Efficient Systems Biology Algorithms for Biological Networks over Multiple Time-Scales: From Evolutionary to Regulatory Time

by

Antonina Mitrofanova

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Department of Computer Science
Courant Institute of Mathematical Sciences
New York University

September 2009

Advisor: Bud Mishra

To Nicole

&

My family

Acknowledgements

I would like to thank all the people in my academic and personal life who motivated, supported, and, most importantly, believed in me.

First, I would like to thank my advisor, Bud Mishra, whose belief, encouragement, and advice have been a driving force of my work. He is a person of a great knowledge, beautiful mind, and deep wisdom, and I am honored to have been privileged to work with him for the last several years.

Additional thanks go to my advisors at Rutgers University and defense committee members, Martin Farach-Colton and Vladimir Pavlovic, for shaping me as a young scientist, teaching how to work hard, and supporting and encouraging throughout my academic life.

I would like to give special thanks to my early undergraduate advisor, Aggelos Kiayias, for motivating me to become a scientist and keeping his belief in me even in the most challenging periods of my life.

I would like to express my deep gratitude to all other members of my thesis committee: namely, Simon Kasif, Alan Siegel, and Ernest Davis. Thank you for your advice and useful comments on my thesis.

I would like to thank all my teachers whom I did not list. You have been my role models, as you have encouraged me to enter an academic career and to fall

in love with teaching.

I am thankful to my dear friends and colleagues at the graduate school. Thank you so much for listening, giving advice, and sharing sweets, books, and moments of fun. Without you graduate school would not have been the same.

Finally, I would like to thank my family, especially my husband Vadim, my parents (Nataliya and Nikolay); my grandparents (Antonina, Nina, and Fedor); my family-in-law (especially Elena, Yevgeniy, Alex, and Aljona). Thank you for your love, endless support, and for believing in me. Most of all, I would like to thank my daughter Nicole, for whom I have always been a special person and mom. I know that I would never have done so much if it were not for you, Nicole.

Thank you.

Abstract

Recently, Computational Biology has emerged as one of the most exciting areas of computer science research, not only because of its immediate impact on many biomedical applications, (e.g., personalized medicine, drug and vaccine discovery, tools for diagnostics and therapeutic interventions, etc.), but also because it raises many new and interesting combinatorial and algorithmic questions, in the process. In this thesis, we focus on robust and efficient algorithms to analyze biological networks, primarily targeting protein networks, possibly the most fascinating networks in computational biology in terms of their structure, evolution and complexity, as well as because of their role in various genetic and metabolic diseases.

Classically, protein networks have been studied statically, i.e., without taking into account time-dependent metamorphic changes in network topology and functionality. In this work, we introduce new analysis techniques that view protein networks as being dynamic in nature, evolving over time, and diverse in regulatory patterns at various stages of the system development. Our analysis is capable of dealing with multiple time-scales: ranging from the slowest time-scale corresponding to evolutionary time between species, speeding up to intra-species pathway evolution time, and finally, moving to the other extreme at the cellular

developmental time-scale.

We also provide a new method to overcome limitations imposed by corrupting effects of experimental noise (e.g., high false positive and false negative rates) in Yeast Two-Hybrid (Y2H) networks, which often provide primary data for protein complexes. Our new combinatorial algorithm measures connectivity between proteins in Y2H network not by edges but by edge-disjoint paths, which reflects pathway evolution better within single species network. This algorithm has been shown to be robust against increasing false positives and false negatives, as estimated using variation of information and separation measures.

In addition, we have devised a new way to incorporate evolutionary information in order to significantly improve classification of proteins, especially those isolated in their own networks or surrounded by poorly characterized neighbors. In our method, the networks of two (or more) species are joined by edges of high sequence similarity so that protein-homologs of different species can exchange information and acquire new and improved functional associations.

