRAP value for the PPI between A and B.

First, we set the value for node A to 1 and the values for the remaining nodes to 0. The edge (A, B) is removed from the graph. After the first iteration, node values are propagated from A to C, A to D, and A to F. The edges (A, C), (A, D) and (A, F) are thus removed from the network. In the second iteration, node values are propagated from C to B, D to E, D to G, and F to G, and the edges (C, B), (D, E), (D, G) and (F, G) are removed from the network. In the final iteration, the node value is propagated from E to B, and the edge (E, B) is removed. This results in an empty graph, and the process terminates at this point with the $IRAP(A, B)$ given by the value at B.

Note that the path <*A-F-G-D-E-B*> was not traversed as the algorithm

Algorithm 1 AlternativePathFinder

Input: PPI network $G = (V, E)$;
Output: Set of $IRAP(v_i, v_j)$ for all edges $(v_i, v_j) \in E$;
Let $weight(v_i, v_j)$ denote the weight of edge $(v_i, v_j) \in E$, $W^{(t)}(v)$ denote the weight of node $v \in V$ in iteration t, p_v denote a path connecting v_A and v, P denote the set of paths connecting v_A and v_B, and p denote the strongest alternative path between v_A and v_B;
for each pair of interacting proteins (A, B) denoted by (v_A, v_B) **do**
  Set $t = 0$; $W^{(t)}(v_A) = 1$; $P = \emptyset$;
  for each node $v \in V - \{v_A\}$ **do**
    Set $W^{(t)}(v) = 0$;
  end for
  Remove edge (v_A, v_B) from E;
  repeat
    for each $(v_i, v_j) \in E$ & $W^{(t)}(v_j) > 0$ **do**
      if $v_j = v_B$ **then**
        Skip edge (v_i, v_j);
      end if
      $IRAP = W^{(t)}(v_j) \times weight(v_i, v_j)$;
      Remove edge (v_i, v_j) from E;
      if $IRAP > W^{(t)}(v_i)$ **then**
        $W^{(t+1)}(v_i) = IRAP$;
        $p_{v_i} = append(p_{v_j}, v_i)$;
        if $v_i = v_B$ & $overlap(p_{v_i}, P) = false$ **then**
          $IRAP(v_A, v_B) = IRAP$;
          $P = P + \{p_{v_i}\}$;
        end if
      end if
    end for
    $t = t + 1$;
  until (no more edge is removed) OR (all the edges connected to v_B have been removed)
end for

automatically selects the shorter path when the paths share some common nodes. In this case, path <*A-F-G-D-E-B*> shared two common nodes D and E with path <*A-D-E-B*>. Hence, only two paths <*A-D-E-B*> and <*A-C-B*> are traversed by the algorithm to propagate node values from A to B. The target node B is assigned the larger weight product of the two paths.

Next, we prove that the path p chosen by the algorithm has the maximum weight product among all the non-reducible paths between two nodes in G.

Theorem 3.5.1. *The algorithm AlternativePathFinder finds the strongest alternative path p such that*

$$\prod_{(u,v)\in p} weight(u,v) = IRAP(v_A, v_B)$$

Proof: Let v_A and v_B be the nodes that correspond to a pair of interaction proteins. Let $PATH = \{p_1, p_2, \ldots, p_k\}$ denote the set of paths between v_A and v_B that have at least one node in common. Let $\{v_1, v_2, \ldots, v_q\}$, $q \geq 1$ be the common nodes of the paths. We can partition $PATH$ into $q+1$ sets of subpaths, $PATH_1$,$PATH_2$, ..., $PATH_{q+1}$, where $PATH_1$ consists of paths from node v_A to v_1, $PATH_2$ consists of paths from v_1 to v_2, ..., $PATH_{q+1}$ consists of paths from v_q to v_B.

Consider the set $PATH_1$. Only the first path that reaches v_1 is allowed to propagate values to the nodes beyond v_1. This is because the algorithm removes an edge after propagating a value through the edge. Thus, the first path reaching the common node v_1 removes all the edges connecting to v_1. This path is also the shortest path from v_A to v_1 since the procedure propagates values from node v_A to the next nodes simultaneously along all the paths.

Similarly, in the sets $PATH_2$, ..., $PATH_{q+1}$, the first path that reaches the end node from the start node is also the shortest path. Hence, the algorithm finds all the shortest path(s) from node v_A to node v_B among all the paths in $PATH$. Given the definition of the non-reducible path in section 3.3, the shortest path(s) found by the algorithm in $PATH$ is non-reducible.

Thus, the algorithm finds all the non-reducible paths from node v_A to node v_B. Among them, the algorithm chooses p which has the maximum weight product. □

3.6 Heuristic IRAP

While IRAP has shown great promises, the *AlternativePathFinder* algorithm employed to determine IRAP is computationally expensive and cannot scale well. It has to traverse the PPI network once for each target interaction in order to find all non-reducible path(s). Figure 3.5 shows the runtime requirement of the *AlternativePathFinder* algorithm (on a 800 MHz Pentium III PC with 256 MB RAM) versus the increase in network size. We find that for 8,454 interactions, the

program requires more than half an hour. Clearly, this approach is infeasible for a large PPI network such as the *D. melanogaster*.

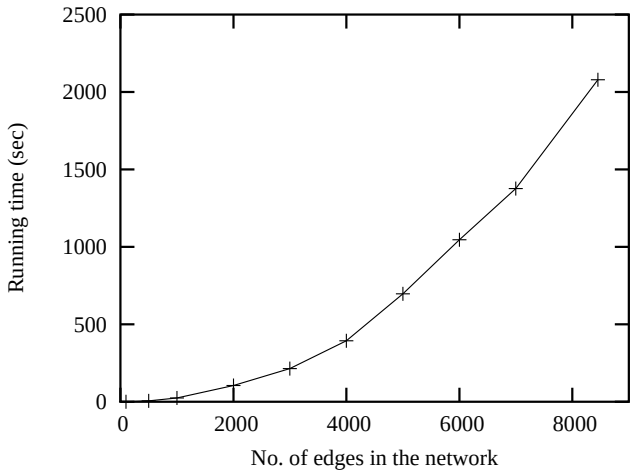

Figure 3.5: Running time of *AlternativePathFinder* versus network size.

To overcome this limitation, we introduce the use of heuristic search to speed up the computation of IRAP in large PPI networks. The essential idea is to make use of a well designed cost function to guide the search for the most promising path. Here, we adopt the best first search strategy. Each node that has been visited is assigned two values: The first value g is the cost from the source node to the node n, and the second value h is the estimated distance from n to the destination node. The node with the lowest $g + h$ value is the most promising node and will be visited first.

The key to ensuring a good speedup using heuristic search lies in the design of the cost function, namely the function to estimate the h value. In recent years, analysis of interactome data has highlighted the apparent scale-free behaviour of the observed PPI network [JMBO01]. Scale-free networks are characterized by an uneven distribution of connectedness. A selected number of nodes will serve as "very connected" hubs, while the rest of the nodes in the network will have very few neighbours. We call the former a hub node. The defining feature of scale-free networks is that the degrees of vertices (k) are distributed according to a power law: $f(k) \propto k^{-\gamma}$, where $\gamma > 0$ and $k = 0, 1, \ldots$. Hence, a plot of $log(degree)$ by

Degree	No. of hubs (Percentage)	Paths involve hub	Paths not involve hub	Path avg len w/ hub	Path avg len w/o hub
≥40	25 (0.6%)	61.2%	38.8%	5.23	5.28
≥30	39 (0.9%)	69.8%	30.2%	5.23	5.31
≥25	52 (1.2%)	74.7%	25.3%	5.21	5.41
≥20	87 (2.0%)	82.7%	17.3%	5.20	5.74

Table 3.2: Statistics on hubs in a PPI network.

$log(frequency)$ will show a decreasing linear trend.

Such kind of degree distribution greatly influences the way the network operates. Table 3.2 shows a summary of the percentage of paths involving at least a hub node, the percentage of paths not involving a hub node, and their average path lengths respectively. We observe that at 2% of hub nodes, the reduction in the average lengths of the paths with and without hub nodes is rather significant. A non-reducible path is highly likely to pass through a hub node. In other words, an alternative path involving a hub node is likely to be shorter than a path without any hub node. With this in mind, we design a function to estimate h (see Algorithm 2 and 3).

In Algorithm 2, we select the top 2 % nodes with highest degree as hub nodes and store them in a set V'. For each node $v_i \in V'$ $(1 \leq i \leq k)$, we use a breadth first search strategy to compute its distance to all other nodes u, $dist(v_i, u)$, in the graph G.

Once the distances from a hub node to other nodes have been computed, we proceed to perform the heuristic search. Suppose the source node is v_A and the destination node is v_B. For each node u in G, the estimated length of the remaining path, h, is given by the estimation function as detailed in Algorithm 3. Lines 4-6 deal with the case when v_B is a neighbor of u, then h is 1. Lines 8-11 check to see if u is a hub node and use the pre-computed distance if u turns out to be a hub node. Lines 12-14 perform the computation of the distance of u through all the hub nodes. Finally, in Line 15, we select the smaller of the shortest distance through the hub nodes and $(2D - g)$ where D is the diameter of graph G, and g is the sum of the length of path thus far. $(2D - g)$ is the upper bound of h as it denotes the longest path length $(2D)$ subtracting the path length from source node to node u.

We implemented the heuristic IRAP in C++ and evaluated its performance

Algorithm 2 SelectHubs

Input: PPI network $G = (V, E)$, number of hub nodes k;
Output: Set of selected k hub nodes V', and the distance $dist(v_i, u)$ for each $v_i \in V'$ and $u \in V - V'$;
for each node $v \in V$ **do**
    $degree(v)$ = No. of neighbours of v;
end for
Sort nodes with their degrees from the largest to smallest;
Let $V' = \{v_1, v_2, \ldots, v_k\}$;
for each node $v_i \in V'$, $1 \leq i \leq k$ **do**
    Compute the distance $dist(v_i, u)$ for all nodes $u \in V - V'$ with a breadth first search strategy;
end for

Algorithm 3 Estimate

Input: PPI network $G = (V, E)$, current node u, the initial node v_A and the destination node v_B;
Output: Estimated length of remaining path h for node u;
Let D be the diameter of graph G, and g be the sum of the length of path thus far;
if v_B is a neighbor of u **then**
    $h = 1$;
    return h;
end if
if $u \in V'$ **then**
    $h = dist(u, v_B)$;
    return h;
end if
for each $v_i \in V'$ **do**
    $h_i = (dist(v_i, u) + dist(v_i, v_B))$;
end for
$h = \min(h_1, h_2, \ldots, h_k, 2D - g)$;
return h;

with the *AlternativePathFinder* algorithm. Two sets of experiments are performed on the yeast PPI network. The first set of experiment aims to determine the speedup of heuristic IRAP over the *AlternativePathFinder* algorithm. The second set of experiments aims to show the accuracy attained by the heuristic IRAP as compared to the *AlternativePathFinder* algorithm.

Figure 3.6 indicates that as the network size increases, the ratio of the runtime for the AlternativePathFinder algorithm over the heuristic IRAP increases from 1.01 to 1.40. In other words, at 16,000 interactions network, we achieve a speedup of 40%. This indicates that it is feasible to run the algorithm on larger PPI networks, such

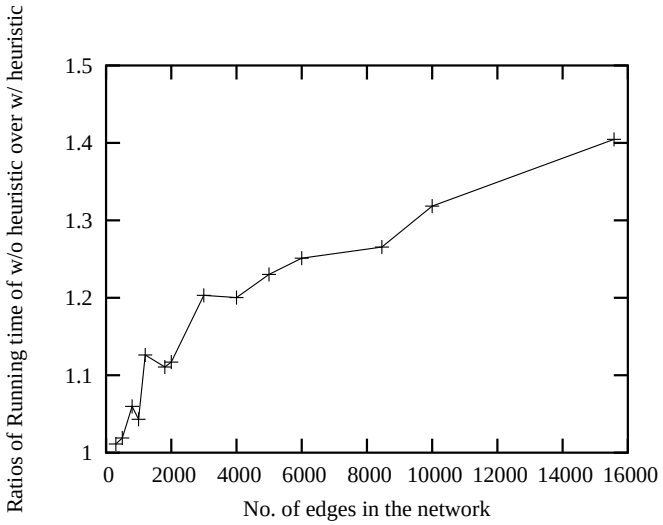

Figure 3.6: Speedup of heuristic search over AlternativePathFinder algorithm.

as *D. melanogaster* which has more than 20,000 PPIs.

However, while the speedup achieved is impressive, one worry is that the heuristic search may miss the optimal solution too often to make the results inaccurate. Therefore, an accuracy measure should be taken to show the quality of the heuristic IRAP. The next experiment examines the effect of network size on the accuracy of the heuristic IRAP. The accuracy measure is the number of PPIs, for which the heuristic IRAP correctly finds the strongest alternative paths, divided by the total number of PPIs.

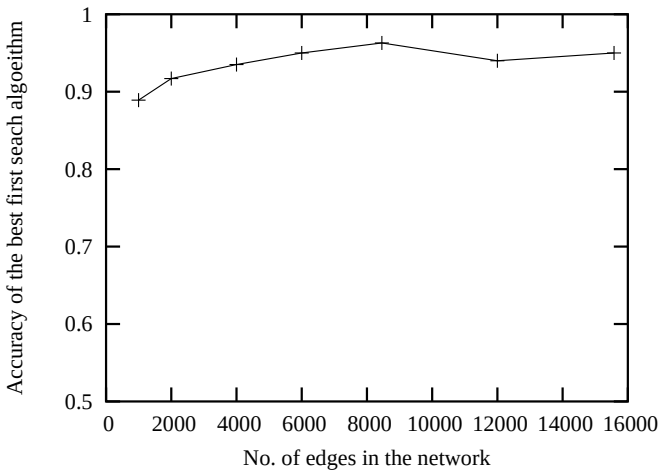

Figure 3.7: Accuracy of the heuristic IRAP.

Figure 3.7 shows the result of the experiment. We observe that once the network size exceeds 5000 interactions, the accuracy of the heuristic IRAP is relatively stable at a high degree of accuracy of around 95 %.

3.7 Experimental Results

We implemented[2] the alternative path finder algorithm and its heuristic IRAP algorithm in *C++*, and applied them to compute the IRAP values of protein interactions in large protein interaction networks generated by data from high-throughout genome-wide biological experimental methods. Then we validated the effectiveness of IRAP in the following ways.

3.7.1 Data Preparation

The experimental data were combined from the following publicly available yeast protein interaction datasets:

1. **Ito** *et al.* **[ICO$^+$01].** We downloaded the core dataset containing 841 protein-protein interactions available from the BRITE web site at KEGG [KGKN02]. The core set of Ito is formed by cases, in which the interactions have been detected more than three times by the two-hybrid assay; and

2. **Uetz** *et al.* **[UGC$^+$00].** We downloaded a dataset of 957 protein-protein interactions, also from the BRITE web site; and

3. **Munich Information Center for Protein Sequences (MIPS) [MFG$^+$02].** We obtain a dataset of 10,413 interactions (from the MIPS_PPI_120803 data file).

After combining these three datasets and removing redundancy from them, we had 8,454 protein-protein interactions involving 4,319 proteins. Note that this was a much larger set of interaction than the interaction dataset that Saito *et al.* previously

[2]The computer has a pentium IV 1.1G Hz CPU and 1G memory. The operating system is Linux Fedora Core 3.0.

used to evaluate their IG2 measure in [SSH02a]—much new interaction data have since been added to the above databases. For comparison, we also implemented the IG1 and IG2 algorithms as described in [SSH02b, SSH02a].

3.7.2 Validation of IRAP

We carried out a series of experiments to evaluate the effectiveness of using the computed IRAP values to detect reliable protein-protein interactions.

1. *Experimentally reproducible interactions.* We use protein interactions that have been detected by multiple independent experiments as the desired "gold standards". We show that the proportion of reproducible interactions increases in IRAP-filtered protein interaction data;

2. *Annotated functional associations.* By the 'guilt-by-association' principle [Oli00], true interacting proteins should share at least a common functional role. Here, we show that the proportion of interacting proteins with a common functional role increases in IRAP-filtered interaction data;

3. *Gene expression correlations.* Genes that are co-expressed indicate that their gene products (the proteins) partake in the same pathway—the corresponding proteins are thus highly likely to be interacting. Here, we check whether the *IRAP*-filtered interactions can be confirmed by co-expression at the mRNA level;

4. *Cellular localization cross-talks.* For two proteins to be interacting *in vivo*, they should at least be at a common cellular localization. We check here that the rate of cellular localization cross-talk is decreased in *IRAP*-filtered interactions, indicating a reduced degree of biologically irrelevant interactions in the post-IRAP data.

5. *Biologically interacting cross-talkers.* Biologically genuine cross-talkers, such as the proteins involved in signal transduction pathways, share same functions but are not co-localized. We check the IRAP model and found that a

large proportion of the cross-talking interactions with high IRAP values have functional matches.

6. *Many-Few interaction trend in protein networks.* Maslov *et al.*[MS02] found that there is a "many-few" interaction pattern in protein interaction networks. The proposed IRAP model indicates that as the IRAP thresholds increased, the proportion of "many-few" interactions also increases in the IRAP-filtered interaction data. This result provides yet another biological validation of IRAP.

Experimentally-Reproducible Interactions

Protein interactions that are confirmed by multiple independent experiments[3] are often regarded as highly reliable. In the combined dataset, 2,394 (that is, ∼28%) experimentally reproducible interactions are confirmed by at least two independent experiments. We use this set of reproducible interactions as the "gold standard" to estimate the degree of true positives in our IRAP-filtered interaction data.

Figure 3.8 shows the ratios of experimentally-reproducible (reliable) interactions over the non-reproducible ones found in sets of protein interactions filtered with various IRAP values. The proportion of reliable experimentally reproducible interactions increased with higher IRAP values, as more of the unreliable experimental interactions were filtered away by the higher IRAP thresholds. The curve in figure 3.8 indicates that IRAP is effective in detecting reliable protein interactions from high-throughput experimental data.

We also compared the performance of IRAP with Saito *et al.*'s Interaction Generality measures IG1 and IG2 based on their average values in the class of reproducible interactions and non-reproducible interactions. Table 3.3 shows the different mean values and standard deviation values for IG1, IG2, and IRAP. The difference between the mean values of IRAP for reproducible interactions and non-reproducible interactions was much more pronounced than the corresponding differences between the mean values for IG1 and IG2 (0.13 *vs.* 0.07 and 0.05). The differences indicate

[3]This includes interactions that are symmetrically detected in yeast-two-hybrid screens: namely, protein A(bait)–protein B(prey) and protein B(bait)-protein A(prey) are both positive.

that the performance of IRAP is clearly better than IG1 and IG2 for identifying reproducible and non-reproducible protein interactions. The excellent performance of IRAP could be explained by the topological nature of the alternative paths. The alternative path approach provides a comprehensive interaction reliability measure that does not impose any restriction on the number of intervening proteins. In contrast, IG1 and IG2 are local measures since their topological context involves only directed neighbors or five topological components of a neighbor C. As such, both the IG1 and IG2 measures do not consider the underlying system-wide topological structure of the entire interaction network as IRAP does to determine the reliability of the discovered protein interactions.

We note that table 3.8 also shows that IRAP has a relatively higher standard deviation value(0.28) than IG1(0.05) and IG2(0.08). This is because about 14% overlapping interactions in the target network have no alternative path, and thus result in zero IRAP value. By excluding these interactions, the corresponding standard deviation value for IRAP decreases to 0.14, which is close to the standard deviation value of IG1 and IG2. The experiment shows that IRAP is better than IG1 and IG2 if the target protein interacting pair has at least one alternative path.

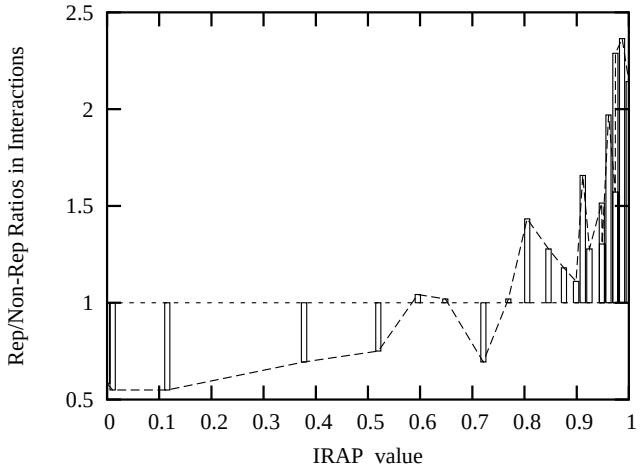

Figure 3.8: Ratio of experimentally reproducible interactions ("rep") over the non-reproducible ones ("non-rep") increases as PPIs are filtered with higher IRAP values.

	Reproducible		Non-Rep		Diff-
	Mean	Dev	Mean	Dev	erence
IG1	0.9564	0.05	0.8967	0.12	0.0597
IG2	0.9190	0.09	0.8487	0.15	0.0703
IRAP	0.7467	0.28	0.6162	0.36	0.1304

Table 3.3: Mean and standard deviation values for IG1, IG2 and IRAP.

Functional Associations

The 'guilt-by-association' approach [Oli00] has been used widely to infer the functional roles of unknown proteins by using the principle that interacting proteins should share at least a common functional role. Here, we used this principle to evaluate the performance of IRAP in filtering false positives from large sets of experimental protein interaction data. By the 'guilt-by-association' principle, we expect that as the value of IRAP increases, the proportion of interacting proteins with a common functional role also increases in the resulting IRAP-filtered data.

We referred to the Comprehensive Yeast Genome Database at MIPS[MFG+02] available at *http://mips.gsf.de/genre/proj/yeast* for reference functional annotations of the yeast proteins. We used the MIPS annotation dated 03-06-25 in our experiment here. Out of the 4,319 proteins and 8,454 interactions in our original dataset, 3,150 proteins and 4,743 interactions had functional annotations. Only 61% of the interactions share at least one common cellular role. **[E1, E2]**In Figure 5.14, we show the effect of IRAP as a filtering measure: as the IRAP threshold was increased, the proportion of interacting pairs with common cellular roles increased from 61% to 87%, **[E3, generalization]**indicating an increased proportion of true protein interactions in the filtered interaction data by the 'guilt-by-association' principle. The experiment results matched what we predicted very well.

For comparison, we also showed the performance of IG1 and IG2 in the figure. With IG2, the proportion of interacting pairs with common functional roles increased from 61% to about 73%; and with IG1, the proportion increased from 61% to 68%. The performance of IRAP is clearly better than IG1 and IG2 for identifying true protein interactions. This is because, unlike IG1 and IG2, IRAP

considers topological content involving the proteins that are not directly connected to, but have at least one solid path to the target proteins. These proteins may form circular contigs to provide higher level functions. The circular contigs with known proteins have been detected in macromolecular complexes as well as signal transduction pathways [WSL+00, WBV00].

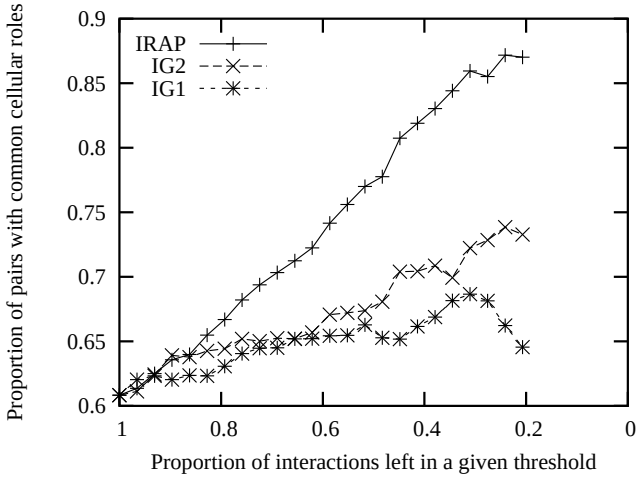

Figure 3.9: Proportion of interacting proteins with common cellular functional roles increases at different rates under different interaction reliability measures.

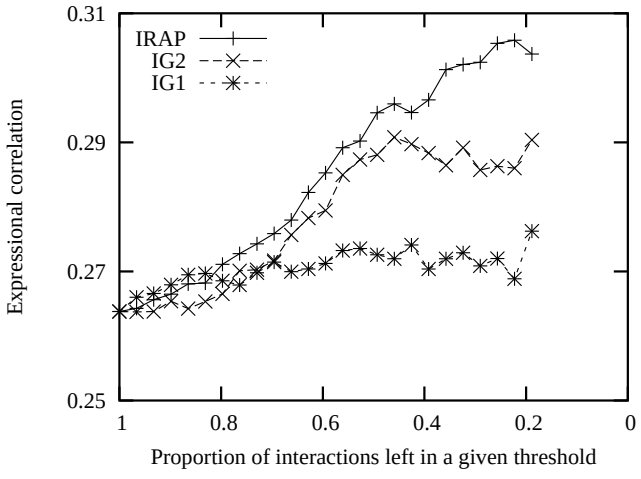

Figure 3.10: Overall correlation of gene expression for interacting proteins increases at different rates under different interaction reliability measures.

Gene Expression Correlations

Studies have also shown that the average correlation coefficient of gene expression profiles that corresponds to interacting protein pairs is significantly higher than those that correspond to random pairs [Gri01, DSXE02]. If IRAP is an effective measure for assessing protein interactions, then we should find that interacting protein pairs with higher IRAP values are more likely to be co-expressed.

To evaluate if this is true with our dataset, we downloaded the yeast gene expression dataset from Eisen's Lab [ESBB98](*http://rana.lbl.gov/EisenData.htm*). The dataset comprised expression vectors from 80 experiments on 6,221 yeast genes, 4,287 of which had their corresponding proteins in our interaction dataset. We computed the average correlations of gene expression for protein partners with different IRAP thresholds, and show in figure 5.16 that gene expression correlations increased from 26.4 to 30.5, as the IRAP threshold increased. The experimental results indicate that protein interactions with higher IRAP values also have higher gene expression correlation. Because high gene correlation corresponds to the true interacting protein pair, the high IRAP values imply an increased possibility of true positives.

We also compared the performance of IRAP with that of IG1 and IG2 in figure 5.16. As the threshold of IG1 and IG2 increased, gene expression correlations increased from 26.4 to 27.6 and 29.0 respectively. The result shows that IRAP once again performs better than IG1 and IG2. A reasonable explanation is that in a small-world protein interaction network, high clustering coefficient property predicates that proteins are likely to form dense clusters by interactions[Wag01]. Therefore, the denser the alternative paths between two nodes are, the more likely that the edge connecting the two nodes is true positive.

Cellular Localization Cross-talks

An experimentally-detected PPI can still be a false positive in the biological sense. An example is an interaction involving two proteins in different cellular localization—it is most likely an *in vivo* false positive. We use this principle to check if the rate of

cellular localization cross-talk is decreased in IRAP-filtered interactions, which will indicate that IRAP is an effective measure for reducing the degree of false-positives.

We refer again to the MIPS[MFG+02] database for the cellular localization annotation dated 03-03-21 of the yeast proteins. Out of the 4,319 proteins in our 8,454 interaction dataset, we have 2,588 proteins with known cellular localizations and 4,188 PPIs involving these proteins. Only 49.5% of our original interactions involved proteins with an annotated cellular location, and 85.3% of them share a common cellular location.

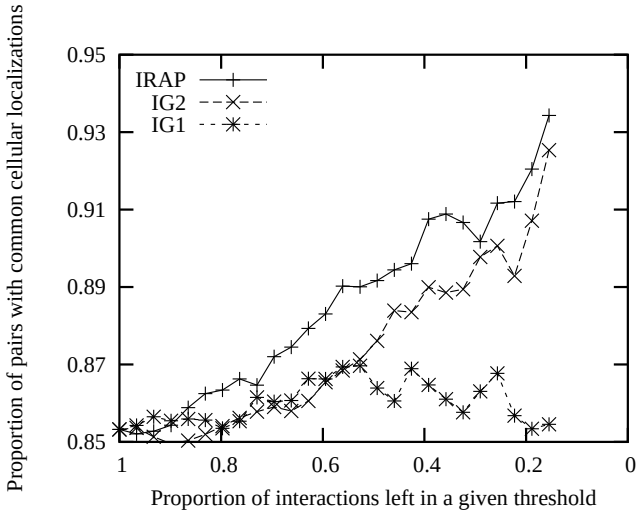

Figure 3.11: Proportion of interacting proteins with common cellular localizations increases at different rates under different interaction reliability measures.

Figure 5.15 shows that as the IRAP threshold is increased, the proportion of interacting pairs with common cellular localization increases from 85.3% to 93.4%, indicating that the rate of potential cellular localization cross-talk has decreased in PPI data filtered with IRAP values. The corresponding performance for IG1 and IG2 is also shown for comparison. Again, IRAP is a better indicator for true PPIs under the cellular localization cross-talk criterion, consistent with the results in all the other experiments.

Biologically interacting cross-talkers

From our cellular co-localization experiment in the previous section, we also observe that there are 257 protein pairs with very high IRAP values (≥ 0.95) that do

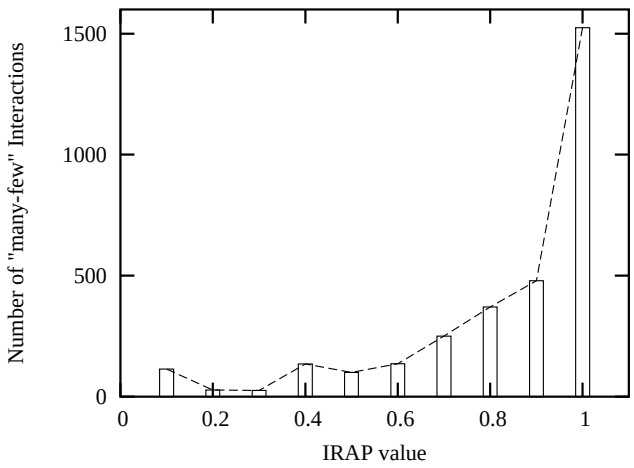

Figure 3.12: Distribution of "many-few" interactions increases with higher IRAP values. Protein with less than 10 interacting partners is a "few" protein; otherwise it is a "many" protein.

not co-occur in the same cellular localization. On closer inspection, we find that a large proportion (53%) of these cross-talking interactions have functional matches based on MIPS [MFG+02] [4], suggesting that these interactions are highly likely to be biologically genuine cross-talkers, such as those involved in signal transduction pathways. Signal transduction refers to the movement of biological signals from outside the cell to inside by proteins that can interact *in vivo* with partners across subcellar boundaries (i.e., they are not co-localized). As in the previous evaluation (Section 3.7.2) where we have used co-localization as a necessary criterion for interaction, many current PPI prediction methods also exclude non-co-localized protein pairs in their training data [JYG+03]. As a result, they are often inadequate for detecting the cross-talkers.

Our IRAP method can be useful for recognizing cross-talking protein pairs by detecting high IRAP interactions that involve non-co-localized protein pairs. Table 3.4 shows some examples of non-co-localizing PPIs with high IRAP values that are involved in a common functional pathway such as signal transduction.

[4]In comparison, only 14% of the interactions with low IRAP value (< 0.1) have functional matches.

ProteinA	Cellular Localization	ProteinB	Cellular Localization	Functional Pathway
YDR299w	nucleolus-protein transport	YLR208w	cytoplasm-release of transport vesicles from ER	Vesicular transport (Golgi network)
YOL018c	endosome, ER-syntaxin SNARE	YMR117c	spindle pole body-spindle pole component	Cellular import
YDL154w	nucleus-recombination	YBR133c	cytoplasm- neg. regulator of kinase	Meiosis and budding
YGL192w	nucleus-put. Adenosine methyltransferase for sporulation	YBR057c	cytoplasm-meiosis potentially in premeiosis DNA synth	Development of asco-basido -zygo spore
YDR299w	nucleolous- protein transport	YPL085w	cytoplasm,ER-veiscle coat protein interacts cytoplasm, with sec23p	both in vesicular transport
YEL013w	vacuole-phosphorylated protein which interacts with Atg13p for cyto to vacuole targeting vacuole targeting	YFL039c	cytoskeleton-actin	Protein targeting and budding

Table 3.4: Examples of interactions with high IRAP values (≥ 0.95) between non-co-localized proteins ("cross-talkers") involved in the same cellular pathway

Many-Few interaction trend in protein networks

Maslov *et al.*[MS02] quantify the correlations between the connectivities of interacting nodes in protein networks and compare them to a null model of a network, in which all the links are randomly rewired. Protein with less than 10 interacting partners is defined as a "few" protein; otherwise it is a "many" protein. They find that there is a "many-few" interaction pattern in PPI networks—that is, links between highly connected proteins are systematically suppressed, whereas those between a highly connected and low-connected pairs of proteins are favored. Biologically, this effect decreases the likelihood of cross talk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of deleterious perturbations.

Saito *et al.*[SSH02a] report that they could not confirm with their IG2 values Maslov *et al.* [MS02]'s recent findings about the observed specificity and stability of protein networks. We test the IRAP model on our PPI data and find that unlike IG2, the IRAP values are consistent with Maslov *et al.*'s "many-few" interaction trend. As shown in Figure 3.12, as the IRAP thresholds increase, the proportion of "many-few" interactions also increases in the IRAP-filtered (reliable) interaction data. This result provides yet another biological validation of our IRAP model over other alternative models.

Discover interactions within a protein complex

A well studied protein complex is chosen to see how many true interactions could be discovered inside it. Because the size of protein complex is much smaller than the whole PPI network, we could test every possible links and select the most reliable ones.

In the experiment, we selected *Saccharomyces cerevisiae* 26S Proteosome Complex as a case to study. Out of the 36 proteins in Proteosome Complex, 34 proteins could be found in our dataset. This complex is selected because it is well studied and has 3D interaction information in PDB database.

Saccharomyces cerevisiae 26S Proteosome Complex contains 36 proteins in BIND database. It also contains 4 structures in PDB covering 13 proteins and 23 physical interactions, only 1 exists in our dataset.

In the experiment, IRAP detects 45 interactions between 13 proteins: 14 of these are the physical interactions recorded in PDB (c.f. 1 exists in the original dataset); another 12 are experimentally detected interactions in BIND only; 19 new interactions are also predicted.

3.8 Conclusions

The dissection of the protein interactome is important for extracting invaluable biological knowledge for understanding the molecular mechanism of our cellular system, and eventually leading to the discovery of new drugs and drug targets for various human diseases. Thus far, most of the recent technological advance in this field has focused on the high throughput detection of PPIs in order to map the tremendously vast protein interactome. Unfortunately, the interaction data that have been generated in large-scale experimental studies using the high throughput technologies have very high error rates. In this work, we therefore focused on tackling the problem of high false positive rates in high-throughput experimental PPI data.

We proposed the use of a novel measurement—*Interaction Reliability by Alternative Path (IRAP)*—to computationally assess the reliability of candidate PPIs by using the topological properties of the underlying interaction network. We de-

veloped an algorithm called *alternative path finder* to compute the IRAP values efficiently in large, interconnected, and loopy PPI networks. Results from our extensive experiments showed consistently that our IRAP measure is an effective way for discovering reliable PPIs in large datasets of error-prone experimentally-derived PPIs. Our results also indicated that IRAP is better than IG2, and markedly better than the more simplistic IG1 measure, which shows that a global, system-wide approach—such as our IRAP measure that considers the entire interaction network instead of merely local neighbors—is a much more promising approach for assessing the reliability of PPIs.

Our IRAP measure is currently based on the "strongest alternative path" model. A candidate interaction that is not accompanied by a strong alternative path of interactions in the overall interaction network is considered to be unreliable under this model. While this may not be true for all the biologically relevant PPIs, we have performed an analysis on our yeast-two-hybrid PPI datasets and found that more than 80% of PPIs in our experiments do have at least one alternative path. This suggests that a significant proportion of PPIs is captured by the current IRAP model. Our next step is to develop further network models to capture PPIs associated with more sophisticated topological characteristics than alternative paths. Combined with our current IRAP model, we hope to be able to detect errors in interaction data effectively. This will facilitate the rapid construction of PPI networks that will help scientists in understanding the biology of living systems.

CHAPTER 4

IRAP*: Repurify protein interactomes

Experimental limitations in high-throughput protein-protein interaction detection methods have resulted in low quality interaction datasets that contained sizable fractions of false positives and false negatives. Small-scale, focused experiments are then needed to complement the high throughput methods to extract true protein interactions. However, the naturally vast interactomes would require much more scalable approaches.

We describe a novel method called IRAP* [CHLN06c] as a computational complement for repurification of the highly erroneous experimentally-derived protein interactomes. Our method involves an iterative process of removing interactions that are confidently identified as false positives and adding interactions detected as false negatives into the interactomes. Identification of both false positives and false negatives are performed in IRAP* using interaction confidence measures based on network topological metrics. Potential false positives are identified amongst the detected interactions as those with very low computed confidence values, while potential false negatives are discovered as the undetected interactions with high computed confidence values. Our results from applying IRAP* on large-scale interaction data sets generated by the popular yeast-two-hybrid assays for yeast, fruit fly and worm showed that the computationally repurified interaction data sets contained poten-

tially lower fractions of false positive and false negative errors based on functional homogeneity.

4.1 Introduction

Although recent progress in high-throughput experimental techniques [UGC+00, ITM+00, MKS+02] has provided us with much more protein-protein interaction (PPI) data than those accumulated using traditional detection methods from the past decades, we are still far from being able to unravel the actual interactomes completely. This is because while the new experimental methods have allowed PPIs to be detected *en masse*, it was done at the expense of data quality. As stated in the previous chapter, the PPI data generated by the high-throughput methods are by no means at the same level of quality as those that were painstakingly generated by conventional small-scale, focused experimental approaches. For example, recent surveys [LWG01, MKS+02, SSM03] have revealed that the interaction data detected by the popular yeast-two-hybrid (Y2H) assay may contain as much as 50% false positives. At the same time, the false negative rate of the Y2H-constructed interaction map for *S. cerevisiae* interaction maps has also been estimated to be as high as 70% [DMSC02]. Such alarmingly high levels of errors in terms of both false positives and false negatives greatly diminish the potential usefulness of the experimental data made available by technology.

As a result, further carefully-focused small-scale experiments are often needed to complement the large-scale methods to validate the detected interactions. However, the vast interactomes require much more scalable and inexpensive approaches. In this chapter, we explore the possibility of computationally "repurifying" the highly erroneous experimentally-derived interaction data sets. We propose a computational method that uses only network topological metrics to sort the PPIs according to their computed reliability [CHLN06c]. Our proposed method can then identify false positives amongst the detected interactions that have low reliability values, and false negative from the putative interactions with high computed reliability values. By iteratively removing false positives from the interactomes and replacing them with

interactions that are identified as false negatives, we will show that the computationally repurified PPI data sets contain potentially lower proportions of false positive and false negative errors than the original experimentally-derived interactomes.

The rest of this chapter is structured as follows. In Section 4.2, we describe the related work and motivation for our topological approach. In Section 4.3, we describe IRAP*, our iterative computational repurification method, and how it uses network topological metrics to identify false positives and false negatives in the experimentally detected interactomes. We report, in Section 6.4, evaluation results from applying IRAP* on actual large-scale interaction data sets generated by the popular Y2H assays for yeast, fruit fly and worm that show that the computationally repurified interaction data sets contain potentially lower fractions of false positive and false negative errors. We then conclude in Section 6.6 with a summary and discussions about possible further work.

4.2 Background

The potential benefits of the current technological advances in large-scale PPI detection have been largely diminished by the abundant presence of both false positives and negatives in the resulting experimental data. Other than resorting to small-scale, focused experiments to selectively validate the PPIs of interest in the interactomes, researchers have also begun to consider various bioinformatics approaches to identify the false negatives and false positives in the PPI data sets.

The problem of false negatives was typically addressed by PPI prediction methods. Researchers have proposed numerous approaches to predict PPIs using a variety of biological information from genome sequences[TO00, MVR$^+$01, WS01] to 3D structures[ABC$^+$04]. However, these prediction methods depended on, and are limited by, the availability and richness of biological information; few have attempted to make use of the existing interactions in a PPI network to predict new interactions for the interactome. In this work, we focus on the latter by directly discovering the false negatives amongst the missing interactions in the experimentally derived PPI networks, using only the topological information from the PPI networks.

In terms of false positive detection, a naive approach would be to use the intersection of PPI data sets derived from various independent methods to detect reliable interactions. However, because of the different and limited coverage of various PPI detection methods, this operation would leave only few interactions [DSXE02] and it is therefore suitable only for identifying a small set of highly confident interactions. As such, some researchers have recently started to explore alternative approaches, such as the use of expected topological characteristics of PPI networks to assess the reliability of the experimentally detected interactions mathematically. Saito *et al.* pioneered such approach by developing a series of computational measures called *interaction generalities* (IG1 and IG2) [SSH02b, SSH02a] that used local topology and the statistics of adjacent interactions to detect false positives in a PPI network.

Motivated by the success of interaction generalities, we have proposed an alternative topologically-based quantitative measure called "Interaction Reliability by Alternative Path" (IRAP) [CHLN04, CHLN05b] in the previous chapter. Instead of using small-sized, predefined network motifs such as those in the interaction generalities, IRAP uses the observation that alternative paths are often present in many real-world networks (for example, there is often more than one way to fly from one city to another in an airline network), and computes the reliability of a detected PPI with respect to the presence of alternative reliable interaction paths in the underlying network Our evaluation results reported in our previous works showed that IRAP outperformed the IG measures in the system-wide detection of false positives in the yeast interactome.

Both the problems of false positives and false negatives must be addressed together in order to improve the usability of experimentally-derived interactomes. In this work, we extend our previous IRAP scoring method to detect both false positives and false negatives, using different topological metrics as the initial weights. We call the IRAP-based iterative approach for computational interactome repurification IRAP*, which stands for *I*nteractome *R*epurification by *A*lternate *P*aths, with the asterisk indicating that it is an iterative process. We will show in this work that our approach works well in other species' interactomes in addition to the commonly evaluated yeast interactome.

4.3 Method

Computationally, an experimentally-derived protein interactome can be modeled using an undirected network $G = (V, E, w)$. Each node in the network represents a unique protein in the species' proteome. An edge exists between two nodes v_A and v_B if there is a detected interaction between the corresponding proteins A and B. The PPI network is modelled as a weighted graph to account for the existence of experimental errors in the interactome, with $w(v_A, v_B)$ as the weight of edge (v_A, v_B) that serves also as a reliability index for the PPI.

In last chapter, we used a quantifiable measure called IRAP to evaluate the reliability of a detected PPI with respect to the presence of a reliable alternative interaction path between the two proteins in the global PPI network.

As stated in the previous chapter, IRAP is defined as follows:

Definition 3. IRAP. *The reliability of a candidate PPI* (A, B)*,* $IRAP(A, B)$*, is indicated by the collective reliability of the strongest alternative path of interactions connecting the two proteins in the underlying PPI network.*

$$IRAP(A, B) = \max_{\phi \in \Phi(A,B)} \prod_{(u,v)\in\phi} w(u, v) \tag{4.1}$$

where $w(u, v)$ *denotes the weight value for edge* (u, v) *in the PPI network* G*;* $\Phi(A, B)$ *denotes the set of non-reducible paths.*

4.3.1 False Positive Detection

As mentioned earlier, some proteins in Y2H assays tend to activate transcription of the reporter gene without actually interacting with their partners, leading to an excess number of candidate partners (false positives) detected for the proteins [SG01]. The IG1 topological metric proposed by Saito *et al.* [SSH02b] was designed to detect these easily-identified 'sticky' proteins. Since in this work, we are focusing on Y2H-derived experimental data sets (Y2H being the most popular hight-throughput PPI detection method), for false positive detection, we use IG1 values as a basis for the PPI network's initial weights:

$$IG1^G(v_A, v_B) = 1+ \\ |\{(v'_A, v'_B) \in E | v'_A \in \{v_A, v_B\} \& \ deg^G(v'_B) = 1\}| \tag{4.2}$$

$$w_+(v_A, v_B) = 1 - \left(\frac{IG1^G(v_A, v_B)}{IG1^G_{max}} \right) \tag{4.3}$$

where $IG1^G_{max}$ is the maximum IG1 value in the interaction network, and $w_+(v_A, v_B)$ is the initial weight for edge (v_A, v_B).

Using (4.1) on the resulting interaction graph $G = (V, E, w_+)$, we compute an interaction reliability index for all the detected PPIs in the interactome. The IRAP scores for each of the edges in the interaction graph can be computed efficiently using the algorithm in our previous work[CHLN04]. Those PPI's that are low in the IRAP reliability index can then be considered as potential false positives.

4.3.2 False Negative Detection

While IG1 was useful as initial weights for false positive detection, it turns out that a different topological metric is required for false negative detection. This is because IG1 tends to assume interaction links to be true unless there are topological evidence from the immediate neighbors to suggest otherwise. As such, while it was well-designed for detecting false positives in Y2H data, when used for false negative detection, IG1 will tend to overestimate the reliability for the missing links during the false negative detection process. For example, under IG1, all missing orphan links will be identified as false negatives since they will all have the lowest (strongest) IG1 values because there are no immediate neighbors to suggest otherwise. As such, for effective false negative detection, we would need a more stringent topological metric that assumes the missing links to be true negatives unless there are topological evidence to suggest otherwise.

Therefore, for false negative detection, we adopt a non-IG1 topological metric that is based on a different observation that interactions between proteins having a large number of common neighbors tend to be true interactions themselves. This was shown previously[KI05] to be a useful measure for predicting interactions in genetic networks. Here, we apply such a common neighbor counting approach to

compute the initial weight for edges (v_A, v_B) in a PPI network for false negative detection:

$$ComNbr^G(v_A, v_B) = |N(v_A)| \cap |N(v_B)| \tag{4.4}$$

$$w_-(v_A, v_B) = \frac{ComNbr^G(v_A, v_B)}{ComNbr^G_{max}} \tag{4.5}$$

where $ComNbr^G_{max}$ is the maximum number of common neighbors in the interaction network, and $w_-(v_A, v_B)$ is the initial weight for edge (v_A, v_B).

We then use IRAP as defined in (4.1) on the resulting interaction graph $G = (V, E, w_-)$ to assign a refined reliability value to each candidate false negative. Those missing links having high IRAP values can then be treated as potential false negatives.

However, unlike false positive detection, a major challenge for false negative detection is the huge number of candidate false negatives. For instance, the size of the yeast proteome is about 6,000 proteins and the current PPI network detected for it is about 10,000. The number of false negative candidates to be considered could be as many as $(6000 \times 5999)/2 - 10000 \approx 1.8 \times 10^7$. It is clearly impractical to evaluate the reliability of every possible link. Fortunately, as we are interested in only the top $\frac{k}{100} \times |E|$ false negatives at each iteration of computational repurification of the interactome with IRAP* (see Section 4.3.3 later), there is no need to consider all the false negative candidates. As such, we use the following heuristic approach to consider the most possible candidates first.

First, each node v_i in network $G = (V, E, w_-)$ is assigned a value h_i as:

$$h_i = 1 - \prod_{j \in N(v_i)} (1 - w_-(v_i, v_j)) \tag{4.6}$$

where $N(v_i)$ is the direct neighbor set of node v_i in network G. Then, the top Υ proteins with the highest h values are pairwise joined to form a candidate set I. In a sparse network, where the number of the edges is linear to the number of the nodes, the action to add α new links is of the same scale to add new links to connect at most 2α nodes. We can thus define the Υ as: $\Upsilon = 2 \times \frac{k}{100} \times |E|$. We can then

compute the IRAP scores for those protein pairs that are in I but not in E.

The edges in I are good false negative candidates since their associated alternative paths are likely to have high IRAP values given the high neighbor edge weights of the nodes. This heuristic is effective in reducing the candidate space—in such a sparse network as the PPI network, the h value for a large proportion of nodes are 0.

4.3.3 IRAP*: Iterative Refinement of Interactome

The previous chapter focused on using IRAP for false positive detection. In this chapter, we extend IRAP to also detect false negatives by using a different network topological metric as the initial scoring scheme. Next, we will describe IRAP*, an iterative application of IRAP for increasing the confidence of protein interactomes by computationally "repurifying" the interactomes from its experimental errors. This is achieved by iteratively removing false positive and false negative interactions from the interactomes using IRAP. In each iteration, we remove top $k\%$ of the pseudo-false positive interactome. Then, we add an equal number of new interactions that have been identified as top pseudo-false negatives into the interactome. In other words, each iteration consists of the following steps:

- *Step 1: False Positive Detection.* Compute the reliability index for all the interactions in $G = (V, E, w_+)$ using IRAP. Replace E with E', the set of $(1 - \frac{k}{100})|E|$ interactions that have the highest IRAP values.

- *Step 2: False Negative Detection.* Compute the IRAP-based reliability index for all the putative interactions in I derived from $G' = (V, E', w_-)$ as described above (Section 4.3.2), and determine the new edge set E'', the set of $\frac{k}{100}|E|$ new interactions that have the highest IRAP values in G'.