Finally, we have integrated many of these techniques into one tool to create a novel analysis of malaria parasite *P. falciparum*'s life-cycle at the scale of reaction-time, single cell level, and encompassing its entire inter-erythrocytic developmental cycle (IDC). Our approach allows connecting time-course gene expression profiles of consecutive IDC stages in order to assign functions to unannotated Malaria proteins and predict potential targets for vaccine and drug development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Biological systems are complex machineries driven by active biochemical entities, such as genes and proteins. Every living cell carries its genetic code in DNA molecules, where genetic alphabet is represented by four characters of DNA bases. Genetic alphabet encodes genes, which operate, regulate, and maintain living cells. Further, DNAs can be transcribed into a more portable form, namely, RNA, which carries genetic information closer to cellular sources, and produces substances encoded by genes - proteins. *Proteins* are the basis of life. Anything in a living cell is made *of* proteins or *by* proteins. Proteins participate in majority, if not all, biological processes interacting with other proteins, genes, and smaller molecules. Within a biological system, all these data (starting from DNA transcription and ending at protein regulating biochemical reactions in a living cell) are processed, integrated, and executed through a complex network of interactions.

Various relationships between proteins, such as physical interactions, regulatory and metabolic pathways, similarity in sequence motifs, gene expression profiles, cellular localization etc., define *Protein Interaction Networks*. In Protein-Protein Interaction Networks, proteins correspond to nodes and relationships between proteins correspond to edges, embedding a complex net of functional and regulatory dependencies. These networks, vibrant in nature and diverse in regulatory patterns, are a complex system with highly non-linear relationships and rapid dynamics. Such networks have attracted attention of a diverse clan of scientific communities because of their structure, complexity, and methodology, broadly applicable to biochemical, metabolic, phylogenetic, financial, internet, and social networks.

Protein-protein networks exhibit a rich variety of topological structures and dependencies, which include hierarchical structures, overlapping and non-overlapping communities, multiple types of edges, causal relationships between regulators and "regulatees", etc. These structures and dependencies are poorly understood, insufficiently characterized, and become even more challenging if considered over various time scales. With the increasing interest in network topologies and relations, it is likely that the understanding of these structures will produce a shift in insight on network organization and causal hierarchies and relationships of the networks.

However, the complexity in biological networks is often approached in a static and time-invariant manner. Such approaches describe network relationships only at an instant during the evolution of a rather complex system. For example, protein-protein interaction networks, as a rule, are treated as static, without taking into account rapidly changing regulatory mechanisms as well as acquired

evolutionary relationships between proteins. Such time-dependent forces in the protein-protein interaction networks can entirely re-define network topology leading to completely different functional relationships between proteins at various times and states of the system. These time-dependent topological and functional changes in the network can be crucial for identifying malfunctioning regulatory pathways at different disease stages or extreme cell conditions.

At the same time, if we look at one of the most widely used bio-data, namely, gene expression profiles, relationships between genes are commonly analyzed with Pearson correlation coefficient of their activities, which ignores temporal relationships and averages over rapidly changing regulations between them. Nevertheless, gene expression regulatory patterns exhibit a far more complex behavior, which evolves over time, and may completely re-define regulator/regulatee relationships between genes at each time step. In fact, it is possible for a single gene to regulate multiple genes, or for a group of genes to regulate one gene. Very often a gene-regulator can only be active in a presence of another gene/protein etc., and the regulated gene can become a regulator at the next time step. Time-dependent regulator/regulated patterns define a long and complex cascade of regulatory relationships, which can be used for controlling cell's response to starvation, growth, antibiotics, immune response, and disease progression (such as cancer, autism, atherosclerosis, etc).