- *Step 3:* Set $E = E' \cup E''$ for the next iteration.

4.3.4 Step-by-Step Example of IRAP*

We illustrate here how IRAP[CHLN04, CHLN05b] and IRAP* works using real PPIs from our *Saccharomyces cerevisiae* dataset. For clarity, we only show the subset of PPIs between 14 proteins. The original interaction (sub)network between these proteins, as detected by Y2H screens, is shown in the figure below.

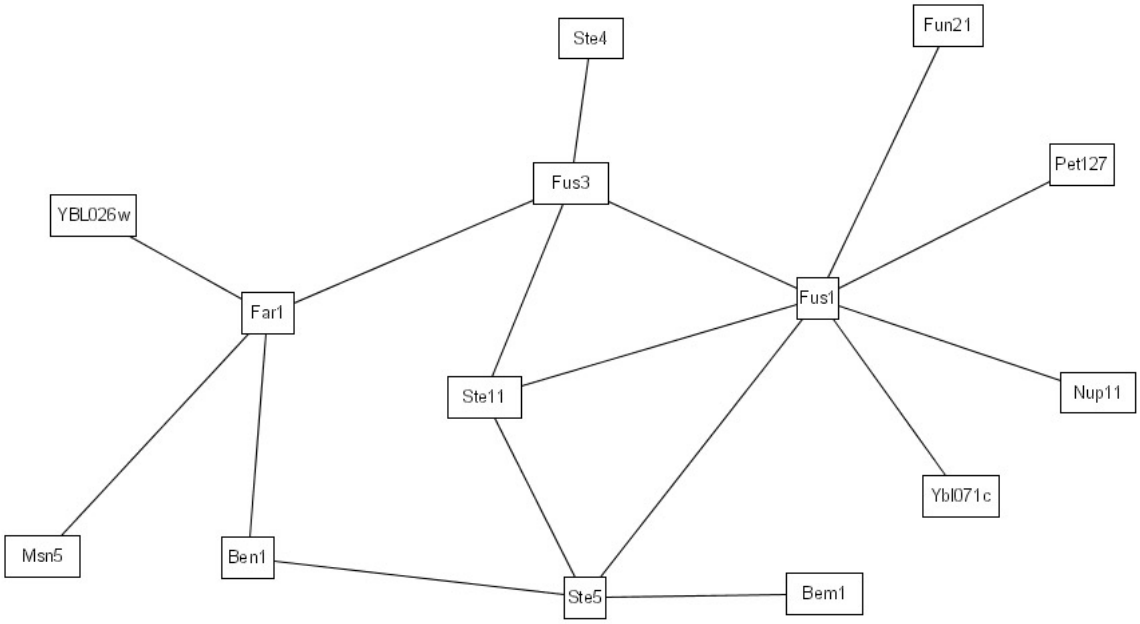

Figure 4.1: The subset of PPIs between 14 proteins.

4.3.5 IRAP - Single-Pass False Positive Detection

The previous chapter on IRAP only does a single-pass evaluation on the interactome to detect potential false positives. First, it ranks the various detected interactions with an IG1-based initial weight as follows:

We then perform the IRAP algorithm (see [CHLN04, CHLN05b] for details) to compute the interaction reliability values for the various interactions based on their alternative paths. The IRAP ranking of the interactions are as follows:

The above IRAP ranking can then be used as a reliability index to filter potential false positives (those with low IRAP values) from the detected interactome. The previous chapter reports biological evidence based on functional homogeneity,

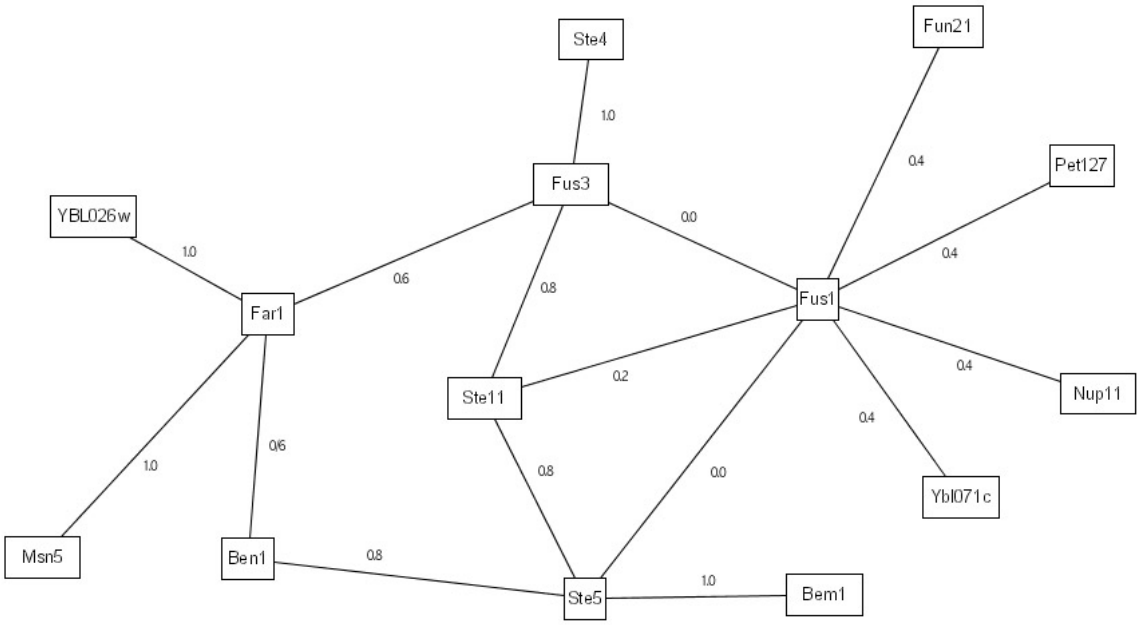

Figure 4.2: The subset of PPIs with IG1 weight.

cellular co-localization, and gene co-expression that the IRAP-ranking is superior to corresponding rankings by IG1 and IG2.

4.3.6 IRAP* - Iterative Removal of False Positives and False Negatives

With IRAP*, we built on IRAP to formulate an iterative framework for removal of both false positives as well as false negatives. Removal of false positives is carried out in a similar fashion as the above - using IRAP with IG1-based initial weights. Removal of false negatives is carried out by computing a similar weight—in this case, it is IRAP with common neighbor counting instead of reversed IG1—for each of the undetected interactions in the interactome. Potential false positives are identified amongst the detected interactions as those with very low computed confidence values, while potential false negatives are discovered as the undetected interactions with high computed confidence values. Figure 4.4 shows the differences in IRAP and IRAP*:

To continue with our previous example, in IRAP*, the interactions are first

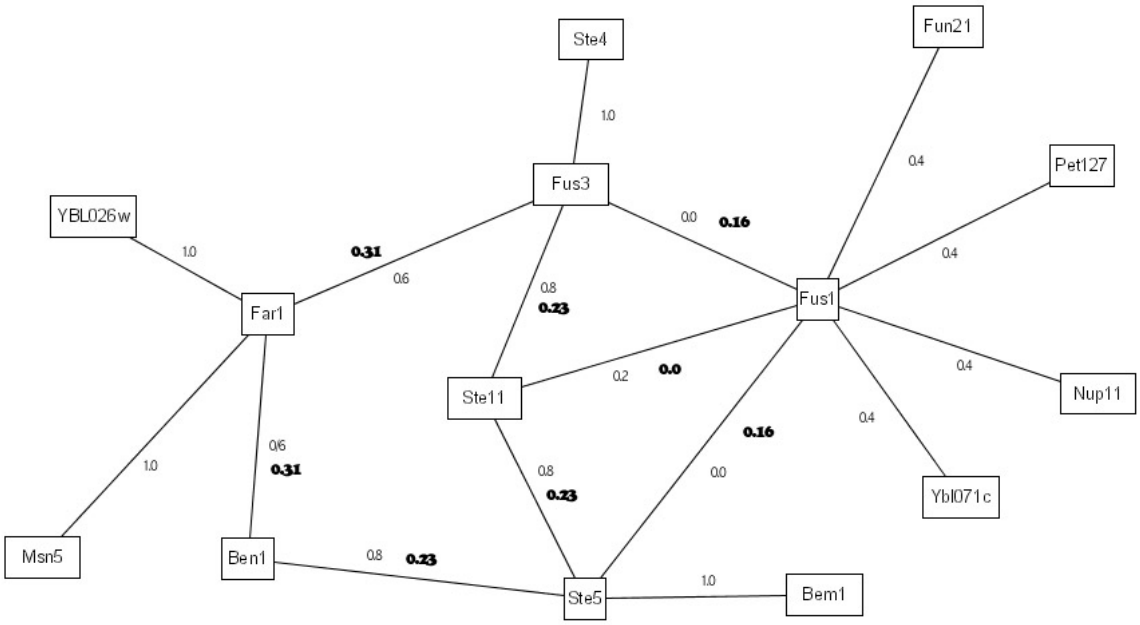

Figure 4.3: The subset of PPIs with IRAP (bold) and IG1 weight.

ranked as before. The bottom 1 interaction in the entire interactome is removed from the interactions. In our example, the interaction ⟨Fus1,Ste11⟩ belonged to the bottom spectrum and were removed from the network as false positives.

Next, IRAP* computes the confidence values for the missing interactions using a different initial weight that is based on common neighborhood counting. In the following example, there are a total of 3 potential false negatives. The table below shows the initial and final weights of these interactions.

Protein A	Protein B	initial weight	final weight
Fus3	Ste5	1.0	0.0625
Far1	Ste5	0.5	0
Fus3	Ben1	1.0	0

Table 4.1: 3 potential false negatives

Since we have removed 1 interaction from the network, we replaced it with the potential false negatives by, in our current work, inserting the top new interactions into the network. In our example, the top interaction in the above table belonged to the overall top interaction ⟨Fus3,Ste5⟩ and is thus added to the network.

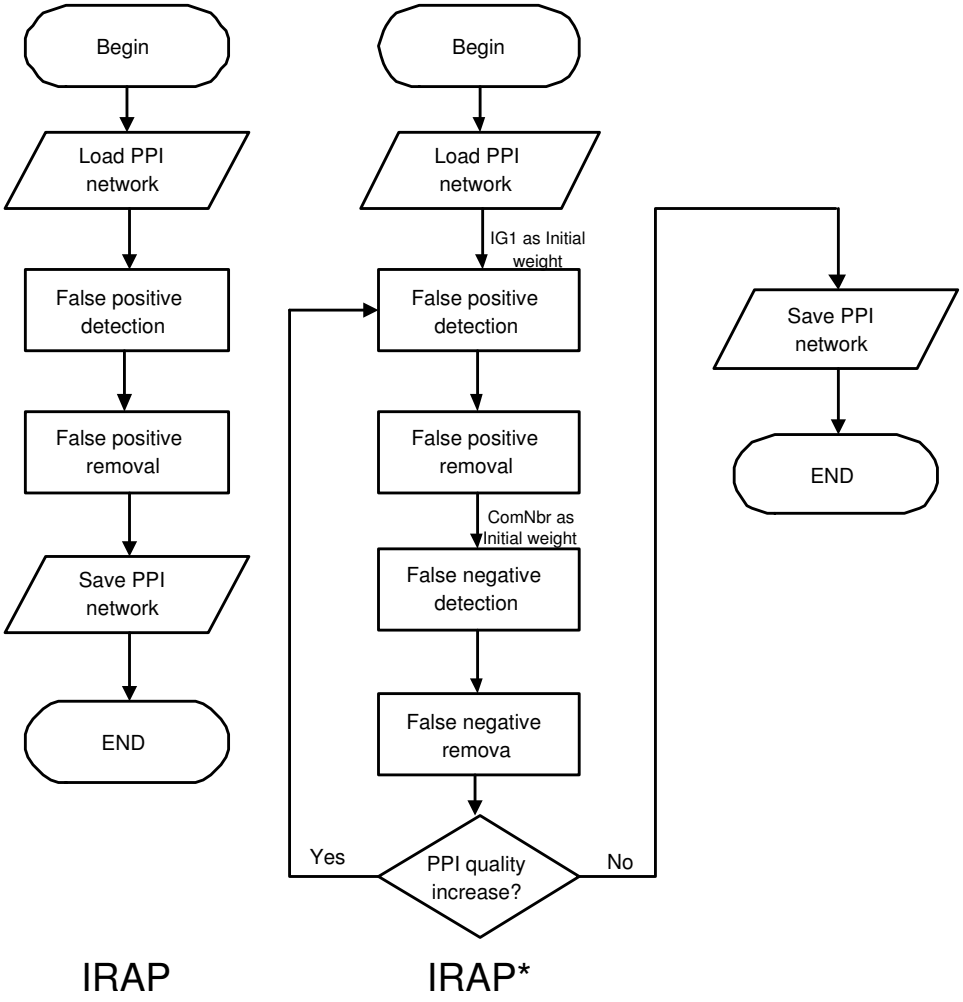

Figure 4.4: Flowcharts for IRAP and for IRAP*.

After 3 iterations of such false positive and false negative removals, we ended up with 14 interactions for the 14 proteins in our example, 3 of the original interactions were detected as false positives and hence removed from the repurified interactome, while 1 new interactions that were undetected by the Y2H screen were added to the final interactome. Our experimental evaluations reported in the manuscript showed biological evidence that the final interactome contains more high confidence interactions that the original interactome.

We will show, in the next section, that by iteratively removing false positives from the interactomes and replacing them with interactions identified as false negatives, the reliability of the resulting interactomes improves (based on functional

homogeneity) after each iteration.

4.4 Evaluation

We perform various evaluation experiments to ascertain that the application of IRAP* can improve the confidence of experimentally derived protein interactomes. First, we validate that IRAP is effective in identifying false positives and false negatives. After that, we show that the iterative refinement process in IRAP* leads to increasingly better interactomes in terms of functional homogeneity of the protein partners in the interactions.

4.4.1 Datasets

We perform our evaluation on experimentally derived interaction data for three different species, namely *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), and *Drosophila melanogaster* (fruit fly). We focus here only on interactomes that are derived by the popular high-throughput assays such as Y2H, which are well-known for their high experimental error rates. For *Saccharomyces cerevisiae*, the experimental Y2H-derived interactome was downloaded from the BIND database [GDC03] that comprises 7686 non-redundant Y2H interactions between 4141 of the yeast proteins. For *Caenorhabditis elegans*, we obtained an interaction network of 5025 non-redundant high-throughput PPIs between 2911 of the worm proteins from the BIND database. For *Drosophila melanogaster*, we obtained a much larger interaction network of 22437 non-redundant high-throughput PPIs between 7621 of the fly proteins from the BIND database. The three PPI datasets used in this work can also be found from our website.

We evaluate the quality of the interactomes by the degree of cellular functional homogeneity amongst the interacting protein pairs. Under the oft-quoted 'guilt-by-association' principle[Oli00], we would expect that as the rate of true positive interactions increases in the resulting interactomes processed by IRAP*, the proportion of interacting proteins with a common functional roles should also increase correspondingly.

4.4.2 False Positive Detection

The previous chapter reported on the performance of IRAP for false positive detection on a combined set of *Saccharomyces cerevisiae* interaction data that included interactions detected via both high-throughput and low-throughput methods. To better illustrate how IRAP can be a useful computational complement to current high-throughput experimental assays for PPIs, we report here the performance on the *Saccharomyces cerevisiae* interaction data that were derived *only* using high-throughput Y2H experimental assays.

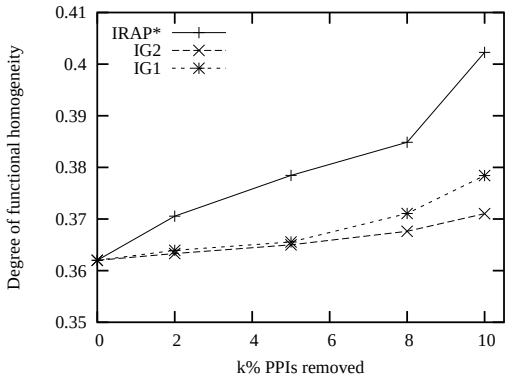

Figure 4.5: Degree of functional homogeneity increases at different rates as potential false positives are removed from the yeast interactome under different interaction reliability measures.

Recall that in each iteration of IRAP*, we seek to identify the most likely false positives amongst the worst $k\%$ of the interactions in terms of their IRAP values. We verify here that the interactions in the lower-end spectrum of the IRAP-indexed interactome indeed contained a larger proportion of biologically unlikely (i.e. functionally non-homogeneous) interactions than those indexed by other such topological metrics as IG1 and IG2. Figure 4.15 shows that IRAP is the best in detecting false positives—as more interactions that were detected as potential false positives were removed from the interactome, the degree of functional homogeneity in the resulting interactome increases at a faster rate than using other topological filtering metrics.

4.4.3 False Negative Detection

Similarly, for false negative detection, we verify whether the top new interactions proposed with IRAP exhibit a higher degree of functional homogeneity than those detected using other topological measures. Figure 4.6 shows that the new interactions proposed by IRAP were indeed of better quality than the corresponding sets proposed by IG1 and IG2.

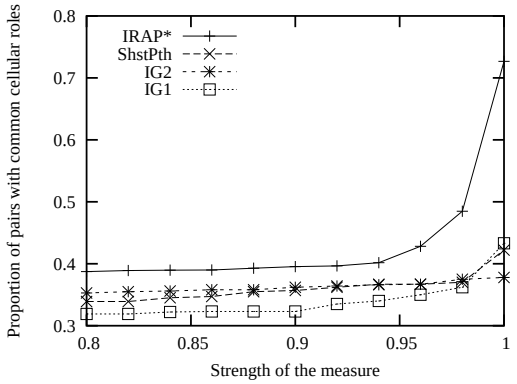

Figure 4.6: Different degrees of functional homogeneity in the various proportions of potential false negative PPIs to be added to the yeast interactome under different interaction reliability measures.

4.4.4 Iterative Refinement by IRAP*

Finally, we evaluate whether the iterative refinement process of IRAP* leads to incrementally improved interactomes. We perform two sets of experiments here. First, we verify whether true "gold standard" interactions are retained in the interactome over the iterations, while the hidden true gold standard interactions are rediscovered. Then, we verify whether the degree of functional homogeneity of the repurified interactomes consistently improves over iterations. For this, we will present the results for interactomes of different species.

For evaluation, we ran IRAP* with parameter k for sufficient iterations to repurify the PPI networks while keeping the number of interactions unchanged in the procedure. To determine the value for parameter k, we tested with different values of k on the *Saccharomyces cerevisiae* PPI network. Figure 4.7 shows the

maximal similarity scores for GO molecular function annotations within 15 iterations for $k = 1$ to 15. We use $k = 5$ since it gives the best overall similarity score.

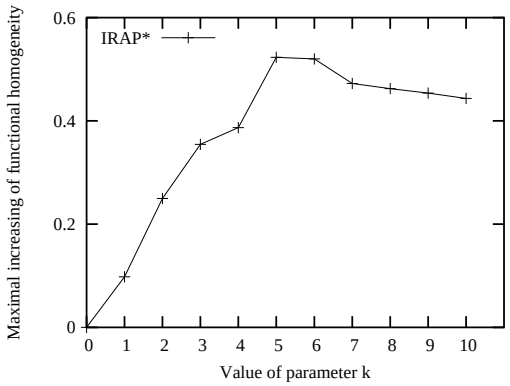

Figure 4.7: Maximal increasing of functional homology in 15 iterations on the *Saccharomyces cerevisiae* interactome varies with the parameter k.

As there are currently no comparable measures that can detect both false positives and false negatives, we compared IRAP* with a similar iterative refinement process, which we will call IG1+ComNbr, that uses only those measures that were used in IRAP* as initial weights, namely, IG1 for false positive detection[1] and ComNbr for false negative detection in each iteration. In other words, in each iteration of IG1+ComNbr, the initial weights of IG1 and ComNbr are directly used for filtering the false positives and false negatives without being further refined with IRAP. In this way, we show the effect of IRAP in the repurification process.

Persistence and Rediscovery

First, we are interested in finding out how well IRAP* retains true interactions and also how well it can rediscover those true interactions that were hidden from the experimentally detected interactome. We use the yeast PPIs in BIND that were indicated as having been obtained using low-throughput experimental methods as the set of "gold standard" interactions here. Out of 1,313 such low-throughput interactions found in BIND, 410 interactions were also found in our experimentally derived yeast interactome (using high throughput Y2H assays). We randomly hide

[1]We use IG1 instead of IG2 since it has comparable performance as IG2 for the Y2H dataset in the lowest 10% interactions (see Figure reffigure:low), and IG2 is much slower to compute.

50% of the 410 gold standard PPIs from the network, to see whether IRAP* could rediscover these lost interactions and retaining the other 50% in its refined interactome. This hiding and re-discovery process was repeated randomly for 100 times. Figure 4.8 shows the average results of the persistent and rediscovery rates for each iteration of IRAP*. In the final IRAP*-repurified interactome, an average of 139 out of 205 gold standard interactions were retained in the PPI network, while and 77 out of 205 hidden gold standard interactions were rediscovered. IG1+ComNbr performed significantly less well. After iterating for 10 times, only 102 gold standard interactions were still kept in the repurified interactomes, and a mere 21 hidden interactions were rediscovered. For reference, Figure 4.8 also shows the results of a baseline process that randomly added and removed PPIs from the network.

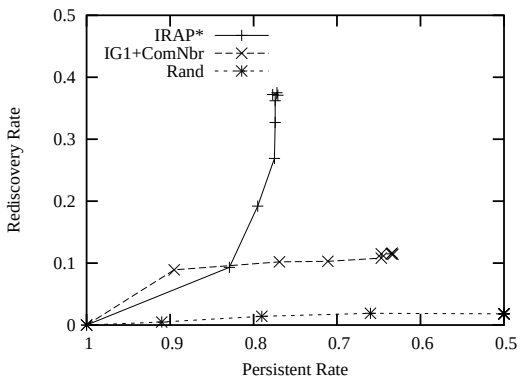

Figure 4.8: Persistent and rediscovered rates for IRAP*, IG1+ComNbr, and the baseline random process.

Functional Homogeneity

Finally, we check whether the quality in terms of functional homogeneity of the repurified interactomes improves over each iteration of IRAP*. To demonstrate IRAP*'s consistent performance across different kinds of interactomes, we show the results of IRAP* on three different species: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster* in Figures 4.9, 4.10, and 4.11 respectively. For reference, we also show the degree of functional homogeneity of the "gold standard" yeast interactions used in section 4.4.4 in Figure 4.9.

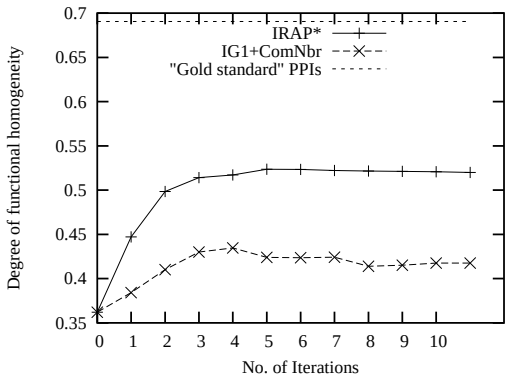

Figure 4.9: PPI similarity score based on enriched GO terms increases at different rates with IRAP* and IG1+ComNbr on the *Saccharomyces cerevisiae* interactome.

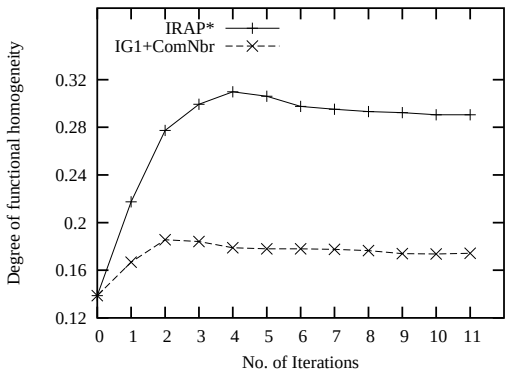

Figure 4.10: PPI similarity score based on enriched GO terms increases at different rates with IRAP* and IG1+ComNbr on the *Caenorhabditis elegans* interactome.

Figure 4.9 shows that with IRAP*, the average functional homogeneity score of the interacting *Saccharomyces cerevisiae* proteins increases from the original 0.362 to 0.523 whereas with IG1+ComNbr, the similarity score in its repurified yeast interactome increases to only 0.435. For the sparser *Caenorhabditis elegans* interactome, the improvement with IRAP* was 0.139 to 0.310 while IG1+ComNbr obtained an increase to only 0.189. As for the much larger *Drosophila melanogaster* interactome, IRAP* improved the degree of functional homogeneity quality from 0.123 to 0.206 whereas IG1+ComNbr could only increase it to 0.158. These results show that IRAP* is effective as a computational complement for the repurification of various experimentally-derived interactomes with different densities and sizes. Note also

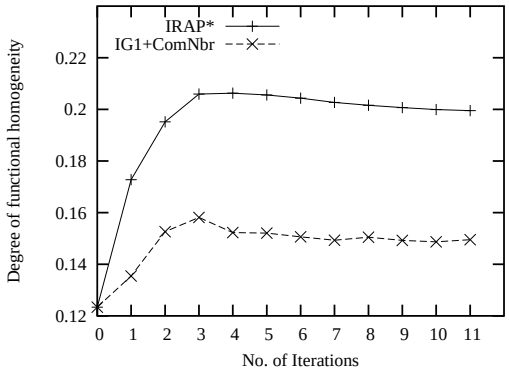

Figure 4.11: PPI similarity score based on enriched GO terms increases at different rates with IRAP* and IG1+ComNbr on the *Drosophila melanogaster* interactome.

that in all cases, only a small number of iterations was required since the quality of the repurified interactomes tend to converge in less than 10 iterations.

4.4.5 Cross-talkers

Thus far, we have evaluated the repurified interactomes based on the degree of functional homogeneity of the interactions. In the other related works[SSH02b, SSH02a, PZ05], the degree of cellular co-localization in the interactions was also used as a quality measure. However, as there are many important biological interactions (such as those in signaling pathways) that occur across cellular localizations, we have chosen not to use cellular co-localization in our evaluation.

In fact, we found that a number of the PPIs in our repurified interactomes that have been assigned high confidence values actually involved protein partners from different cellular localizations. Figure 4.12 shows that the degree of co-localization in the protein pairs detected by IRAP* decreases in each iteration until about 7-8 iterations. Under the cellular co-localization evaluation measure, these interactions would have been considered as errors. However, on inspecting the corresponding cellular functional roles of the non co-localizing protein interaction partners, we found that many of them actually shared the same function. For example, 32% of the non co-localized pairs discovered by IRAP* involved functionally homogeneous protein partners. These "cross-talkers" are thus likely to be actual biological interactions

in pathways that occur across cellular sublocalizations. Figure 4.13 shows some examples of cross-talkers discovered by IRAP* in *Saccharomyces cerevisiae* PPIs.

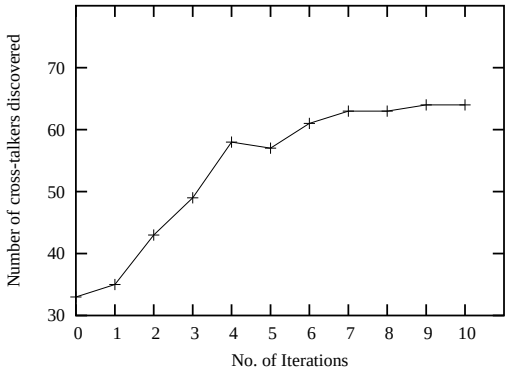

Figure 4.12: Degree of co-localization decreases in each iteration.

Protein A	Protein B	Com. Function	Localization A	Localization B
MCM1	CDC43	mitotic cell cycle and cell cycle control	nucleus	cytoplasm
FIG1	GPA1	pheromone response, mating-type determination, sex-specific proteins	integral membrane / endomembranes	plasma membrane
STE24	SLA2	incompatibility	mitochondrial matrix	actin cytoskeleton
END3	YCK2	cytokinesis (cell division) /septum formation	cytoskeleton	plasma membrane
MCM10	PRT1	mitotic cell cycle and cell cycle control	nucleus	cytoplasm
TPS2	SED1	stress response	cytoplasm	plasma membrane

Figure 4.13: Examples of interactions between non co-localized proteins ("cross-talkers") that are involved in the same cellular pathways as discovered by IRAP*.

4.4.6 IRAP* v.s. IG1/2 in each iteration

We do the experiment to compare IRAP with IG1/2 with the enriched GO term comparison method instead of the exact GO term matching. By considering the GO term relationships, IRAP outperforms IG1 and IG2.

We also tested the top 10% PPIs in the first 5 iterations of the repurification process (we select 5 because IRAP* to over-fit after 5 iterations). Results are shown in Figure 4.14. The "increase of functional homology" refers to the increase of similarity score based on the enriched GO term in the whole network after removing the bottom 10% interactions according to IRAP* or IG1/IG2. We see that IRAP* is consistently better than IG1 and IG2 in detecting false positives.

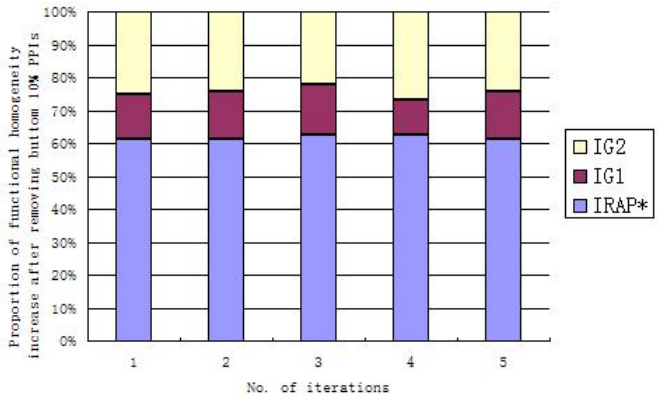

Figure 4.14: The increase of the degree of cellular functional homogeneity in the first 5 iterations at different rates as the bottom 10% protein interactions are removed from the yeast interactome under different interaction reliability measures.

4.4.7 False Positive Detection by IRAP* v.s. PathRatio

A new path-based measure, PathRatio, has been proposed by Pei and Zhang[PZ05] recently as an alternative to our IRAP measure. Basically, PathRatio is a topological measure to select reliable interactions by all the length specified paths connecting the two target proteins. PathRatio was shown to perform better than our IRAP in the top spectrum of the indexed interactome, suggesting that PathRatio was better in detecting true positives. Note that the PathRatio results reported in [PZ05] was obtained using a different set of initial weights from the IG1 values used in IRAP; as such, it is unclear whether the difference in performance was solely due to the relative performance of the best-path approach adopted by IRAP versus the all-path approach adopted by PathRatio.

However, in our computational repurification application, the performance of detecting false positives is more important than detecting true positives since we are removing a small portion of interactions that have to be confidently deemed as false positives in each iteration. In this aspect, IRAP actually performed comparably if not slightly better than PathRatio. In addition, unlike our IRAP which adopts an efficient best-path approach, PathRatio uses an all-paths approach which is therefore computationally much more intensive.

Figure 4.15 shows the superiority of IRAP as a false positive filtering measure: as the IRAP threshold is increased, the enriched GO term similarity increases from 0.362 to 0.402, indicating an increased rate of true positives in the filtered interaction data. For comparison, we also show the performance of IG1, IG2 and PathRatio in the figure. With PathRatio, the enriched GO term similarity increases from 0.362 to 0.392, with IG2, the score only increases to 0.378; and with IG1, the proportion only increases to 0.371.

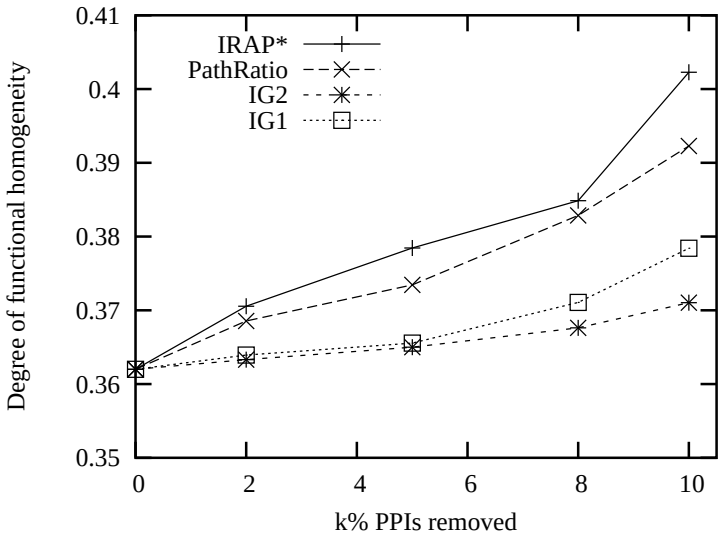

Figure 4.15: Degree of functional homogeneity increases at different rates as potential false positives are removed from the yeast interactome under different interaction reliability measures.

4.5 Conclusions

Much of the recent technological advance has been focused on the high throughput detection of PPIs in order to map the tremendously vast protein interactome. Unfortunately, the potential benefits of these technological advances cannot be fully realized as the PPI data generated using the high throughput technologies have very high error rates. In this work, we have proposed a novel computational complement for the repurification of the experimentally-derived interactomes. We iteratively refine an interactome by removing interactions that are identified as false positives

and adding interactions detected as false negatives into the interactome. The computationally repurified interaction data sets were shown to contain potentially lower fractions of false positive and false negative errors. Additionally, biologically interesting interactions such as cross-talkers may also be discovered using our method.

Note that in this work, the detection of the potential experimental errors was intentionally done using only the topological information that were mathematically derived from the underlying interaction graphs. This is to allow us to clearly illustrate the potential usefulness of such a topological approach. In practice, other biological information (e.g. gene expression correlation and even the functional and co-localization information that were used for evaluation here) should be incorporated to improve the quality of the repurified interactomes. As future work, it will be thus interesting to investigate how other topological measures, as well as additional biological information, can be incorporated for interactome repurification in practice.

CHAPTER 5

Network Motif Discovery

It is not necessary for a true interaction to have at least one alternative path. The strategy of the alternative path may lose some interactions which are true positives but do not have alternative paths. For example, in the yeast PPI network, about 14% PPI pairs do not have at least one alternative paths [CHLN05b]. As a consequence, many true positives will be wrongly remove by the methods IRAP or IRAP* introduced in the previous two chapters.

Recent works in network analysis have revealed the existence of network motifs in biological networks such as the protein-protein interaction (PPI) networks. Such topological patterns do not limited the existence of alternative paths. Hence the recall to use network motif to validate protein interactions may be higher than the alternative path related approaches.

However, existing motif mining algorithms are not sufficiently scalable to find meso-scale network motifs. Also, there has been little or no work to systematically exploit the extracted network motifs for dissecting the vast interactomes.

We describe an efficient network motif discovery algorithm, NeMoFinder, that can mine meso-scale network motifs that are repeated and unique in large PPI networks. Using NeMoFinder, we successfully discovered, for the first time, up to size-12 network motifs in a large whole-genome *S. cerevisiae* (Yeast) PPI network.

We also show that such network motifs can be systematically exploited for indexing the reliability of PPI data that were generated via highly erroneous high-throughput experimental methods.

5.1 Introduction

Recent works in network analysis [MSOI+02] have revealed that the topology of complex natural networks such as protein-protein interaction (PPI) networks are far from random. Many of these networks have been shown to exhibit such common global topological features as the "small world" and "scale free" properties. It turns out that in addition to these global topological characteristics, many local topological patterns can also be detected in the large complex natural networks. For example, Milo *et al.*[MSOI+02] discovered various significant patterns of local connections that occurred more frequently in complex networks than in randomized networks. They called these recurring local topological substructures as "network motifs". While relatively less widely studied than the global topological features, such network motifs can lead to better understanding about various classes of complex networks, as some network motifs may be particular to specific classes of networks. For example, certain triad or tetrad motifs are specific topological patterns that are found to appear in biological networks rather than in other networks [MSOI+02]. The presence of such network motifs also reveals the basic structural elements that underlie the hierarchical and modular architecture of such complex natural networks as PPI networks.

Researchers have only recently begun to employ network motifs in exploring the interactomes; for example, Saito *et al.* [SSH02b, SSH02a] used manually derived network motifs to detect false positives in highly erroneous PPI networks, while Albert *et al.*[AA04] used them to predict PPIs. These pioneering works have achieved promising results even though the network motifs used in these works were rather limited—Saito *et al.* had used only 5 predefined network motifs of size 3 in their latest work on false positive detection [SSH02a], while Albert *et al.* had used only 4 predefined small network motifs for predicting interactions. This shows that the

network motifs can provide a framework for the effective dissection of the complex PPI network based on the underlying structural principles.

As many of the relevant processes in biological networks have been shown to correspond to the meso-scale (5-25 genes or proteins) [SM03], it would be interesting to investigate if it is advantageous to use network motifs that are of equivalent sizes. However, existing network motif discovery algorithms [MSOI+02, KIMA04] are not applicable as they are mostly enumeration-based and limited to extracting smaller network motifs (up to size 8) for the following reasons:

1. The number of network motifs candidates increases exponentially with the motif size [IWM00, KK04a].

2. Interesting network motifs are typically repeated and unique [MSOI+02], that is, the motifs occur repeatedly in the PPI network but not in the randomized networks. Such motifs do not have the downward closure property and *Apriori* algorithms are not applicable here.

3. The graph isomorphism problem, which is the essential technique to identify different network motifs, is an NP problem [For96].

Such limitations impact the applicability of motif discovery approach for biological applications, as meso-scale network motifs are beyond the reach of existing exhaustive enumeration algorithms.

In this chapter, we present an efficient graph mining algorithm called *NeMoFinder* [CHLN06b] to discover meso-scale repeated and unique network motifs in a large, genome-scale PPI network for biological applications. The proposed algorithm utilizes repeated trees to partition a network into a set of graphs. We introduce the notion of graph cousins to facilitate the candidate generation and frequency counting processes. Experiment results indicate that NeMoFinder is scalable and outperforms existing network motif discovery methods. We also use the network motifs that are mined from the real-life biological networks to detect false positives in the highly erroneous PPI network obtained from biological experimental methods. The experimental results demonstrate that the actual meso-scale network motifs extracted

from the biological interaction networks can achieve better performance than using small, predefined ones for assessing the reliability of PPIs from conventional high-throughput experiments.

The rest of this chapter is organized as follows. Section 5.2 introduces the basic concepts. Section 5.3 describes the related work in network motif algorithms. In Section 5.4, we describe the proposed NeMoFinder algorithm. Section 5.5 presents the comparative results of using NeMoFinder for discovering network motifs for *S. cerevisiae* PPI networks. In Section 5.6, we show how the extracted network motifs can be used to validate the interactions in a PPI network. Finally, we conclude in Section 5.7.

5.2 Definitions

In this work, we model a PPI network as an undirected graph $G = (V, E)$ where each vertex in V represents a unique protein, and each edge in E between two vertices v_A and v_B indicates that there is an interaction detected between the corresponding proteins A and B. We exclude self-loops from G here, as we are only interested to see the effectiveness of graph topology between proteins (see section 5.6).

By definition, a network motif is a frequently occurring subgraph pattern in a network (in this case, a large genome-wide PPI network such as the Yeast PPI network that consists of 4341 vertices and 10199 edges). The class of network motifs that we are interested in extracting from the interactomes are unique non-random subgraphs [MSOI$^+$02] that occur repeatedly in the underlying biological network.

Let f_g be the number of occurrences of a subgraph g in a PPI network G. We say that g is *repeated* if f_g is more than some user-specified value F.

Let $f_{g\_rand_i}$ be the frequency of g in a randomized network G_{rand_i}, for $1 \leq i \leq N$, where N is the number of the randomized networks. Let s_g be the number of times f_g is equal or greater than $f_{g\_rand_i}$, $1 \leq i \leq N$, over the total number of randomized networks N. We say that g is *unique* if its s_g is more than some user specified value S.

Definition 4. Network Motif. *A network motif g in a PPI network G is a connected, unlabelled and undirected topological pattern of inter-connections that is* repeated *and* unique *in G.*

Note that it is common for proteins and their interactions in complex biological networks such as PPI networks to participate in multiple biological functional modules. It is therefore perfectly possible for multiple vertex- or edge-overlapping subgraphs to be simultaneously active at any time. However, PPI networks contain many highly dense clusters and high percentage of false positives. Therefore, it is preferable to count every occurrence differing in at least one vertex, allowing arbitrary overlaps in edges and vertices. Hence, during the subgraph counting process, we must consider patterns with arbitrary overlaps of vertices and edges, and at least one vertex difference.

This results in a computationally more complex problem as the frequency of network motifs does not have the downward-closed property in this case.

In addition, the uniqueness property of a network motif is also not downward-closed as a result of allowing vertex- and edge-overlap in the network motifs. When a motif g extends (or reduces) to its supergraph (or subgraph), the decrease (or increase) of f_g and f_{g_rand} is non-deterministic. This means that given a network motif g, we cannot directly infer whether the supergraphs and subgraphs of g are unique. In fact, even when we have found a non-unique motif, we still have to generate its supergraphs and check for their frequencies and uniqueness. This implies that determining the uniqueness value of a motif is also computationally expensive.

5.3 Related Work

In terms of biological network motif mining, the pioneering work by Milo *et al.* employed an exhaustive search algorithm that counts all the subgraphs of a given number of vertices. As such, they could only discover small network motifs in the form of 3-vertex and 4-vertex subgraphs. Kashtan *et al.*[KIMA04] developed a more efficient sampling method to estimate the relative frequencies of subgraphs. Their

method was useful for analyzing very large networks and for detecting high-order motifs since the runtime is independent of the network size. However, the sampling approach cannot be guaranteed to discover the complete set of network motifs. It also does not scale for large-size network motifs (the algorithm takes about 2 hours to find a size-8 motif in the network of transcriptional regulation of *E. coli* with 423 vertices and 519 edges).

On the other hand, the computationally savvy graph mining community has also been diligent in developing various algorithms to efficiently discover frequent subgraphs. The initial algorithms, notably the AGM [IWM00] and FSG [KK04a], were devised to find all the frequent subgraphs in a large graph database efficiently through the extension of the market basket analysis. The algorithms utilize the *Apriori* property to discover frequent subgraphs level by level. The gSpan [YH02] algorithm discovers frequent substructures by using a DFS-based canonical representation of graphs and enumerated the search space in a depth-first order. The FFSM [HWP03] method improves the performance of gSpan by reducing redundant subgraph candidates through a vertical search scheme with join and extension operations. Finally, the SPIN [HWPY04] algorithm overcomes the problem of cycles in graph by generating the frequent substructures hierarchically in two steps: starting from trees, and then extending the frequent trees to graphs.

All the above works have focused on mining subgraphs from a *collection* of graphs, and considered only the frequency but not the uniqueness property of subgraphs. Furthermore, in these works, the frequency of a subgraph is determined by the number of global graphs that the subgraph occurs in, regardless of whether the subgraph occurs many times within a particular graph. This is computationally easier than the network motif discovery problem where the frequency of a motif is determined by the number of occurrences, including vertex- and edge-sharing ones, within one large and complex graph.

Kuramochi *et al.*[KK04b] designed two methods hSigGram and vSigGram to look for frequent subgraphs in a sparse graph. These methods first determine the number of edge-disjoint occurrences of a subgraph based on approximate and exact maximum independent set computations and then use it to prune infrequent

subgraphs. However, the methods are not suitable for biological applications where a protein or an interaction can participate in multiple functional modules, in other words, the occurrences of a motif can overlap arbitrarily in a graph, which is a much more computationally challenging counting problem.

The FPF method by Schreiber *et al.*[SS04] extends hSigGram and vSigGram to find frequent subgraphs with arbitrary overlap. FPF uses the concepts of pattern tree and generating parent to prune redundant subgraph candidate generation. However, the method is expensive as it has to perform subgraph isomorphism test for all candidates. Furthermore, it is unable to prune the non-promising subgraphs as the frequency counting does not satisfy the downward closed property.

5.4 NeMoFinder: Network Motif Discovery Algorithm

In this work, we propose a network motif discovery algorithm called NeMoFinder to discover repeated and unique meso-scale network motifs in a large PPI network (Algorithm 4). The algorithm utilizes repeated trees to partition a network into a set of graphs, then use graph cousins (discuss later) for efficient candidate generation and frequency counting.

The input to the algorithm is a PPI network G, a user defined frequency threshold F, a user defined uniqueness threshold S, and a user defined maximal network motif size K. The output of the algorithm is a set U of repeated and unique motifs from size 3 to size K. Note that a subgraph with k vertices is said to be a size-k subgraph. The proposed algorithm consists of three main steps. First, we find repeated subgraphs in the PPI network (Lines 4-15). Then we check the frequency of the repeated subgraphs in the randomized networks (Lines 16-21). Finally, we determine the uniqueness values of the the repeated subgraphs (Lines 22-28).

We illustrate the algorithm using the example graph G in Figure 5.1. Suppose we want to find all the motifs up to size 5 (*i.e.,* $K = 5$) from G. We let the frequency

threshold $F = 2$, and the uniqueness threshold $S = 0.95$.

Algorithm 4 NeMoFinder

Input: G - PPI network;
N - Number of randomized networks;
K - Maximal network motif size;
F - Frequency threshold;
S - Uniqueness threshold;
Output: U - Repeated and unique network motif set;
$D \leftarrow \emptyset$;
for motif-size k from 3 to K **do**
  $T \leftarrow FindRepeatedTrees(k)$;
  $GD_k \leftarrow GraphPartition(G, T)$
  $D \leftarrow D \cup T$;
  $D' \leftarrow T$;
  $i \leftarrow k$;
  while $D' \neq \emptyset$ and $i \leq k \times (k-1)/2$ **do**
    $D' \leftarrow FindRepeatedGraphs(k, i, D')$;
    $D \leftarrow D \cup D'$;
    $i \leftarrow i + 1$;
  end while
end for
for counter i from 1 to N **do**
  $G_{rand} \leftarrow RandomizedNetworkGeneration()$;
  for each $g \in D$ **do**
    $GetRandFrequency(g, G_{rand})$;
  end for
end for
$U \leftarrow \emptyset$;
for each $g \in D$ **do**
  $s \leftarrow GetUniqunessValue(g)$;
  if $s \geq S$ **then**
    $U \leftarrow U \cup \{g\}$;
  end if
end for
return U;

Step 1. Discover Repeated Subgraphs.

The discovery of repeated size-k subgraphs in a PPI network, $2 < k \leq K$, involves the following three steps:

Step 1.1 Find Repeated Size-k Trees.

Algorithm NeMoFinder starts by finding the size-2 tree t_2 in G. Then the

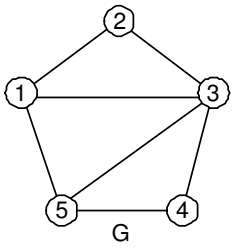

Figure 5.1: Example graph G.

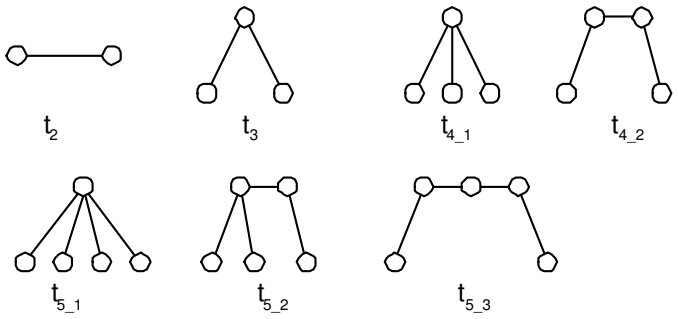

Figure 5.2: Size 2 to size 5 trees.

algorithm extends t_2 to a size-3 tree, size-4 trees, *etc.,* until size-K trees are obtained. Figure 5.2 shows all the size-2 to size-5 trees. Note that we have two size-4 trees (t_{4_1}, t_{4_2}) and three size-5 trees (t_{5_1}, t_{5_2}, t_{5_3}).

When a size-k tree t_k is formed, NeMoFinder counts its occurrences in G. If the occurrences of tree t_k is more than the user given threshold, then t_k is a repeated tree, and it is added to the set T_k.

In our example, the occurrences/frequencies of the various size trees are as follows: $f_{t_2} = 7$, $f_{t_3} = 13$, $f_{t_{4\_1}} = 6$, $f_{t_{4\_2}} = 17$, $f_{t_{5\_1}} = 1$, $f_{t_{5\_2}} = 5$, $f_{t_{5\_3}} = 7$. All frequency values except for the frequency of t_{5_1} are more than the user given threshold of 2. Thus we have $T_2 = \{t_2\}$, $T_3 = \{t_3\}$, $T_4 = \{t_{4\_1}, t_{4\_2}\}$ and $T_5 = \{t_{5\_2}, t_{5\_3}\}$.

Step 1.2 Use Repeated Size-k Trees to Partition Graph.