The goal of this thesis is to understand the structure and behavior of biological networks, over multiple time scales: starting with deep evolutionary time-scales connecting species and populations, but then moving on to generation-by-generation time-scales modulating evolution-induced relationships inside single specie, and ultimately ending with fast regulatory time-scales over cell's life cy-

cle, etc. We illustrate our approaches to multiple time scales, starting with evolutionary associations between Yeast and Fly and ending with 48-time period of a Malaria life cycle into protein-protein networks, and show how our techniques identify specific structural and functional relationships of the biological system. For this purpose, we integrate ideas of algorithmic theory, artificial intelligence, Bayesian analysis, model checking, and causality analysis techniques, all still largely unexplored in biology.

Our goal is to map topology of the networks (as they dynamically evolve over time) to the biological properties of their building blocks (proteins, genes, small molecules, protein complexes etc). In particular,

1. We develop a new method to overcome limitations of noise in Yeast Two-Hybrid (Y2H) network and its interpretation. In fact, high false positive and false negative rates in the data can obscure network connectivity and make search for highly connected groups of proteins (termed, protein complexes), a challenge. We propose a new combinatorial algorithm, which measures connectivity between proteins in Y2H network not by edges but by edge-disjoint paths, which potentially reflects pathway evolution within a single species network. This algorithm proves to be robust against increasing false positives and false negatives, as measured with variation of information and separation.

2. We devise a new way to incorporate evolutionary information to significantly improve classification of proteins, especially those isolated in their own network and surrounded by poorly characterized neighborhood. In our method, the networks of two (or more) species are joined with edges

of high sequence similarity so that proteins-homologs of different species can exchange information (as is done by message passing algorithms for graphical models) and acquire new and improved functional associations.

3. Finally, we focus on understanding the role of time (both at deep evolutionary scale as well as faster reaction-time scale) by approaching it at multiple spatio-temporal scales. We consider malaria parasite (*P. falciaprum*) in a single cell scale, during its inter-erythrocytic developmental cycle (IDC). Understanding time-course data is crucial for defining regulatory relationships between genes. In fact, parasite's genes convey a complex pattern of adjusting gene expression and rapidly changing regulatory relationships as inter-erythrocytic developmental cycle evolves. We develop a novel approach to connect time-course gene expression profiles to assign functions to un-annotated Malaria proteins and predict potential target for vaccine development.

To summarize, this thesis aims to introduce novel robust methods to define and improve protein and gene classifications, time-dependent relationships, and causal time-depended inferences in protein-protein networks. Such advances can have a tremendous future impact on many biomedical applications: personalized medicine, drug and vaccine discovery, designing tools for diagnostics and therapeutic interventions, etc.

1.2 Thesis Outline

The thesis is organized as follows. In Chapter 1 we present the algorithm for the detection of protein complexes from noisy Yeast-Two Hybrid experiments by measuring network connectivity not with edges but with edge-disjoint paths. In Chapter 2 we present a probabilistic graphical method which connects two networks of different but related species with links of high homology to improve protein classification of isolated or sparsely-connected proteins. In Chapter 3 we address an important issue of protein function prediction for malaria parasite and show that dynamic data, such as time-course gene expression profiles, can have a crucial effect on biological process classification of malaria proteins.

Chapter 2

Predicting Protein Complexes and Functional Modules from Noisy Data by using Gomory-Hu trees

2.1 Introduction

Two-Hybrid (Y2H) Protein-Protein interaction (PPI) data suffer from high False Positive and False Negative rates, thus making searching for protein complexes in PPI networks a challenge. To overcome these limitations, we propose an efficient approach which measures connectivity between proteins not by edges, but by edge-disjoint paths. We model the number of edge-disjoint paths in terms of a network flow problem and efficiently represent it in a Gomory-Hu tree. By manipulating the tree, we are able to isolate groups of nodes sharing more edge-disjoint paths with each other than with the rest of the network, which are our putative protein complexes. We examine the performance of our algorithm with

Variation of Information and Separation measures and show that it belongs to a group of techniques which are robust against increased false positive and false negative rates. We apply our approach to Yeast , Mouse, Worm, and Human Y2H PPI networks, where it shows promising results. We demonstrate that our algorithm outperforms previously described methods in the quality of produced clusters. On Yeast network, we identify 38 statistically significant protein clusters, 20 of which correspond to protein complexes and 16 to functional modules of proteins.