Next, we use the size-k trees in T_k to partition the graph G into a set of graphs GD_k such that each graph G_{k_j} in GD_k embeds a size-k tree in T_k, $2 \leq k \leq K$ and

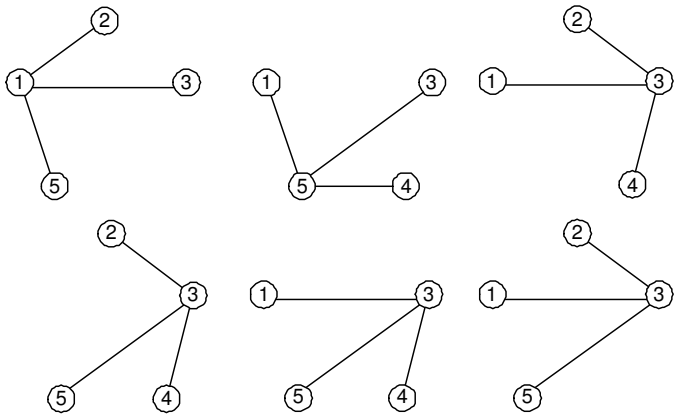

Figure 5.3: Occurrences of t_{4_1} in G.

$1 \leq j \leq |GD_k|$.

Consider the trees t_{4_1} and t_{4_2} in Figure 5.2. Figure 5.3 and 5.4 shows the occurrences of t_{4_1} and t_{4_2} in G. We use t_{4_1} and t_{4_2} to partition the PPI network G to obtain the set of graphs $GD_4 = \{G_{4\_1}, G_{4\_2}, G_{4\_3}, G_{4\_4}G_{4\_5}\}$ (Figure 5.5). Note that each graph in GD_4 embeds the tree t_{4_1} and/or t_{4_2}.

Step 1.3 Perform graph join operation to find repeated size-k graphs.

For each tree t in T_k, we generate size-k subgraphs with $k-1$ edges (the rules for generating the subgraphs are given in Section 5.4.1). Then we join t with each of these subgraphs to generate size-k subgraphs with k edges. The latter are added to the candidate set C_k.

Figure 5.6 shows the 4-vertex 3-edge subgraphs, $h_1, \ldots, h_5$, generated from the two size-4 trees t_{4_1} and t_{4_2} in T_4. We join t_{4_1} with h_1 and h_2, and join t_{4_1} with h_3, h_4 and h_5 to generate 4-vertex 4-edge subgraphs. Figure 5.7 shows the subgraphs obtained after joining t_{4_1} with h_1, and t_{4_2} with h_3. The non-redundant subgraphs g_{1_1} and g_{1_2} are added into the candidate set C_4.

For each subgraph $g \in C_k$, we check its occurrences in GD_k. If the occurrences of g is more than the threshold F, we add g to the set D_k. In our example, $f_{g_{1\_1}} = 2$ and $f_{g_{1\_2}} = 5$. Thus, g_{1_2} is a repeated subgraph and is added to the set of frequent subgraphs D.

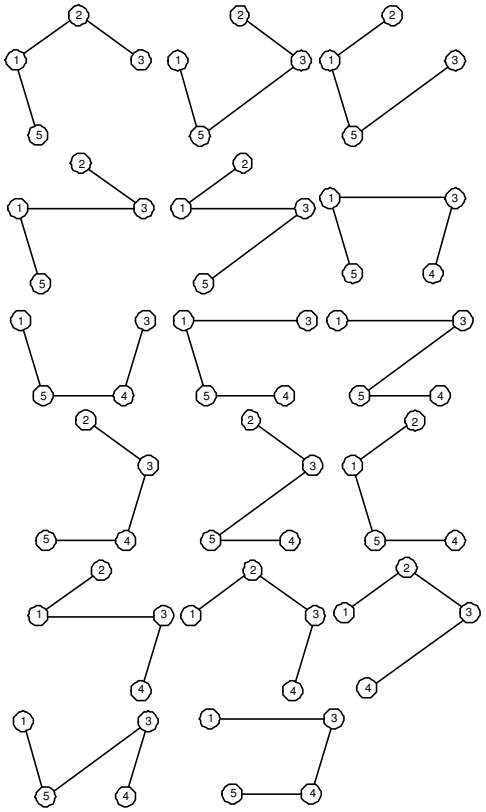

Figure 5.4: Occurrences of t_{4_2} in G.

Next, we use the repeated subgraphs obtained to generate all the possible k-vertex and k-edge subgraphs. These repeated subgraphs are joined with the newly generated subgraphs to get $(k+1)$-edge subgraphs. The repeated $(k+1)$-edge subgraphs are added to D. This process continues until a complete graph of $k*(k-1)/2$ edges is obtained, or no repeated subgraph can be found.

Figure 5.8 shows the 4-vertex and 4-edge subgraphs, h_6 and h_7, generated from the repeated subgraph g_{1_2}. We join g_{1_2} with h_6 and h_7 to get a 4-vertex and 5-edge subgraph g_2 (see Figure 5.9). Since the frequency of g_2 in GD_4 is not more than 2, it is not a repeated subgraph and the algorithm stops.

At the end of Step 1, the algorithm outputs the set D which contains all the repeated trees and subgraphs from size-2 to size-K.

Step 2. Determine Subgraph Frequency in Randomized Networks.

Next, we use the Markov-chain algorithm [MS02] to generate randomized

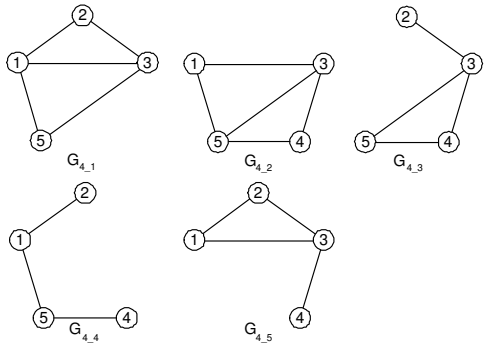

Figure 5.5: Set of graphs GD_4; each graph in GD_4 embeds t_{4_1} and/or t_{4_2}.

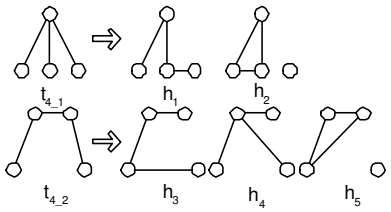

Figure 5.6: Generate 3-edge subgraphs from size-4 trees.

networks G_{rand_i} $(1 \leq i \leq N)$ by swapping randomly selected interactions, as was done in [MSOI+02]. This ensures that the randomized networks have the same single-vertex characteristics as the PPI network, *i.e.,* each vertex in the randomized networks has the same number of neighbors as the corresponding vertex in the PPI network. We check the frequency of the subgraphs in D in each of the randomized networks G_{rand_i} $(1 \leq i \leq N)$. The procedure is similar to Step 1.

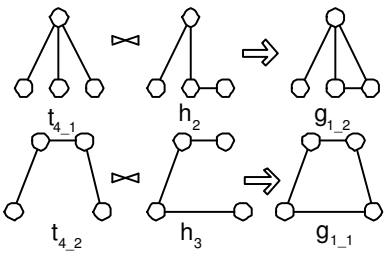

Figure 5.7: Examples of graph join operations for 3-edge subgraphs.

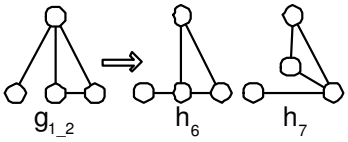

Figure 5.8: Generate 4-edge subgraphs from repeated 4-edge subgraphs of G.

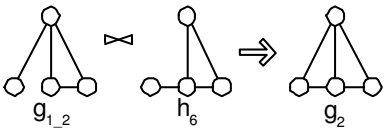

Figure 5.9: Examples of graph join operations for 4-edge subgraphs.

Step 3. Compute Uniqueness of Subgraphs.

Finally, we compute the uniqueness value for each subgraph in D based on its frequencies in the input PPI network and the randomized networks.

NeMoFinder is scalable because the repeated trees naturally partitions the network into a set of graphs GD. Hence, the problem of counting the frequency of a size-k subgraph g in the network is reduced to the problem of finding the number of graphs in GD that contain the subgraph g, which is naturally downward closed.

In order to reduce the computational complexity, NeMoFinder adopts the idea in SPIN [HWPY04] to search for repeated trees and then extend them to subgraphs. However, NeMoFinder differs from SPIN in the following:

1. The notion of frequency in SPIN is different from our NeMoFinder. SPIN simply checks whether a subgraph occurs in a graph; it is not interested in counting how many times the subgraph occurs in the graph. In contrast, NeMoFinder considers occurrences of a subgraph in a network, including arbitrary overlaps.

2. SPIN uses equivalence classes to find maximal labelled frequent subgraphs in a set of graphs. In contrast, NeMoFinder is focused on discovering repeated unlabelled subgraphs from a single graph. Hence, our NeMoFinder is able to utilize the symmetry property of unlabelled trees to further reduce the number of candidate trees enumerated.

5.4.1 Candidate Generation using Graph Cousins

Finding repeated subgraphs involves generating candidate subgraphs and frequency counting (see Algorithm 5). The standard method to generate a subgraph candidate g_k from a tree t_k is to add a new edge to t_k and check whether the resulting graph is already in the candidate set C_k. However, C_k can become very large for meso-scale subgraphs, and checking whether a graph exists in C_k requires graph isomorphism test which is a NP problem.

Given that the network motifs are meso-scale, we use adjacency matrices to represent the subgraphs so as to facilitate the graph join operation to generate candidate subgraphs. A graph g with n vertices can be modelled using a $n \times n$ matrix M. An entry $m_{i,j}$ in an adjacency matrix is set to 1 if there is an edge from vertex i to j, and 0 otherwise. The code of M, denoted as $code(M)$, is a sequence formed by linking the lower triangular entries of M in the following order: $m_{1,1}m_{2,1}m_{2,2}...m_(i,j)...m_{n,1}m_{n,2}...m_{n,n}$ where $(0 < j \leq i \leq n)$.

We can transform any adjacency matrix into a unique representation called canonical adjacency matrix (CAM) [For96]. Then two subgraphs that are isomorphic to each other have the same CAM, and vice versa. The canonical adjacency matrix (CAM) of a subgraph g, denoted as $CAM(g)$, is the adjacency matrix of g with the maximal code. The last edge entry of $CAM(g)$ is the rightmost non-zero edge entry in $code(CAM(g))$. By removing the edge which corresponds to the last edge entry of $CAM(g)$, we obtain a subgraph of g. We call the adjacency matrix of such a subgraph as $subCAM(g)$ defined as follows:

Definition 5. subCAM of a graph. *Let $CAM(g)$ be canonical adjacency matrix of a graph g. Then $subCAM(g)$ is a matrix obtained by setting the last edge entry in $CAM(g)$ to 0.*

Given two subgraphs g and h, if $subCAM(g) = subCAM(h)$, then we say that h is a *cousin* of g. There are three types of cousin relationship between g and h:

- **Type I: Direct Cousin** h is isomorphic to a subgraph g' which has the same number of vertices and edges as g, and $g \neq g'$;

- **Type II: Twin Cousin** h is isomorphic to subgraph g;
- **Type III: Distant Cousin** h is a disconnected subgraph.

Figure 5.10 shows the adjacency matrices for the size-4 trees t_{4_1} and t_{4_2} and the generated subgraphs $h_1, \ldots, h_5$ in Figure 5.6. From the above definitions, we see that h_1 is a Type I direct cousin of t_{4_1} since it is isomorphic to t_{4_2}; h_2 is a Type III distant cousin of t_{4_1} since it is a disconnected subgraph; h_3 is a Type II twin cousin of t_{4_2} since it is isomorphic to t_{4_2}; h_4 is a Type I direct cousin of t_{4_2} since it is isomorphic to t_{4_1}; h_5 is a Type III direct cousin of t_{4_2} since it is a disconnected subgraph.

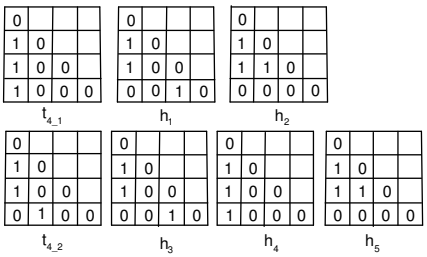

Figure 5.10: Adjacency matrices for the graphs in Figure 5.6.

We now show how the subgraph generation and frequency counting are efficiently carried out based on the cousins of a graph.

Given a repeated subgraph g of size k, we first find its set of cousins, H. Then we join g with each graph $h \in H$ to form new subgraphs of size k that have one more edge than g. Let $CAM(g)$ be CAM of g and $CAM(h)$ be CAM of h, then the adjacency matrix M of the new subgraph candidate is a $k \times k$ matrix and

$$m_{i,j} = \begin{cases} 1 & \text{if } CAM(g)_{i,j} = 1 \text{ or } CAM(h)_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

Algorithm 6 gives the pseudo code for the candidate generation procedure.

The following theorem proves that the join operation generates the complete set of candidate subgraphs.

Algorithm 5 FindRepeatedGraphs(k, i, D')

Input: D' - Set of repeated subgraphs with k vertices and $i-1$ edges;
Output: D'' - Set of repeated subgraphs with k vertices and i edges;
$C \leftarrow CandidateGeneration(k, i, D')$;
$D'' \leftarrow FrequencyCounting(k, i, C)$;
return D'';

Algorithm 6 CandidateGeneration(k, i, D')

Input: D' - Set of repeated subgraphs with k vertices and $i-1$ edges;
Output: C - Set of candidates with k vertices and i edges;
$C \leftarrow \emptyset$;
for each $g \in D'$ **do**
  $H \leftarrow GetCousin(g)$;
  for each $h \in H$ **do**
    $g' \leftarrow join(g, h)$;
    $C \leftarrow Can \cup \{g'\}$;
  end for
end for
return C;

Theorem 5.4.1. *Given all the subgraphs $g \in C_k$ which has k vertices and l edges ($l \geq k-1$), the join operation generates the complete set of subgraphs C'_k, where each $g \in C'_k$ has k vertices and $l+1$ edges.*

Proof: Let M be an adjacency matrix of a subgraph $g \in C'_k$ and e_1 be the last edge entry in M, such that matrix $M_1 = M - \{e_1\}$ is a CAM of a subgraph g_1. Let e_2 be the last edge entry in M_1. Since M_1 is a connected graph, its corresponding subgraph g_1 must be in C_k.

Let $M_2 = M_1 - \{e_2\} + \{e_1\}$ and M_2 be an adjacency matrix of a subgraph g_2, we have $g_1 \bowtie g_2 \Rightarrow g$. Based on the definition of graph cousins, if g_2 is isomorphic to g_1, g_2 is a Type II twin cousin of g_1; if g_2 is connected but not isomorphic to g_1, then g_2 is a Type I direct cousin of g_1; if g_2 is disconnected, g_2 is a Type III distant cousin of g_1.

Since the join operation joins g_1 with all its cousins, each $g \in C'_k$ is generated from C_k. □

Algorithm 7 FrequencyCounting(k, i, C)

Input: GD_k - Set of graphs generated by partitioning G with size-k repeated trees;
C - Set of subgraph candidates with k vertices and i edges;
F - Frequency threshold;
Output: D'' - Set of repeated subgraphs with k vertices and i edges;
$D'' \leftarrow \emptyset$;
for each $g' \in C$ **do**
  Get the join parameter of g': g and h;
  $L_g \leftarrow$ set of graphs in GD_k embedding g;
  $L_h \leftarrow$ set of graphs in GD_k embedding h;
  if $f_g < F$ or $f_h < F$ **then**
    $f_{g'} \leftarrow 0$;
  else if type of h = Type I direct cousin **then**
    $f_{g'} \leftarrow |L_g \cap L_h|$
  else if type of h = Type III remote cousin **then**
    $f_{g'} \leftarrow |L_g \cap L_h|$
  else if type of h = Type II twin cousin **then**
    $f_{g'} \leftarrow CheckAllOccurances(g)$;
  end if
  if $f_{g'} > F$ **then**
    $D'' \leftarrow D'' \cup \{g'\}$;
  end if
end for
return D'';

5.4.2 Frequency Counting

A straightforward method to count the frequency of a size-k subgraph g in a graph G is to check all the graph in GD_k. However, this is an NP-complete subgraph isomorphism problem. Given that the discovery of network motifs requires checking the frequency of the candidate subgraphs in both the PPI network as well as the large number of randomized networks, it is critical for us to reduce the computational time of the frequency counting process. This can be achieved by leveraging the properties of the different types of cousins.

Theorem 5.4.2. *Let L_x denote the set of graphs in GD_k such that each graph in L_x embeds x. Let h be a Type I direct cousin of a size-k subgraph g and g' be the subgraph obtained by joining g and h. Then we have $L_{g'} = L_g \cap L_h$, and the frequency of g' is given by $|L_g \cap L_h|$.*

Proof: Each graph in $L_{g'}$ must embed g and h since g' contains all the edges of both g and h. Thus, we have $L_{g'} \subseteq L_g \cap L_h$.

On the other hand, each graph in $L_g \cap L_h$ embeds both g and h. Hence, the graph must embed g', since each edge in g' is in either g or h. Thus, we have $L_{g'} \supseteq L_g \cap L_h$.

Therefore, we have $L_{g'} = L_g \cap L_h$ and the frequency of g' is given by $|L_g \cap L_h|$.

□

Let us consider t_{4_1} and h_2 in Figure 5.7. We have $L_{t_{4\_1}} = \{G_{4\_1}, G_{4\_2}, G_{4\_3}, G_{4\_5}\}$ and $L_{h_2} = \{G_{4\_1}, G_{4\_2}, G_{4\_3}, G_{4\_4}, G_{4\_5}\}$ (see Figure 5.5). Then, for subgraph g_{1_2} which is generated by joining t_{4_1} and h_2, the graphs in GD_4 that embed g_{1_2} are $L_{g_{1\_2}} = L_{t_{4\_1}} \cap L_{h_2} = \{G_{4\_1}, G_{4\_2}, G_{4\_3}, G_{4\_5}\}$. Hence, the frequency value of g_{1_2} is 4.

Similarly, we can prove that if h is a Type III distinct cousin of a size-k subgraph g, the frequency of g' (the subgraph obtained by joining g and h) is also given by $|L_g \cap L_h|$.

However, if h is a Type II twin cousin of a size-k subgraph g, then h is isomorphic to g. In order to determine the frequency of the subgraph obtained by joining g and h, we have to check all the graphs in GD_k that embeds g. This frequency counting involves the NP-complete subgraph isomorphism test. Hence, given that the same subgraph can be generated by joining g with its various types of cousins, we choose to join g with its Type I or Type III cousin whenever possible to avoid the subgraph isomorphism test. Algorithm 7 gives the pseudo-codes for the frequency counting process.

For the complexity analysis of NeMoFinder, please refer to our technical report TRC6/06 (June 2006) [CHLN06a].

5.5 Performance Study

We have implemented our NeMoFinder algorithm in C++ and carried out experiments to compare NeMoFinder with existing network motif discovery algorithms such as the enumeration method [MSOI+02], sampling method [KIMA04], and FPF [SS04].

We use two real-life datasets, the Uetz dataset and the original MIPS CYGD

dataset. The Uetz dataset [UGC+00] contains 957 PPIs and 1004 proteins of *S. cerevisiae* and can be downloaded from the BRITE website. The MIPS CYGD dataset [MFG+02] is the whole-genome PPI network of *S. cerevisiae* from the Munich Information Center for Protein Sequences. After removing redundancy and orphan links, this dataset contains 10199 PPIs involving 4341 proteins that have been detected with high-throughout genome-wide biological experimental methods.

First, we evaluate the runtime of the four network motif discovery methods (enumeration, sampling, FPF, NeMoFinder) in finding network motifs of varying sizes in the Uetz dataset. We set the frequency threshold to 50, the uniqueness threshold to 0.95, and the number of randomized networks to 100. Figure 5.11 shows that NeMoFinder consistently gives the best performance, with 20- to 100-fold speed up. We also observe that only NeMoFinder manages to find all the motifs within a reasonable amount of time.

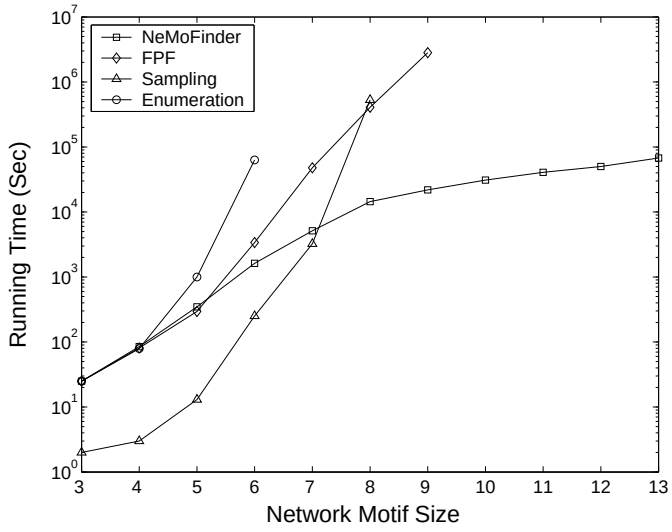

Figure 5.11: Comparison of computational times to find network motifs of varying sizes in Uetz PPI network.

Next, we evaluate the performance of NeMoFinder under varying frequency thresholds. We set the uniqueness threshold to 0.95, the number of randomized networks to 100, and the maximal size of network motif to 9. The enumeration method and sampling method have been excluded from this experiment because they could not scale up to size-9 motifs. Figure 5.12 indicate that NeMoFinder is

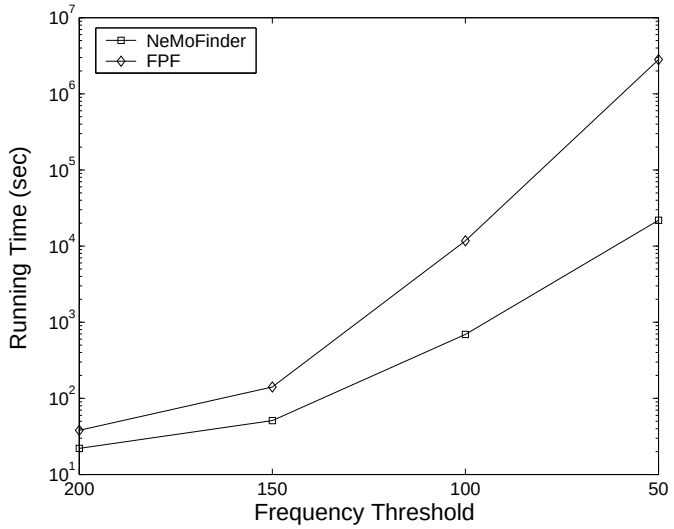

Figure 5.12: Comparison of computational times to find network motifs in Uetz PPI network under varying frequency thresholds.

able to achieve up to 100-fold speedup over FPF.

We also compare the maximal motif size and the total number of identified motifs by the four algorithms to find network motifs of varying sizes in the MIPS dataset, which is much larger than the Uetz dataset. We set the frequency threshold to 50, the uniqueness threshold to 0.95, the number of randomized networks is set to 1000. Figure 5.13 shows that NeMoFinder was able to extract network motifs up to size 12, while the maximum sizes of the motifs discovered by FPF, sampling method and enumeration method are 9, 8 and 5 respectively. In addition, NeMoFinder was able to find a total of 11140 motifs, while FPF, sampling method and the enumeration method discovered only 1112, 848 and 21 network motifs respectively. The limited number of network motifs found by FPF, sampling and enumeration methods was due to the limitation of the motif size that these algorithms could handle.

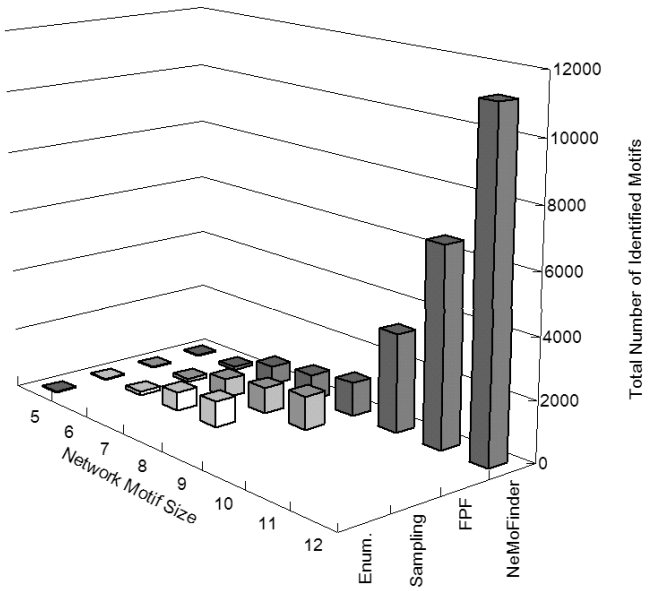

Figure 5.13: Comparison in size and number of network motifs that can be found by four algorithms in MIPS PPI network.

5.6 A Motif Application: PPI Validation

Previous works in biological network motifs have focused mostly on motif discovery; there has been little or no work in showing how the network motifs can be systematically exploited. In this section, we describe how we can exploit the extracted network motifs in PPI validation. Our results show that the inclusion of the larger meso-scale network motifs indeed leads to better results.

Technological developments in high-throughput PPI detection methods such as *yeast-two-hybrid* and *protein chips* have enabled biologists to experimentally detect PPIs at the whole genome level for many organisms. For example, currently more than 15000 PPIs have already been detected and deposited in biological databases for *S. cerevisiae*. The abundant number of PPIs enables scientists to analyze these organisms at the genome level. However, a significant proportion of the PPI networks obtained from high throughput biological experiments has been found to contain false positives. Recent surveys have revealed that the reliability of the

popular high-throughput *yeast-two-hybrid* assay can be as low as 50% [MKS+02]. These errors in the experimental data may lead to spurious discoveries that can be potentially costly, e.g., wrong drug targets for diseases.

Positive results from the various experiments conducted by Saito *et al.* [SSH02b, SSH02a] suggest that the use of even a seemingly primitive network motif in dissecting genome-wide PPI networks is helpful in increasing the reliability of currently erroneous experimental interaction data. In this section, we investigate whether using the actual network motifs can indeed give better performance than using the simple, predefined ones such as those employed in IG1 and IG2.

5.6.1 Motif Strength

We have seen how NeMoFinder is able to discover a much more comprehensive set of network motifs as compared to the other methods (Section 5.5). For it to be useful in practice, it is important that the set of network motifs can provide sufficient coverage of the vast interactome. We found that 96% of PPIs in the MIPS dataset was indeed covered by at least one network motif discovered by NeMoFinder.

First, we rank the network motifs in terms of their contribution to the PPI network with respect to their individual sizes, frequencies and uniqueness. For simplicity, we assume that the motifs are independent here. We define the strength $MS^k(g)$ for each motif g as:

Definition 6. MotifStrength. *The strength of a size-k motif g, denoted as $MS^k(g)$, is the frequency value of the motif times its uniqueness value over max_k, where max_k is the maximal value of $s(g) \times f(g)$ of all size-k motifs.*

$$MS^k(g) = \frac{s(g) \times f(g)}{max_k} \tag{5.2}$$

where $s(g)$ and $f(g)$ are the uniqueness value and the frequency value of subgraph g.

5.6.2 Evaluation based on motif strength

Having defined the MotifStrength, we score each interaction in the PPI network by combining the strengths of the network motifs that contain the interaction (edge).

Definition 7. Reliability Index of PPI *The reliability index of a PPI* (A, B)*, denoted as* $I(A, B)$*, is the sum of the MotifStrength of all the motifs that contain the edge* (A, B)*.*

$$I(A, B) = \sum_{k=2}^{K} \sum_{i=0}^{n} MS^k(g_i) \times k \tag{5.3}$$

where $g_i, 1 \leq i \leq n$ are the motifs where edge (A, B) occurs and k is the size of g_i.

We apply our method, as well as IG1 and IG2, on the MIPS CYGD dataset described in Section 5.5 to compute reliability indices for the 10199 *S. cerevisiae* PPIs in the dataset. We then compare the quality of the various reliability indices in the following three different aspects:

1. **Function Homogeneity.** The cellular functions of the protein partners in a genuine biological interactions are likely to be similar. As such, we would expect an interactome that has been sorted with a good reliability index to exhibit a high degree of functional homogeneity in the interactions with high reliability scores. We use the Comprehensive *S. cerevisiae* Genome Database (dated 2005-06-20) at MIPS [MFG+02] as the ground truth for the proteins' functional annotations. Out of the 4341 proteins in the MIPS CYGD interaction dataset, 3150 proteins have functional annotations and 4743 interactions involve the annotated proteins.

2. **Localization Coherence.** With the exception of the proteins involved in cellular pathways such as the signalling pathway, the cellular localizations of the protein partners in a genuine biological interactions are expected to be the same. As such, a better reliability index would exhibit a higher degree of cellular co-localization amongst the protein partners in the sorted interactions.

We use the cellular localization annotations of the *S. cerevisiae* proteins in the MIPS database as the basis for comparison in our experiment.

3. **Gene Expression Correlation.** Studies have shown that the average correlation coefficient of gene expression profiles that corresponds to interacting protein pairs is significantly higher than those that correspond to random pairs [Gri01]. As such, we can also use the degree of gene expression correlation to evaluate the relative quality of the PPI reliability indices. For gene expression correlation analysis, we downloaded the *S. cerevisiae* gene expression dataset from Eisen's Lab [ESBB98]. The dataset comprises expression vectors from 80 experiments on 6221 genes.

Figure 5.14 shows that as the reliability index value is increased, the proportion of interacting pairs with common cellular functions also increases, indicating an increase in the number of true positives in the filtered interaction data. The reliability indices generated using the NeMoFinder's network motifs show significant increases (from 61% to 87% and 81%) than those using IG1 and IG2 (from 61% to only 68% and 73% respectively). The reliability indices using up-to 8 vertex network motifs has similar performance as IRAP.

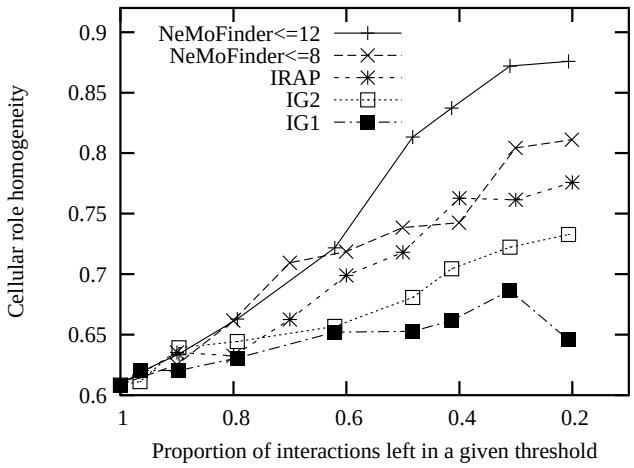

Figure 5.14: Proportion of interacting proteins with common cellular functional roles increases at different rates under different interaction reliability measures.

Figure 5.15 shows the relative performance in terms of cellular localization

coherence. Using the reliability indices computed by the network motifs, the proportion of interacting pairs with common cellular localization increases from 85.3% to 94.0% and 91.7% for the NeMoFinder network motifs, again outperforming IG1 and IG2 (from 85.3% to 87.0% and 90.1%). The reliability indices using up-to 8 vertex network motifs has similar performance as IRAP.

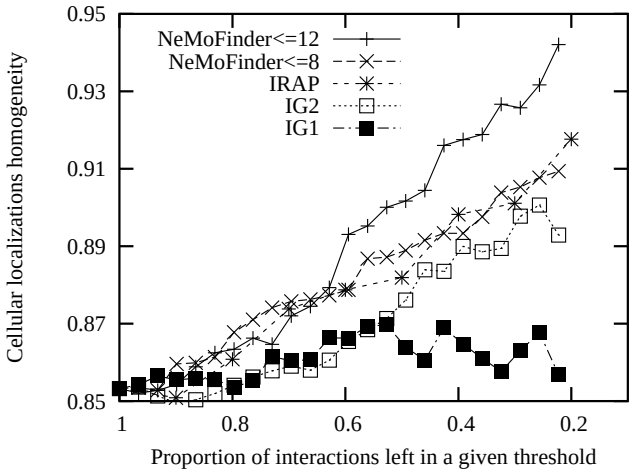

Figure 5.15: Proportion of interacting proteins with common cellular localizations increases at different rates under different interaction reliability measures.

The results based on gene expression correlation, shown in Figure 5.16, exhibit a similar trend. Again, the increase in the average gene expression correlation between the protein partners in the sorted PPIs is much more significant when using reliability indices computed with NeMoFinder's network motifs (from 26.4% to 33.5% and 30.8%) than those generated by using IG1, IG2 and IRAP (from 26.4% to 27.6% 29% and 29.5%).

These results show that the PPI reliability indices computed using the NeMoFinder network motifs are more reliable than those computed using IG1 and IG2, demonstrating the positive effect of using a more comprehensive set of actual network motifs against a small number of simple, predefined motifs. We also found that the reliability indices using up-to 8 vertex network motifs has similar performance as IRAP, but clearly has more coverage than IRAP. Additionally, we also compared the performance of using motifs of different sizes. In all three evaluation experiments, the reliability indexes computed using NeMoFinder network motifs of sizes

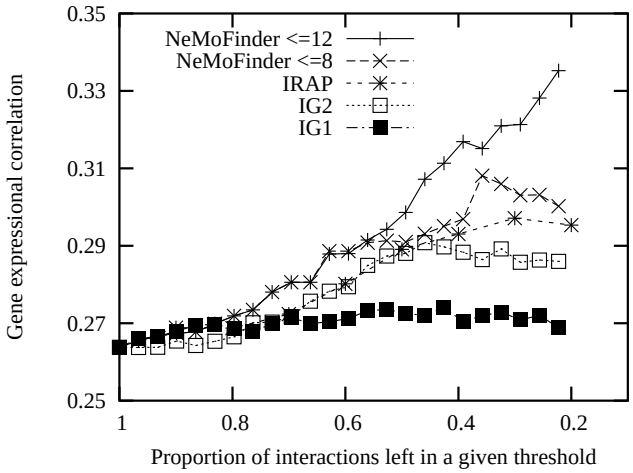

Figure 5.16: Overall correlation of gene expression for interacting proteins increases at different rates under different interaction reliability measures.

up to 12 consistently show superior performance over that computed with motifs of sizes only up to 8. This indicates that it is advantageous to include the larger motifs, justifying the need for discovering meso-scale network motifs.

5.7 Conclusions

Existing network motif discovery algorithms are limited to extracting smaller network motifs and cannot be employed to mine meso-scale level network motifs in large biological networks. In this chapter, we have presented an efficient network motif discovery algorithm called NeMoFinder to discover larger-sized repeated and unique network motifs in PPI networks. The algorithm utilizes repeated trees to partition a network into a set of graphs. We have introduced the notion of graph cousins for efficient candidate generation and frequency counting. We use NeMoFinder to successfully extract, for the first time, up to size-12 network motifs from the whole *S. cerevisiae* PPI network. The network motifs discovered by NeMoFinder provided a good coverage of the PPIs in the vast interactome.

In this work, we also showed an example of how the network motifs can be systematically applied in the validation of the PPIs in an interactome. Our results

confirmed that employing the larger actual network motifs derived from biological networks instead of predefined small-sized network motifs can indeed achieve better results. Future work will include directed network motif discovery and network motif labelling.

CHAPTER 6

Network Motif Labeling

Biological networks such as the protein-protein interaction (PPI) network have been found to contain small recurring subnetworks in significantly higher frequencies than in random networks. Such network motifs are useful for uncovering structural design principles of complex biological networks. However, current network motif finding algorithms model the PPI network as a uni-labeled graph, discovering only unlabeled and thus relatively uninformative network motifs as a result.

Our objective is to exploit the currently available biological information that are associated with the vertices (the proteins) to capture not only the topological shapes of the motifs, but also the biological context in which they occurred in the PPI networks for network motif applications. We present a method called LaMoFinder to label network motifs with Gene Ontology terms in a PPI network. We also show how the resulting labeled network motifs can be used to predict unknown protein functions. Experimental results showed that the labeled network motifs extracted are biologically meaningful and can achieve better performance than existing PPI topology based methods for predicting unknown protein functions.

6.1 Introduction

Motifs in a network are small connected subnetworks that are found to be repeatedly occurring in the network in frequencies that are significantly higher than in random networks. Many complex networks in the real world, such as the gene regulatory network and the protein-protein interaction network, have recently been found to contain such topological patterns of local connections [MSOI+02]. Analysis of network motifs in these naturally occurring networks has led to many interesting results. For example, it was shown that conserved network motifs allow protein-protein interaction predictions [AA04], and that they can be used to discover the underlying network decomposition [IMK05]. As such, network motifs have been gaining increasing attention as a useful concept to uncover structural design principles of complex networks [MSOI+02, MIK04, WR06].

Current approaches in finding network motifs typically consist of two major subtasks:

- **Task 1.** Find which classes of isomorphic subgraphs occur frequently in the input network;
- **Task 2.** Verify which of these subgraph classes are also displayed at a much higher frequency than in random graphs.

The first subtask discovers network motifs that are *frequent* or *repeated* in the network, while the second subtask ensures that they are also *unique*. Clearly, network motif discovery is a computationally challenging problem, but scientists have begun to devise methods for detecting motifs in large networks. For example, the MFINDER by Kashtan *et. al* [KIMA04] supported the detection of network motifs consisting of up to eight vertices, while the latest NeMoFinder by Chen *et. al* [CHLN06b] has enabled the discovery of network motifs with sizes ranging all the way to meso-scale, since many of the relevant processes in biological networks have been shown to correspond to the meso-scale (5-25 genes or proteins) [SM03].

However, the current PPI network motif finding methods are based on a standard graphical model of protein-protein interactions (PPI) as *uni-labeled networks*.

In this model, a species' "interactome" is defined as a network of interactions between the n proteins found in the species (*i.e.* its "proteome"), represented as a graph in which all the vertices (*i.e.* proteins) are *uniquely labeled* with $v_1, \ldots, v_n$. As a result, the network motifs generated by the current motif finding algorithms are "unlabeled", capturing only the topological shapes of the motifs, and not the biological context in which they occurred. While these network motifs have been shown to be somewhat competent for certain biological applications such as protein interaction prediction [AA04], such purely statistical patterns are not informative enough for the more sophisticated biological applications of network motifs that have been envisaged by researchers; for example, in protein function prediction using a dictionary of network motifs and their functional information to predict the functions of unknown proteins [Alo03].

Since the current uni-label model treats each protein in a PPI network as a unique and anonymous entity, it inadvertently ignores any other useful biological information that we may have already known about some of the proteins. In reality, the biologists usually would have performed experimental studies on some of the proteins to determine their biological functional roles and the cellular sublocalization. In fact, there are ongoing systematic efforts to annotate the various proteins in a species' proteome with the known biological information using the Gene Ontology or GO (Section 6.2). This means that the underlying PPI network is actually a partially labeled network, with many of the vertices (i.e. proteins) being already annotated with known functional and cellular sublocalization labels. In order to exploit the availability of such useful biological information associated with the proteins in network motif applications, we introduce a third subtask to the problem of network motif mining:

- **Task 3.** Assign biological labels to the vertices in the network motifs such that the resulting labeled subgraphs also occur frequently in the underlying labeled input network.

The task of labeling the network motifs (formally defined in Section 6.3) turns out to be computationally expensive, due to the sophisticated GO scheme by which

the proteins are annotated. There is often missing information even in the most well-studied model organism. As a result, not all the proteins in the PPI network are annotated with biological information. When they are, many of the proteins would be multiply-labeled since they have complex biological roles. Moreover, the biological labeling scheme is hierarchical, introducing a further element of complexity. As such, even if both the motif size and the number of the motifs are small, it is almost impossible to hand-label the motifs. In fact, the number of possible motifs' labels increases exponentially as we graduate to meso-scale network motifs.

In this chapter, we propose an algorithm, LaMoFinder [CHLN07], which stands for *La*beled *Mo*tif *Finder*, to label the network motifs discovered in a biological network (Section 6.3). Such enrichment of the network motifs enables them to become biologically meaningful for the more sophisticated biological applications such as protein function prediction envisaged by researchers. We apply LaMoFinder to label network motifs mined from the large whole-genome *S. cerevisiae* (Yeast) PPI network for knowledge discovery applications. Our evaluation results show that our labeled network motifs are biologically meaningful (Section 6.4) and can achieve better performance than existing topology-based methods for predicting unknown protein functions using PPI (Section 6.5).

6.2 Gene Ontology

The Gene Ontology (GO) project [GO206] is a collaborative effort initiated since 1998 to construct and use ontologies to facilitate the systematic annotation of genes and their products (e.g. proteins) in a wide variety of organisms. The resulting GO ontologies have now been accepted as the *de facto* language for the description of attributes of genes and gene products, with a rapidly growing number of model organism databases and genome annotation groups contribute annotation sets using GO terms to GO public repository.

The GO ontologies provide a systematic language for the description of attributes of biological entities in 3 key domains that are shared by all organisms, namely molecular function, biological process and cellular component. In each of

these domains, the corresponding GO ontology is structured as a directed acyclic graph (DAG) to reflect the complex hierarchy of biological terminologies. Mathematically, suppose $T = \{t_1, t_2, \ldots, t_n\}$ is a set of GO terms, we say term t_i is a direct child of term t_j, if and only if t_i is an instance ("is-a" relationship) or a component ("part-of" relationship) of t_j ($t_i, t_j \in T$).

To properly model the biological information in different genomes, we also need to take into account that not all the GO terms are equally informative within a certain genome due to their biological differences [LSBG02]. In other words, for each genome, we assign genome-specific weights to the GO terms based on the method suggested by Lord et. al [LSBG02]: the weight of a GO term is defined as the ratio of the number of occurrences of the GO term and any of its descendants' terms in the genome to the total number of terms occurrences in the genome. We denote it as $w(t)$, $\forall t \in T$. By this definition, the GO term weight value is between 0 and 1, and the root node has a weight of 1.

$$w(t) = \frac{N(t) + \sum N(t_{child})}{N_{total}}$$

where $N(t)$ is the number of occurrences of the GO term t, t_{child} is the number of the descendants' terms of t, and N_{total} is the total number of terms occurrences in the genome.

Figure 6.1 shows an illustrative example of a subset of GO. In addition, Table 6.1 shows its protein annotation list. We observe that G04 is a child of G02 following the "is-a" relationship. G06 is a child of G03 following the "part-of" relationship. In addition, the weight of G04 is 0.42 because 245 out of 585 proteins are annotated with G04 or its decedents. Note that it is possible for a child term to have multiple parents in GO. In Figure 6.1, G05 has G02 and G03 as its parents.

Zhou et al [ZKW02] define a GO term as an informative functional class (FC) if the GO term has at least 30 proteins directly annotated with it. In Figure 6.1, G04, G05, G06, G09, and G10 are informative FC. In this work, we are interested in a subset of the informative FC, namely the informative FC with no ancestors that are informative. We call them the *border* informative FC. Border informative FC

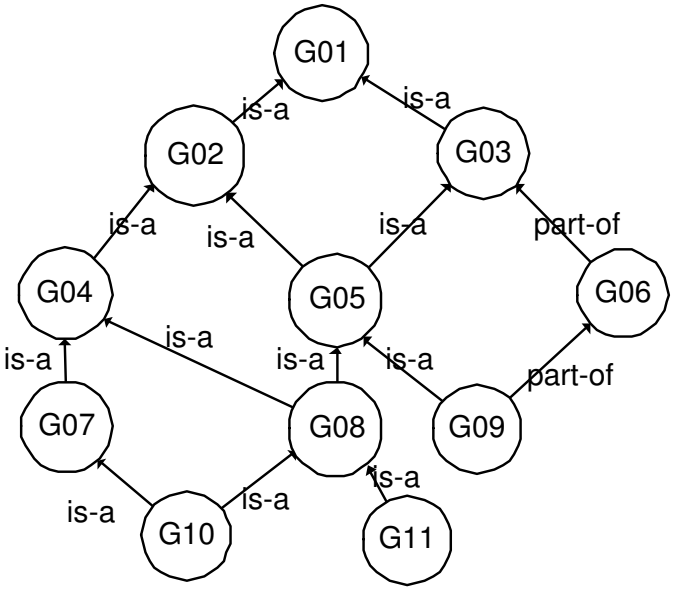

Figure 6.1: Example: a subset of GO.

are used to avoid the generation of labels that would be too general. In our example, G09 and G10 have informative ancestors G05. Hence they are not excluded from the border informative FC.

Having introduced the background of GO annotations, we now illustrate some of the difficulties in labeling network motifs with GO annotations. Figure 6.2 shows an unlabelled network motif g that has been discovered in a PPI network. The occurrences of g in the PPI network G are shown in Figure 6.3 and protein GO annotations are shown in Table 6.2. The task is to label the vertices of g such that the labeling scheme is consistent with some occurrences of g. In other words, the labels must be the same, or more general than the annotation of the corresponding vertex in the occurrence.

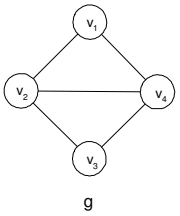

Figure 6.2: Example: network motif g.

GO term t	Num. of proteins annotated with t	Num of proteins annotated with t and its decedents	GO term weight $w(t)$
G01	0	585	1.00
G02	0	415	0.71
G03	20	475	0.81
G04	100	245	0.42
G05	70	280	0.48
G06	150	250	0.43
G07	10	100	0.17
G08	25	135	0.23
G09	100	100	0.17
G10	90	90	0.15
G11	20	20	0.03
SUM	585		

Table 6.1: Example: Weights and the numbers of occurrences of GO terms in Figure 6.1 .

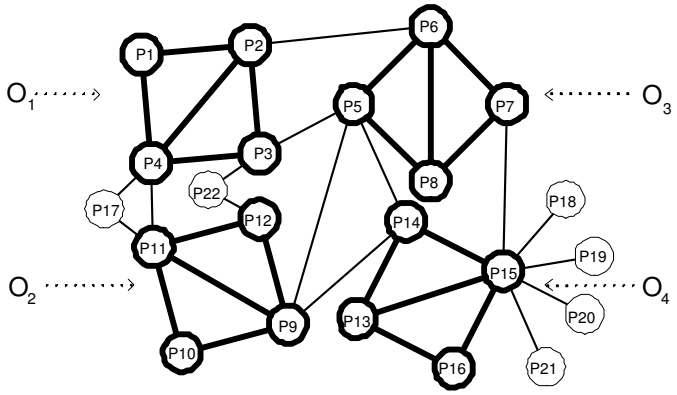

Figure 6.3: Example: 4 occurrences (shown with thick lines) of the network motif g (Figure6.2) in a PPI network G.

For example, suppose we label the vertices $\{v_1, v_2, v_3, v_4\}$ as $\{G04, G08, G04, G05\}$ in Figure 6.3. For occurrence o_1, suppose vertices $\{v_1, v_2, v_3, v_4\}$ are mapped to $\{p_1, p_2, p_3, p_4\}$. We observe that G04 is one of the annotation of p_1 (see Table 2). For p_2, although G08 is not in any of the p_2's annotation, we realize that G10 is in fact a descendant of G08. In other words, assigning G08 to v_2 is appropriate since it is more general than the annotation of p_2 (G10). Similarly, p_3's annotation of G08 is a descendant of G04 and p_4's annotation of G09 is a descendant of G05. We

Protein	GO annotation	Protein	GO annotation
p_1	G04, G09, G10	p_9	G11, G10
p_2	G10, GO3	p_{10}	G03, G05, G07
p_3	G08	p_{11}	G05
p_4	G09, G07	p_{12}	G09
p_5	G03	p_{13}	G11
p_6	G10	p_{14}	G04, G05
p_7	G03	p_{15}	G04
p_8	G05	p_{16}	G04, G09

Table 6.2: Example: GO annotations for proteins in occurrences o_1, o_2, o_3 and o_4.

can conclude that the labeling scheme $\{G04, G08, G04, G05\}$ is consistent with the occurrence o_1.

From this example, we realize that the task of labeling network motifs from biological networks needs to consider the following issues:

1. **Multiple and hierarchical labeling.**

   Biologically, many proteins are involved in multiple cellular processes and they are therefore labeled with more than one GO term, e.g., the proteins in yeast are currently annotated with an average of 9.34 GO terms. Therefore, the number of labeling schemes that are consistent with an occurrence increases exponentially with network motif size. This leads to the scalability issue.

2. **Symmetric vertices.**

   Symmetric vertices are vertices that can be interchanged without affecting the topological structure of the network. For example, the network motif g in Figure 6.2 has two sets of symmetric vertices, $\{v_1, v_3\}$ and $\{v_2, v_4\}$. The existence of these sets of symmetric vertices implies that we need to enumerate all possible mappings between the motif vertices and the occurrence vertices in order to obtain all the possible labeling schemes. Time complexity increases exponentially with the size of the symmetry set. Furthermore, testing whether a graph has any axial symmetry is an NP-complete problem [Man90]. This also increases the complexity of the labeling work.

The LaMoFinder method to be described in Section 6.3 is specifically devised to address the above challenges effectively.

6.3 LaMoFinder

We model a biological network as a graph $G = (V, E)$ where each vertex in V represents a biological entity (e.g., a protein for PPI networks, or a gene for gene regulatory networks), and each edge in E between two vertices v_A and v_B indicates that there exists a biological relation detected between the corresponding proteins/genes A and B. To simplify discussion, we will focus on PPI networks, although our algorithm can be applied to any biological networks.

A network motif g is a frequently occurring non-random subgraph pattern in a network G [MSOI$^+$02]. By definition, g is a connected, unlabeled subgraph that is repeated and unique in G. For each g, there exists a set of occurrences of this network motif in G, denoted as D_g.

Let $T = \{t_1, t_2, \ldots, t_n\}$ be the set of GO terms which will be assigned to the vertices of network motifs as labels. A labeling scheme L of g is said to conform to an occurrence o ($o \in D_g$) if the assigned labels for all vertices of g are either the same or more general than the label of the corresponding vertices in o.

Our goal is to find all possible labeling schemes for the vertices of a network motif g such that they conform to at least σ occurrences in D_g.

A naive approach is to pick an occurrence at random and use its labels as a possible labeling scheme. It then proceeds to determine the number of occurrences that conform to this labeling scheme. If the number of occurrences is less than σ, it picks a combination of vertices at random and generalizes their labels one level up the function hierarchy. With the generalized vertex labels, the total number of occurrences that conforms to the labeling scheme is re-computed. If the number exceeds σ, the scheme is output. The process is repeated till all occurrences have participated in at least one labeling scheme. Clearly, this approach is not scalable. As the network motif size increases, the number of possible vertices combination to generalize increases exponentially. A better approach is needed.

We design a heuristic network motif labeling algorithm called LaMoFinder. Instead of enumerating all possible vertices and their sets of possible generalized labels, we start with the set of occurrences and try to group the occurrences based on their degree of similarity to each other. As the occurrences are grouped, we determine the least general labeling scheme that conforms to all the occurrences in the group. Here, the least general labeling scheme refers to selecting the lowest GO terms that is able to encompass all the occurrences.

In Figure 6.4, suppose we group o_1 and o_2 and assume that $\{p_1, p_2, p_3, p_4\}$ are matched with $\{p_{12}, p_9, p_{10}, p_{11}\}$. The corresponding annotations for $\{p_1, p_2, p_3, p_4\}$ are $\{(G04, G09, G10), (G10, G03), (G08), (G09, G07)\}$; while the corresponding annotations for $\{p_{12}, p_9, p_{10}, p_{11}\}$ are $\{(G09), (G11, G10), (G03, G05, G07), (G05)\}$. Then the least general labeling scheme is $\{(G05, G09), (G08, G10), (G04, G05), (G05)\}$.

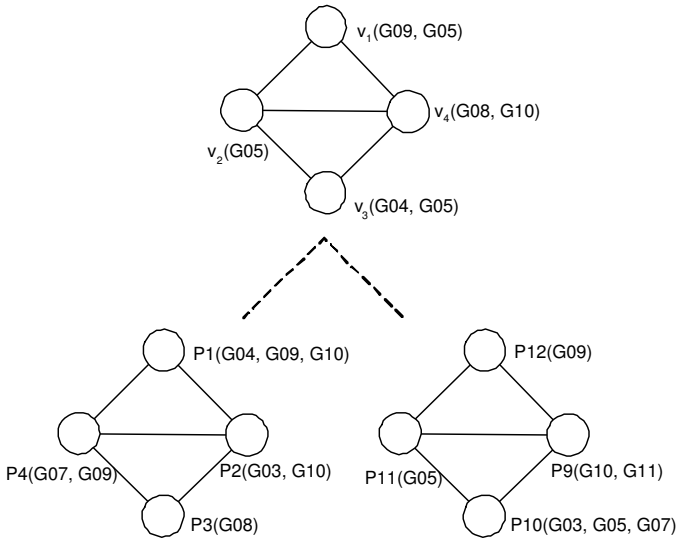

Figure 6.4: Example: The labeling of two occurrences

Two issues immediately surface. The first issue concerns the computation of similarity measures between occurrences. To address this problem, we derive a similarity measure for occurrences based on the GO term similarities. The second issue concerns the grouping criteria. This is dealt with in subsection 6.3.2.

6.3.1 Similarity Measure for Occurrences

As we use GO terms as labels, we first compute the similarity value between any two GO terms. Based on the GO term similarity, we will compute the similarity value between occurrences.

GO Term Similarity. Given two GO terms t_a and t_b and their corresponding weights $w(t_a)$ and $w(t_b)$, we adopt an enriched GO term comparison method [LSBG02] to assign a term similarity score for t_a and t_b, denoted as $ST_{(t_a,t_b)}$.

Recall that GO allows multiple parents for each term. Two terms may share one or more common parents via different paths. For example, in Figure 6.1, G08 and G09 have 2 common parents (G05 and G01). We denote the GO term of the lowest common parent (in our example, this corresponds to G05) as t_{ab}. Then the similarity between GO terms t_a and t_b is defined as:

$$ST(t_a, t_b) = \frac{2 \times \ln w(t_{ab})}{\ln w(t_a) + \ln w(t_b)} \tag{6.1}$$

where $w(t_x)$ is the weight of GO term t_x in T. As $1 \geq w(t_{ab}) \geq w(t_a)$ and $1 \geq w(t_{ab}) \geq w(t_b)$, $ST(t_a, t_b)$ varies between 1 and 0.

Occurrence Similarity. The similarity between any two occurrences o_i and o_j of a network motif g is determined from the similarities between the corresponding vertices of o_i and o_j. The computation of the occurrence similarity has two complications.

The first complication arises from the fact that each vertex of an occurrence may have multiple labels. For any two vertices v_i and v_j, let T_{v_i} and T_{v_j} be the set of GO terms annotated to v_i and v_j respectively, we define the similarity score $SV_{i,j}$ for vertices v_i and v_j as:

$$SV(v_i, v_j) = 1 - \prod_{t_a \in T_{v_i}, t_b \in T_{v_j}} (1 - ST(t_a, t_b)) \tag{6.2}$$

where $ST(t_a, t_b)$ denotes the similarity between GO term t_a and t_b computed with Equation 6.1. Note that $SV(v_i, v_j)$ is close to 1 as long as there is at least one good GO term match among the lists of GO terms in T_{v_i} and T_{v_j}. In other words, two

vertices are considered similar if they share at least one biological feature.

The second complication arises due to the presence of two or more symmetric vertices. In our example, occurrence o_1 has symmetric vertices $\{p_1, p_3\}$ and $\{p_2, p_4\}$ and occurrence o_2 has symmetric vertices $\{p_{12}, p_{10}\}$ and $\{p_9, p_{11}\}$. Let $I_1 = \{v_{11}, v_{12}, \ldots, v_{1t}\}$ be one set of symmetry vertices in o_1 and $I_2 = \{v_{21}, v_{22}, \ldots, v_{2t}\}$ be the corresponding set of symmetry vertices in o_2. We denote $pair(I_1, I_2)$ as the possible pairings of the vertices between the two sets I_1 and I_2. In our example, $pair(\{p_1, p_3\}, \{p_{12}, p_{10}\}) = \{\{(p_1, p_{12}), (p_3, p_{10})\}, \{(p_1, p_{10}), (p_3, p_{12})\}\}$.

Let $\wp_a = \{I_{a1}, I_{a2}, \cdots, I_{ak}\}$ be the set of all sets of symmetric vertices in the occurrence o_i; $\wp_b = \{I_{b1}, I_{b2}, \cdots, I_{bk}\}$ be the set of all sets of symmetric vertices in the occurrence o_j. We define the similarity score of the occurrences o_i and o_j, $SO(o_i, o_j)$, as:

$$SO(o_i, o_j) = \frac{1}{|V|} \sum_{a,b=1}^{k} \left(max\{ \sum_{pair(I_a, I_b)} SV(v_\alpha, v_\beta)\} \right) \tag{6.3}$$

where $|V|$ is the number of vertices in the network motif, and $(v_\alpha, v_\beta) \in pair(I_a, I_b)$, $I_a \in \wp_i$ and $I_b \in \wp_j$.

For example, if we want to compute the pairwise similarity scores for the occurrences o_1 and o_2, we need to find the sets of symmetric vertices. This problem has been proven to be NP-complete by J. Manning in [Man90]. Several heuristics are known to be polynomial in general. Here, we make use of the heuristics provided in the graph algorithm library PIGALE (*http://pigale.sourceforge.net/*).

Table 6.3 shows the occurrence similarity between o_1 and o_2.

6.3.2 Grouping Occurrences

Having worked out the details to compute the similarity of occurrences, the next issue concerns the grouping of the occurrences such that we can find all the possible labeling schemes that encompass the σ number of occurrences.

One possible solution is to use the popular clustering algorithm such as the k-means clustering algorithm to find clusters of size σ. For each cluster, we derive the labeling scheme by assigning to the vertex of the network motif one GO term

occurrence o_1	occurrence o_2	SV score
p_1(G04, G09,G10)	p_{12}(G09)	1.00
p_1(G04, G09,G10)	p_{10}(G03, G05,G07)	0.99
p_2(G03, G10)	p_9(G10, G11)	1.00
p_2(G03, G10)	p_{11}(G05)	0.76
p_3(G08)	p_{10}(G03, G05,G07)	0.80
p_3(G08)	p_{12}(G09)	0.45
p_4(G07, G09)	p_{11}(G05)	0.69
p_4(G07, G09)	p_9(G10, G11)	0.99
SO score		0.87

Table 6.3: Example: Similarity score between occurrences o_1 and o_2

that conforms to all the occurrences of that vertex.

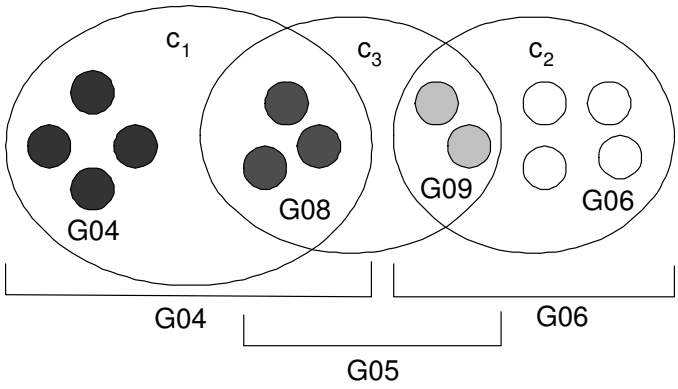

Figure 6.5: Example: Clusters and their labeling schemes.

Unfortunately, this approach does not work well due to the hierarchical structure of the GO ontology. Consider Figure 6.5. We observe that if we use k-means clustering, all the occurrences will be grouped into non-overlapping clusters. We can find 2 labeling schemes c_1 and c_2 with threshold $\sigma = 5$. However, a closer examination shows that there are in fact 3 possible labeling schemes. This example shows that non-overlapping clusters may miss some valid and significant labeling schemes.

In order to discover all the possible labeling schemes for the unlabeled network motifs, we adopt an agglomerative hierarchical clustering method to cluster the occurrences based on the occurrence similarity measures in Section 6.3.1.

In the hierarchical clustering process, each occurrence is initially a cluster by

itself. At each iteration, pairs of the most similar clusters are joined to form a new cluster. The least general labelling scheme of the cluster is derived. If a cluster does not have any occurrence to combine with, it proceeds to the next step. The clustering process stops when the labeling scheme has assigned more than half of the vertices with labels that belong to the border informative FC. If the number of occurrences within the cluster exceeds σ, the cluster's labels are saved as a labeling scheme.

o_1	o_2	common label
G04, G09, G10	G09	G02, G09, G05
G03, G10	G10, G11	G03, G10, G08
G08	G03, G05, G07	G03, G05, G04
G07, G09	G05	G02, G05

Table 6.4: Example: The minimum common father labels of vertices in occurrence o_1 and o_2

The details of LaMoFinder are given in Algorithm 1 and Algorithm 2. LaMoFinder continuously combines the clusters of occurrences until all the labeled network motifs are obtained. In the worst case, LaMoFinder takes $O(|D|^2)$ computational time in the pairwise similarity computation, where D is the size of the occurrence set of network motif g. In the algorithm, the unavoidable graph symmetry process is a NP problem. In this work, we adopted an existing heuristic method that has $O(n^3)$ time complexity, where n is the number of the vertices of g.

6.4 Experiment Results

We implemented LaMoFinder in C++ and carried out experiments on a 3.0GHz single processor Pentium PC with 1GB memory. For evaluation, we applied LaMoFinder on an experimentally-derived (yeast-two-hybrid) interaction data for *Saccharomyces cerevisiae* (yeast) downloaded from the BIND database. The interactome comprises of 7903 Y2H interactions between 4401 of the yeast proteins. After removing redundant links and self-links, the resulting PPI network has 7095 edges and 4141 vertices.

Algorithm 8 LaMoFinder

Input: G - PPI network;
T - the set of GO terms;
g - a network motif of G;
D - occurrence set of g in G;
σ - Frequency threshold;
Output: L - Labeled network motif set;
$L \leftarrow \emptyset$;
$C \leftarrow D$;
$C' \leftarrow \emptyset$;
$\Upsilon \leftarrow getSymmetry(g)$;
while $|C| \neq 1$ and $C \neq C'$ **do**
  $C' \leftarrow C$
  for each cluster $c_i,\ c_j\ \in\ C$ **do**
    $Sim \leftarrow getSimilarity(c_i, c_j, \Upsilon)$;
  end for
  $C \leftarrow Cluster(C', Sim, \Upsilon)$;
end while
for each cluster $c \in C$ **do**
  if $size(c) \geq \sigma$ **then**
    $L \leftarrow c$;
  end if
end for
return L;

Algorithm 9 $Cluster(C', Sim, \Upsilon)$

Input: C' - set of clusters of occurrences of g;
Sim - set of pairwise similarity scores of clusters in C;
Υ - Symmetry vertices set in g;
Output: C - the new set of the clusters;
$C \leftarrow \emptyset$;
for each $c_i \in C'$ **do**
  if less than half of vertices in c_i are border informative FC **then**
    $c_i' \leftarrow c_i$'s closest cluster in C'
    $C \leftarrow Combine(c_i, c_i', \Upsilon)$;
  end if
end for
return C;

We utilized the NeMoFinder algorithm in [CHLN06b] to discover 1367 network motifs from the PPI network. Motifs of sizes up to 20 were discovered by NeMoFinder. All the motifs have frequencies of at least 100 times in the PPI network, with a uniqueness value of more than 0.95 (against random networks).

The GO annotations for the yeast proteins were downloaded from the Gene Ontology database [GO206]. 3554 out of the 4141 yeast proteins are found to have at least one GO biological annotation. There are 3 different branches of GO annotations (function, process and location). We call LaMoFinder 3 times to label the network motifs based on the 3 branches of GO annotations before using them for protein function prediction (Section 6.5).

6.4.1 Meso-scale labeled network motifs

We set the labeled network motif frequency threshold to 10, requiring each labeled network motif to have at least 10 occurrences in the PPI network.

Out of the 1367 unlabeled network motifs, LaMoFinder is able to extract a total of 3842 labeled network motifs from the PPI network. Figure 6.6 shows that the number of labeled network motifs varies with motif size. We observe that the majority of the labeled network motifs are meso-scale. For example, 18.5% labeled network motifs have 16 vertices, and 15.6% labeled network motifs have 17 vertices. This is in accordance to the observation that many relevant processes in biological networks are at the meso-scale (5-25 genes or proteins) level [SM03].

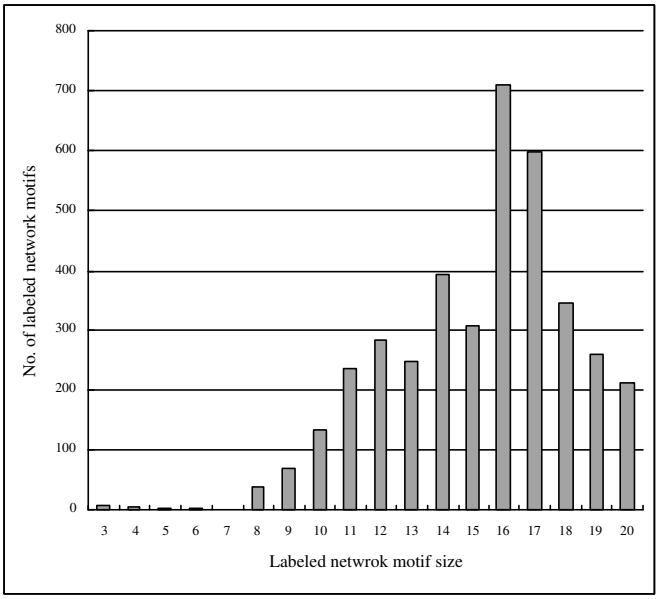

Figure 6.6: Labeled network motif distribution.

6.4.2 Biologically meaningful motifs

We asked a biologist to peruse the different classes of labeled network motifs to verify if there are any motifs discovered by LaMoFinder that would be biologically meaningful.

We first check whether LaMoFinder is able to discover biologically meaningful uni-labeled network motifs, since scientists have observed a notable functional homogeneity in large motifs [WOB03]. Figure 6.4.2 shows a uni-labeled motif g_1 discovered by LaMoFinder that is indeed verified to be commonly found in protein splicing complexes.

Next, we verify whether LaMoFinder is able to discover non-uni-labeled motifs where the vertices have different but biologically related labels. An example is shown in Figure 6.4.2. Unlike g_1, the network motif g_2 is labeled with 3 different function labels. Our biologist has ascertained that g_2 is indeed a biologically meaningful motif because it depicts a interesting biological possibility that a protein with function *"carbohydrate utilization"* can be regulated (via *"mRNA transcription"*) by its indirect neighbor with function *"regulation of carbohydrate utilization"*.

Finally, since we have labeled our motifs with both functional labels as well

as cellular localization labels in this work, we test the biological validity of some of those network motifs that are labeled with these two classes of GO terms at the same time. In fact, just like we have shown in the above non-uni-labeled example, the more complex network motifs can reveal interesting biological insights. For example, the third labeled motif g_3 shown in Figure 6.4.2 illustrates how a parallel-labeled motif can reveal from the PPI network such insightful information as how proteins with different functions may operate in different cellular localizations. The upper triangle of g_3 shows a protein triplet labeled with the same function, suggesting that they are likely to form a protein complex for the purpose of, in this case, rRNA transcription. The other two vertices in the motif depict its functional neighbors that are necessary for this biological process to occur. On closer examination at the parallel cellular sublocalization labels of this motif, we can postulate the various locations in which this complex biological process typically take part.

The above findings illustrate that using LaMoFinder to label network motifs can reveal interesting insights to help biologists better understand the underlying biological processes.

6.5 Application: Protein Function Prediction

Determining protein functions experimentally is an expensive process. As such, even in yeast, the historically most well-studied model organism, only about 60% of yeast proteins have been functionally annotated to-date. Scientists have recently envisaged the accurate prediction of protein functions using a dictionary of network motifs and their functional information [Alo03]. In this section, we describe how this can be achieved with network motifs that have been functionally labeled by LaMoFinder.

6.5.1 Prediction with Labeled Motifs

Suppose we have a labeled network motif $g_{labeled}$ and its set of occurrences O in a PPI network G. We observe that

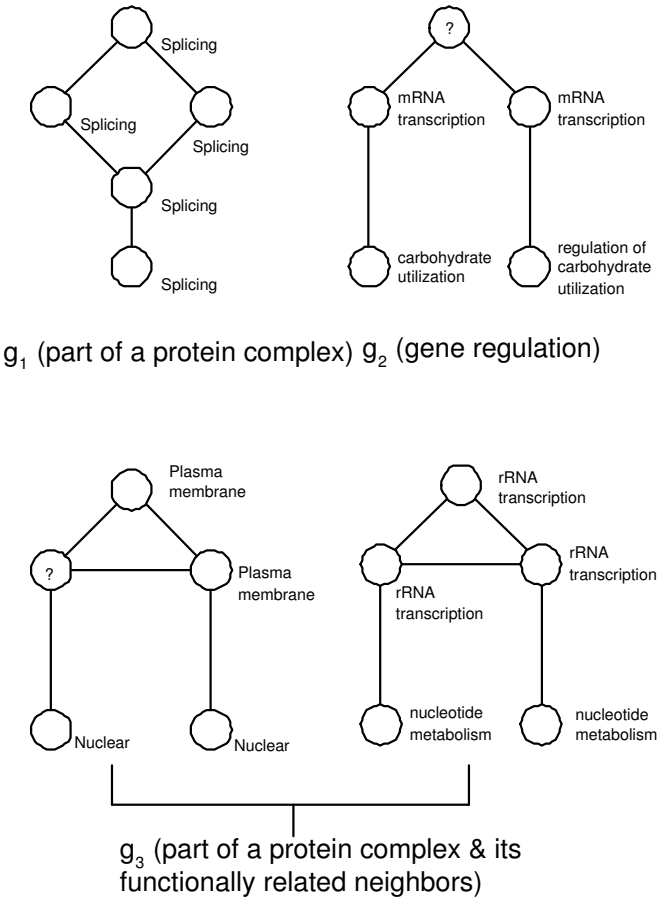

Figure 6.7: Example labeled network motifs.

1. Any protein p in an occurrence $o_i \in O$ is topologically similar to its corresponding proteins in the occurrences $O - \{o_i\}$; and

2. All the proteins in o_i other than p are functionally similar to their corresponding proteins in the occurrences in $O - \{o_i\}$.

Therefore, we propose to predict unknown protein functions by using labeled network motifs as follows:

Given a protein p whose function is unknown, and p is located in an occurrence of a labeled network motif $g_{labeled}$, we can predict the functions of p by using the functions of proteins that are topologically similar to p in the occurrences of $g_{labeled}$.

For example, Figure 6.8 shows an unknown protein p in occurrence o_p. The occurrence o_p is in the cluster of occurrences c_1 which has the labeled motif g_1. We

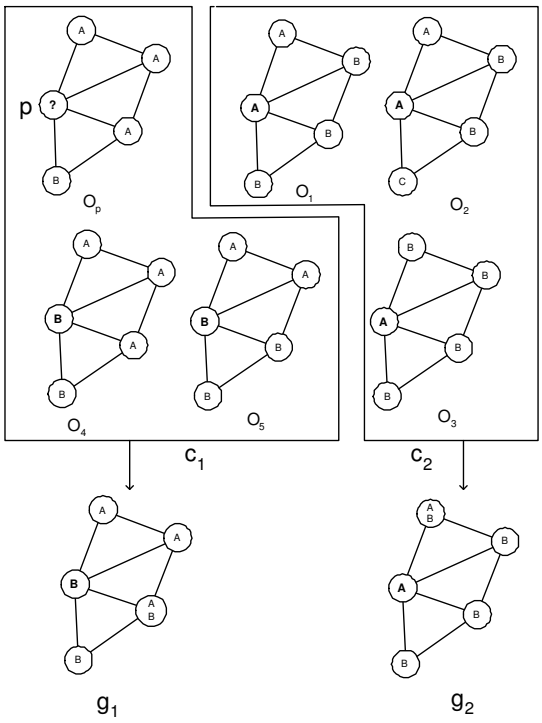

Figure 6.8: Example: predicting function of protein p from labeled motif g_1.

can actually predict that protein p has the function B from the corresponding vertex in the labeled motif g_1.

A straightforward method to predict protein functions using network motifs is to build a dictionary of network motifs and their functions, as suggested in [Alo03]. However, a network motif is likely to have multiple functions, as we have seen in the many non-uni-labeled motifs discovered by LaMoFinder. In order to measure the relation between network motif and protein function more precisely, we define the concept of labeled network motif strength (LMS).

Let g be a network motif, $g_{labeled}$ be a labeled network motif of g. Let $D_{g_{labeled}} = \{o_1, \ldots, o_m\}$ be the set of occurrences of $g_{labeled}$. We say that $g_{labeled}$ is the labeled network motif for a protein p if and only if p is a vertex in o_i ($o_i \in D_{g_{labeled}}$ and $1 \leq i \leq m$).

We can rank the labeled network motifs in terms of their contribution to the PPI network with respect to their individual frequencies and uniqueness. For a labeled network motif $g_{labeled}$, the frequency value is the number of occurrences

in G that conforms to $g_{labeled}$. The uniqueness of $g_{labeled}$ is the number of times g's frequency is equal or greater than its frequency in randomized networks, over the total number of randomized networks [MSOI+02]. For simplicity, we assume that the labeled network motifs are independent of each other. For a labeled network motif $g_{labeled}$, we define the labeled network motif strength $LMS(g_{labeled})$ as:

$$LMS(g_{labeled}) = \frac{s(g_{labeled}) \times |g_{labeled}|}{max_k} \tag{6.4}$$

where $|g_{labeled}|$ is the frequency of $g_{labeled}$; $s(g_{labeled})$ is the uniqueness value of $g_{labeled}$; max_k is the maximal value of $s(g_{labeled}) \times |g_{labeled}|$ of all size-k labeled network motifs.

Given a set of labeled network motifs for protein p, denoted as LG_p, let v be the corresponding vertex of p in a labeled network motif $g_{labeled}$ ($g_{labeled} \in LG_p$), and $x_1, \ldots, x_k$ be the k functions of v. Then the likelihood that protein p has function x is given by:

$$f_x(p) = \frac{1}{z} \sum_{g_{labeled} \in LG_p} (\delta^{g_{labeled}}(v, x) \times LMS(g_{labeled})) \tag{6.5}$$

where $\delta^{g_{labeled}}(v, x)$ returns the frequency of function x on vertex v in $g_{labeled}$. $\delta^{g_{labeled}}(v, x)$ is 0 if x is not a function of v. z is a normalization parameter to ensure that $f_x(p)$ is between 0 and 1.

6.5.2 Results

Previous works have shown that simple topological methods [SUF03, DSC03] could outperform sequenced-based methods, especially in the case of functional similarity without sequence homology. Hence, we expect that using topologically similar proteins will further improve the precision of function prediction. We compare our method with some of the well-known topological associative analysis methods that have been recently shown to be useful in the inference of unknown protein function:

1. The neighbor counting (NC) approach [SUF03] labels a protein with the function that occurs frequently in its neighbors. The k most frequent functions are assigned as the k most likely functions for that protein.

2. Chi-Square (Chi^2) approach is a statistical approach proposed by Hishigaki et al [HNO+01] that makes use of Chi-Square statistics to take into account the frequency of each function in the dataset.

3. PRODISTIN [BCM+03] uses the Czekanowski-Dice distance between each pair of proteins as a distance metric and clusters the proteins using the BIONJ algorithm.

4. The MRF approach proposed by Deng *et. al* [DZM+03] is a global optimization method based on Markov Random Fields and belief propagation to compute a probability that a protein has a function given the functions of all other proteins in the interaction dataset.

All the above prediction methods are based on the functional information of nearby proteins in the network. The proposed use of meso-scale labeled network motifs will enable, for the first time, the exploitation of remote but topologically similar proteins for the functional prediction of unknown proteins.

To facilitate comparison, we use the same PPI dataset employed by the other methods. The PPI dataset was download from MIPS and it comprises 1877 proteins and 2448 physical interactions after removing 120 pairs of self-interactions. We apply NeMoFinder followed by LaMoFinder to discover a set of labeled network motifs for this MIPS dataset. Then, we use a leave-one-out strategy to recognize 13 functional categories of yeast proteins. Figure 6.9 shows the precision and recall of the various methods. The proposed labeled network motif prediction method shows improved accuracy.

6.6 Conclusion

Many biological networks such as the PPI network have been found to contain small recurring subnetworks in significantly higher frequencies than in random networks [MSOI+02]. Scientists have believed that such overabundant topological modules in the network can be useful for uncovering the structural design principles of complex biological networks. However, current network motif finding algorithms invariably

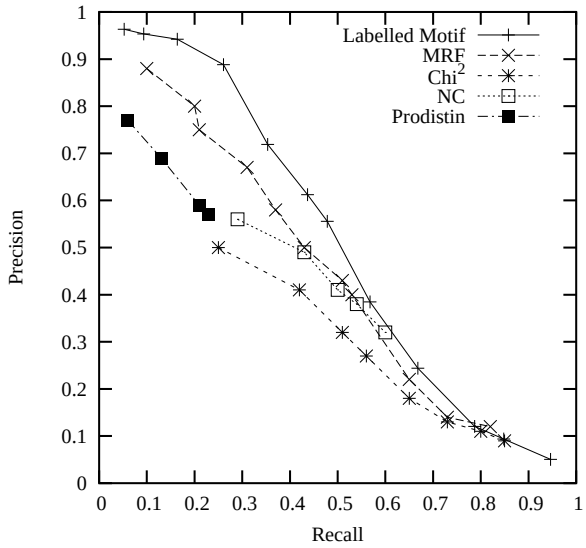

Figure 6.9: Precision vs. Recall for labeled network motif functional prediction

models the PPI network as a uni-labeled graph, limiting themselves to only discovering unlabeled (and uninformative) network motifs. As a result, the currently available biological information that are associated with the vertices (the proteins) cannot be exploited for further knowledge discovery applications.

In this work, we have proposed a method called LaMoFinder to annotate network motifs with the biological information associated with the proteins in the PPI network. Our method was specifically devised to handle the large labeling space as well as the sophisticated scheme (GO) in which the proteins were annotated. As a result, we have captured not only the topological shapes of the motifs, but also the biological context in which they occurred in the labeled network motifs.

We also demonstrated how the network motifs labeled by LaMoFinder can be used to predict the functions of unknown proteins in the PPI network. Our superior performance against other current prediction methods confirmed that the network motifs have indeed been adequately enriched by LaMoFinder for the more sophisticated biological applications such as protein function prediction. For further work, we plan to look into mining labeled and directed network motifs, as many real-world networks can also be modelled with directed graphs.

CHAPTER 7

Discussion

The dissection of the protein interactome is important for extracting invaluable biological knowledge for understanding the molecular mechanism of our cellular system, and eventually leads to the discovery of new drugs and drug targets for various human diseases. Thus far, most of the recent technological advance in this field has focused on the high-throughput detection of protein interactions in order to map the tremendously vast protein interactome. Unfortunately, the protein interaction data that have been generated in large-scale experimental studies using the high throughput technologies have very high error rates, and a large proportion of protein functions are unknown. In this work, we therefore focused on tackling the problem of high false positive rates in high-throughput experimental protein interaction data, and predict unknown protein functions as well using network topologies.

7.1 Review of main findings

We proposed the use of a novel measurement—*Interaction Reliability by Alternative Path (IRAP)*—to computationally assess the reliability of candidate protein interactions by using the topological properties of the underlying protein interaction network. We developed an algorithm called *alternative path finder* to compute the

IRAP values efficiently in large, interconnected, and loopy protein interaction networks. Given the expensive computational requirement of the algorithm, we then devised a heuristic IRAP algorithm that selects the most promising paths via an estimation function. This estimation function was designed to capture the concept of "hub" nodes – a widely recognized scale-free behavior in the protein interaction network.

Results from our extensive experiments showed consistently that the IRAP measure is an effective way for discovering reliable protein interactions in large datasets of error-prone experimentally-derived protein interactions, and the heuristic IRAP is able to achieve remarkable speedup while maintaining a high degree of accuracy. Our results also indicated that IRAP is better than IG2, and markedly better than the more simplistic IG1 measure. The outstanding performance of IRAP showed that a global, system-wide approach—such as our IRAP measure that considers the entire protein interaction network instead of only local neighbors—is a much more promising approach for assessing the reliability of protein interactions.

Beside the high error rates of the high-throughput PPI networks, the false negative rate of the networks have also been estimated to be as high. In this thesis, we have proposed a novel computational complement for the repurification of the experimentally-derived interactomes. We iteratively refine an interactome by removing interactions that are identified as false positives and adding interactions detected as false negatives into the interactome. The computationally repurified interaction data sets were shown to contain potentially lower fractions of false positive and false negative errors. Additionally, biologically interesting interactions such as cross-talkers may also be discovered using our method. Note that in this work, the detection of the potential experimental errors was intentionally done using only the topological information that were mathematically derived from the underlying interaction graphs. This is to allow us to clearly illustrate the potential usefulness of such a topological approach.

We also presented a network motif model to find frequent and unique network motifs in the protein interaction networks, and to evaluate protein interactions with these motifs. By discovering network motifs, protein interaction networks were

broken down into simple units that can help researchers discover unknown principle of complex network. Overcoming the drawbacks of existing algorithms for detecting unique network motifs, The algorithm NeMoFinder could rapidly scale to meso-size network motifs. In the algorithm NeMoFinder, a new framework was designed with the ability to directly scale to motifs with certain size. In the framework, frequent trees were firstly discovered, because tree is a simper topological structure than graph and the number of distinct trees is much less than the number of graphs with the same size. By finding frequent trees, graph G was naturally divided into a set of graphs GD, in which each graph was an embedding of a frequent tree. Then, three kinds of join operations were introduced to reduce the computational time of motif candidate generation and frequency counting in GD. Experimental results showed that NeMoFinder was able to discover meaningful network motifs from the yeast protein interaction network successfully. While running NeMoFinder on yeast data, we discovered about 100 times more network motifs than existing ones. The protein interaction evaluation based on meso-scale network motifs are more reliable than small local motifs (c.f. "IG2").

The performance of meso-scale network motif is similarly accurate as IRAP, but has advantages if network is sparse (i.e., where few alternate paths are present). The results suggest that the two approaches, alternative path and network motif, can facilitate the rapid construction of protein interaction networks that help scientists in understanding the biology of living systems and unknown behaviors of real networks.

Current network motifs are unlabeled (and uninformative). As a result, the currently available biological information that are associated with the vertices (the proteins) cannot be exploited for further knowledge discovery applications. In this thesis, we have proposed a method called LaMoFinder to annotate network motifs with the biological information associated with the proteins in the PPI network. Our method was specifically devised to handle the large labeling space as well as the sophisticated scheme (GO) in which the proteins were annotated. As a result, we have captured not only the topological shapes of the motifs, but also the biological context in which they occurred in the labeled network motifs. We also demonstrated how the network motifs labeled by LaMoFinder can be used to predict the functions

of unknown proteins in the PPI network. Our superior performance against other current prediction methods confirmed that the network motifs have been adequately enriched by LaMoFinder for the more sophisticated biological applications such as protein function prediction.

7.2 Recommendations

The IRAP and IRAP* measures are currently based on the "strongest alternative path" model. A candidate interaction that is not accompanied by a strong alternative path of interactions in the overall protein interaction network is considered to be unreliable. While this may not be true for all the biologically relevant protein interactions, we have performed an analysis on our yeast-two-hybrid protein interaction datasets and found that more than 80% of interactions in our experiments do have at least one alternative path. With a significant proportion of interactions captured by the current IRAP and IRAP* measures, it is acceptable that the measure cannot evaluate the other 20% of protein interactions.

The other measure, network motif, is based on the frequent and unique subgraphs that are found solely in the current protein interaction network. Protein interactions that are captured by at least one significant network motif are considered to be reliable. As this work focuses on the topological significant interactions which are thought to be the most biologically important, the protein interactions with no network motifs involved are lost. The labeled network motifs cover even less protein interactions. The number of the lost interactions varies with the threshold of frequency and uniqueness given by users. Generally, for *S. cerevisiae*, about 96% protein interactions are involved in at least one network motif; for *E. coli*, about 80% protein interactions are involved in at least one network motif.

Therefore, while both of the two approaches capture a large part of protein interactions with distinct approaches, there are still a certain proportion of the protein interactions that cannot be evaluated by current IRAP/motif model. The next step is to develop further network models to capture protein interactions associated with more sophisticated topological characteristics than alternative paths and network

motifs. New models could be developed in the following ways.

7.2.1 Combine IRAP/motif model with other existing models

Besides the IRAP/motif model, there are some existing protein interaction evaluation methods based on the protein interaction network topology. For example, Bader *et al* [BCC04] developed a quantitative method which treated pairs of proteins close together in multi-networks as positive examples, and proteins connected in one network and far apart in the second network as negative examples.

By combining the existing protein interaction evaluation methods with the IRAP/motif model, detection of more protein interactions in the protein interaction data may be possible.

7.2.2 Disconnected Network Motifs

In our network motif model, we focused on finding the simplest topological units that are connected. A network motif is connected if there is a path between every pair of vertices in the motif. However, the current protein interaction network is not only with many false positives but also has a high ratio of false negatives. The false negative problem is critical by the fact that the combination of independent datasets results in a low overlap rate[HF01, MKS$^+$02]. With the missing interactions, an interesting network motif could be separated. Consequently, a disconnected network motif will be overlooked by our network motif model since it focuses only on connected motifs.

Therefore, it would be interesting to develop an algorithm to discover disconnected network motifs with gaps (missing nodes or missing edges). The disconnected motifs could be generated by glue smaller connected motifs that often occur in the protein interaction network with a close distance, or could be discovered directly in a similar way as finding discrete subgraphs or subtrees in a complex network.

7.2.3 Incorporate with protein functional interaction networks

The linkage in the protein functional interaction network indicates that the two connected proteins have the same function. Naturally, the functional network is much larger than the physical protein interaction network that we focused on.

An alternative path that does not appear in the physical interaction network but appears in the functional network may indicate two possibilities. First, the two target proteins are strongly correlated in functional annotations but not physically connect with each other. Second, there exist a physical alternative path, but the path does not exist in the current physical protein interaction network due to the high error rate. Therefore, the interacting pair with only functional alternative path could be assigned a weight based on the number of missing edges in the path. With this approach, we hope to detect more protein interactions in the physical interaction data.

It is also reasonable to assume that there are no false negatives in the functional network, which means the physical protein interaction data is a subset of the functional protein interaction data. Hence, in the disconnected network motif discovery approach introduced in section 7.2.2, a gap in physical network should have its corresponding edge in the functional network. Therefore, the disconnected motif discovery approach could be more effective since the search space is dramatically reduced.

7.3 End note

There is no better way to end the thesis by relating some "history" [CCH+06]. Professor Limsoon Wong first learned, at GIW 2002, of the possibility of ranking the reliability of protein-protein interactions reported in high-throughput Y2H assays from Dr. Rintaro Saito, who was showing a poster of his works on IG1 [SSH02b] and IG2 [SSH02a]. Professor Limsoon Wong was so impressed with the poster that, upon returning to Singapore, he told his colleagues Dr. See-Kiong Ng and Mr.

Soon-Heng Tan about it. Dr. See-Kiong Ng subsequently followed up on the idea with his collaborators A/P Wynne Hsu, Dr. Mong Li Lee and me; and developed improvements including IRAP [CHLN04, CHLN05b, CHLN05a], IRAP* [CHLN06c], and NeMoFinder [CHLN06b, CHLN07]. Mr. Soon-Heng Tan did not follow up on the idea, though he was inspired to work on identification of protein-protein binding motifs [LLTN04]. Professor Limsoon Wong followed up on the paper, and co-authored with Haiquan Li and Jinyan Li a paper on binding motifs [LLW06]. He also co-authored with Dr. Wing-Kin Sung and Mr. Hon Nian Chua a paper on using indirect neighbours to infer protein function [CSW06]. As we can see, the discussion of professor Limsoon Wong and Dr. Rintaro Saito at GIW 2002 has lead to a fruitful chain of results.

BIBLIOGRAPHY

[AA04] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, 2004.

[ABC+04] Patrick Aloy, Bettina Bottcher, Hugo Ceulemans, et al. Structure-based assembly of protein complexes in yeast. *Science*, 303:2026–2029, 2004.

[Alo03] U. Alon. Biological networks: the tinkerer as an engineer. *Science*, pages 1866–1867, 2003.

[BCC04] J.S. Bader, A. Chaudhuri, and J. Chant. Gaining confidence in protein interaction networks. *Nature*, 22(1):78–85, 2004.

[BCM+03] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.*, 5(1):R6, 2003.

[BDH03] G.D. Bader, Betel D., and C.W. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[BJR+02] A. L. Barabasi, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and A. Schubert. Evolution of the social network of scientific collaborations. *Physica A*, 311:590, 2002.

[BR99] Albert-Laszlo Barabasi and Albert Reka. Emergence of scaling in random networks. *Science*, 286, 1999.

[CCH+06] Jin Chen, Hon Nian Chua, Wynne Hsu, Mong Li Lee, See-Kiong Ng, Rintaro Saito, Wing-Kin Sung, and Limsoon Wong. Increasing confidence of protein-protein interactomes. *GIW*, 2006.

[CHLN04] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Systematic assessment of high-throughput experimental data for reliable protein interactions using network topology. *ICTAI*, pages 368–372, 2004.

[CHLN05a] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Assessing reliability of protein interaction data from high-throughput experiments with belief inference(poster). *APBC*, 2005.

[CHLN05b] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Towards discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in medicine*, 35(1-2):37–47, 2005.

[CHLN06a] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Discovering and exploiting meso-scale network motifs in protein interactomes. Technical Report TRC6/06, National University of Singapore, 2006.

[CHLN06b] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Dissecting genome-wide protein-protein interactions with meso-scale network motifs. *SIGKDD*, 2006.

[CHLN06c] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22(16):1998–2004, 2006.

[CHLN07] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Labeling network motifs in protein interactomes for protein function prediction. *ICDE*, 2007.

[CSW06] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623–1630, 2006.

[DBTM+01] A. Davy, P. Bello, N. Thierry-Mieg, et al. A protein-protein interaction map of the caenorhabditis elegans 26s proteasome. *EMBO Rep*, 2(9):821–828, 2001.

[Dij59] E.M. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[DMSC02] Minghua Deng, Shipra Mehta, Fengzhu Sun, and Ting Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12:1540–1548, 2002.

[DSC03] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *PSB*, 2003.

[DSXE02] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, 1:349–356, 2002.

[DZM+03] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Z. Sun. Prediction of protein function using protein-protein interaction data. *J. Comp. Biol.*, 10(6):947–960, 2003.

[EIKO99] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, 95:14863–14868, 1998.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*, pages 251–262, 1999.

[For96] S. Fortin. The graph isomorphism problem. *Technical Report TR96-20, Department of Computing Science, University of Alberta*, 1996.

[FS89] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, 1989.

[G+02] A.C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[GBBK02] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and casual structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63, 2002.

[GDC03] Bader GD, Betel D, and Hogue CW. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, 2003.

[GO206] The gene ontology (go) project in 2006. *Nucleic Acids Res*, 34(Database issue):322–326, 2006.

[GR03] Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100(8):4372–4376, 2003.

[Gri01] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res*, 29(17):3513–3519, 2001.

[HF01] T. R. Hazbun and S. Fields. Networking proteins in yeast. *Proc Natl Acad Sci U S A*, 98(8):4277–4278, 2001.

[HNO+01] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):523–531, 2001.

[HS00] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.

[HWP03] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraph in the presence of isomorphism. *ICDM*, pages 549–552, 2003.

[HWPY04] J. Huan, W. Wang, J. Prins, and J. Yang. Spin: Mining maximal frequent subgraphs from graph databases. *SIGKDD*, 2004.

[ICO+01] T. Ito, T. Chiba, R. Ozawa, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, 2001.

[IMK05] S. Itzkovitz, R. Milo, and N. Kashtan. Coarse-graining and self-dissimilarity of complex networks. *Phys. Rev. E*, 71(016127), 2005.

[ITM+00] T. Ito, K. Tashiro, S. Muta, et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3):1143–1147, 2000.

[IWM00] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph. *PKDD*, pages 13–23, 2000.

[JMBO01] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[JYG+03] R. Jansen, H. Yu, D. Greenbaum, et al. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, pages 449–453, 2003.

[KGKN02] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The kegg databases at genomenet. *Nucleic Acids Res*, 30(1):42–46, 2002.

[KI05] Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23:561–566, 2005.

[KIMA04] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[KK04a] M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. *TKDE*, 2004.

[KK04b] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. *In SIAM International Conference on Data Mining*, 2004.

[LLTN04] Haiquan Li, Jinyan Li, Soon-Heng Tan, and See-Kiong Ng. Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. *PSB*, pages 312–332, 2004.

[LLW06] Haiquan Li, Jinyan Li, and Limsoon Wong. Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.

[LRR+02] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.R. Harbison, C.M. Thompson, Simon I., Zeitlinger J., E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J. Tagne, Volkert T.L., E. Fraenkel, Gifford D.K., and R.A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[LSBG02] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 2002.

[LSBG03] PW Lord, RD Stevens, A Brass, and CA Goble. Semantic similarity

measures as tools for exploring the gene ontology. *In Proceedings of the Pacific Symposium on Biocomputing*, pages 601–612, 2003.

[LWG01] P. Legrain, J. Wojcik, and J. M. Gauthier. Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet*, 17:346–352, 2001.

[Man90] J. Manning. Geometric symmetry in graphs. *Ph.D thesis, Purdue University*, 1990.

[MFCG03] L. Mirabeau, I. Feldman, M. Cokol, and E. Goodnoe. odeling innovation with fitness landscapes: The star network motif. *NECSI*, 2003.

[MFG+02] H. W. Mewes, D. Frishman, U. Guldener, et al. Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–34, 2002.

[MHMF00] S. McCraith, T. Holtzman, B. Moss, and S. Fields. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A*, 97(9):4879–4884, 2000.

[MIK04] R. Milo, S. Itzkovitz, and N. Kashtan. Superfamilies of designed and evolved networks. *Science*, 303(5663):1538–1542, 2004.

[MKS+02] C.V. Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of largescale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

[MP+99] E. Marcotte, M. Pellegrini, et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(751-753), 1999.

[MS02] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[MSOI+02] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[MVR+01] Lisa R. Matthews, Philippe Vaglio, Jerome Reboul, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126, 2001.

[Oli00] S. Oliver. Guilt-by-association goes global. *Nature*, 403:601–603, 2000.

[P+03] S Peri et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363–2371, 2003.

[PB01] J. Park and D. Bolser. Conservation of protein interaction network in evolution. *Genome Inform Ser Workshop Genome Inform*, 12:135–140, 2001.

[PMT+99] M. Pellegrini, E.M. Marcotte, M.J. Thompson, et al. Assigning protein functions by comparative analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96(4285-4288), 1999.

[PWJ04] N. Przulj, D.A. Wigle, and I. Junsica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, 2004.

[PZ05] Pengjun Pei and Aidong Zhang. A topological measurement for weighted protein interaction network. In *CSB*, pages 268–278, 2005.

[RSDR+01] J. C. Rain, L. Selig, H. De Reuse, et al. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–215, 2001.

[SCN+04] Li S, Armstrong CM, Bertin N, Ge H, et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

[SG01] I.G. Serebriiskii and E.A. Golemis. Two-hybrid system and false positives. approaches to detection and elimination. *Methods Mol. Biol.*, 177:123–134, 2001.

[SM03] V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, 2003.

[SOMMA02] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherchia coli. *Nature*, 31:64–68, 2002.

[SS04] F. Schreiber and H. Schwobbermeyer. Towards motif detection in networks: Frequency concepts and flexible search. *NETTAB'04*, pages 91–102, 2004.

[SSH02a] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19:756–763, 2002.

[SSH02b] R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res*, 30:1163–1168, 2002.

[SSM03] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–923, 2003.

[SUF03] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnol*, 18:623–627, 2003.

[TO00] Sophia Tsoka and Christos A. Ouzounis. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet.*, 26:141–142, 2000.

[UGC+00] P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[Wag01] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18:1283–1292, 2001.

[Wag03] Andreas Wagner. Does selection mold molecular networks? *Sci. STKE*, page 41, 2003.

[Wat03] Duncan J. Watts. *SmallWorlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ:Princeton University Press, 2003.

[WBV00] A. Walhout, S. Boulton, and M. Vidal. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast*, 17:88–94, 2000.

[WOB03] S. Wuchty, Z.N. Oltvai, and A.L. Barabasi. *Nature Genetics*, 25:176–179, 2003.

[WR06] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[WS01] Jerome Wojcik and Vincent Schachter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Suppl 1:S296–305, 2001.

[WSL+00] A. Walhout, R. Sordella, X. Lu, et al. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287:116–122, 2000.

[Wuc01] S. Wuchty. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18(9):1694–1702, 2001.

[XRS+00] I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg. Dip: The database of interacting proteins. *Nucleic Acids Research*, 28:289–291, 2000.

[XSD+02] I. Xenarios, L. Salwnski, X.J. Duan, et al. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.

[YH02] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. *ICDM*, 2002.

[YLSea04] E. Yeger-Lotem, S. Sattath, and N. Kashtan et. al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, 101(16):5934–5939, 2004.

[Z+01] H. Zhu et al. Global analysis of protein activities using proteome chips. *Science*, 293:2101–2105, 2001.

[ZKW02] X. Zhou, M. C. Kao, and W. H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U S A*, 99(20):12783–88, 2002.