2.1.1 Motivation

We wish to propose a new efficient and robust algorithm to infer protein complexes correctly from Y2H experiments. If the protein-protein interaction data were flaw-less and error free, then a fairly direct graph-theoretic algorithm working on graphs whose edges represent pair-wise interactions would have sufficed. The intuitively direct algorithms (e.g., clique detection, clustering or density-based methods) tend to be efficient, and work reasonably well with small number of errors that mislabel the edges falsely (both false positive and negative, errors). Our challenge is to devise more sophisticated algorithms that enjoy a comparable computational efficiency, and yet work robustly as the quality of the experimental data degrade substantially, as is common with practically all currently available PPI data. The fundamental conceptual innovation in our algorithm is to analyze structure of the graphs through their collections of edge-disjoint paths that remain relatively immune to the corrupting noises in the experiment, and yet lead to an efficient implementation through Gomory-Hu tree representations. Below,

we further elaborate on these points.

Complexes of proteins are at the heart of many fundamental biological processes, including e.g. RNA metabolism, signal transduction, energy metabolism, and translation initiation. As noted, the process of efficiently purifying [2, 3] protein complexes and identifying their structure and function has remained a challenge. The most common experimental techniques result in the yeast two-hybrid (Y2H) protein-protein interaction (PPI) networks, which encode pair-wise interactions between proteins, and thus hold the promise to yield information about large-scale phenomena such as participation in protein complexes, as examined in [4–8]. It has been a well-known problem that Y2H experiments suffer from noise inherent in the experiments. To overcome these limitations, one needs algorithmic approaches robust against high FP and FN rates. Thus, even when the details of protein complexes become "disguised" by false negatives or become intertwined with each other by false positives, these algorithms could exploit the fact that proteins within complexes still remain connected by adequately many paths in the network. However, this qualitative statement requires a quantitative justification, namely, as the number of false edges (positive or negative) increases, how and when do these algorithms break down? What is the nature of the algorithmic degradation: slow and graceful, or sudden and catastrophic? What is the best algorithmic framework, in which they could be studied? Our main results are as follows:

Algorithmic Results: We devise and implement a novel algorithm based on max-flow and their representations through the classical Gomory-Hu tree data structures. We perform both theoretical and practical complexity analysis. We describe and conduct its performance and robustness analysis with respect to practical data

using Meila's variational information [9] and Separation measure [10].

Robustnes Results: We study the nature of the computational robustness of our algorithm through extensive simulation studies. We propose a prior model consisting of a family of complexes whose sizes are random i.i.d., but distributed according to a power-law. We study the validity of our computational analysis as the underlying graphs evolve incorporating various forms of experimental error.

Experimental results: We consider *Saccharomyces Cerevisiae* as a *model* organism for our study, since its Y2H network as well as its protein complex data are most complete. Data for protein Y2H pairwise interactions and protein complexes were taken from the BIOGRID [11] and MIPS [12] databases. On Yeast network, we identify 38 protein clusters which show p-value $< 10^{-4}$ of being found at random. Among them there are 20 protein complexes and 16 functional modules. Identified protein complexes cover 61% of all existing MIPS complexes, which have sufficient data coverage (or 72% of non-broken complexes, see Section 3.3).

2.1.2 Related work and overview

The Y2H experiments are known for their high false positive and false negative rates: two adjacent proteins might not belong to the same protein complex (False Positives; Figure 2.1 **b**) as well as proteins from the same complex might not share an edge (False Negatives; Figure 2.1 **a**). These phenomena raise questions about the validity of the direct statistical examination of pure Y2H networks.

With current data coverage and high false negative rates, protein complexes of the Y2H PPI networks suffer from low connectivity within complexes. Among all existing Y2H edges, only 6.14% connect protein pairs which participate in the

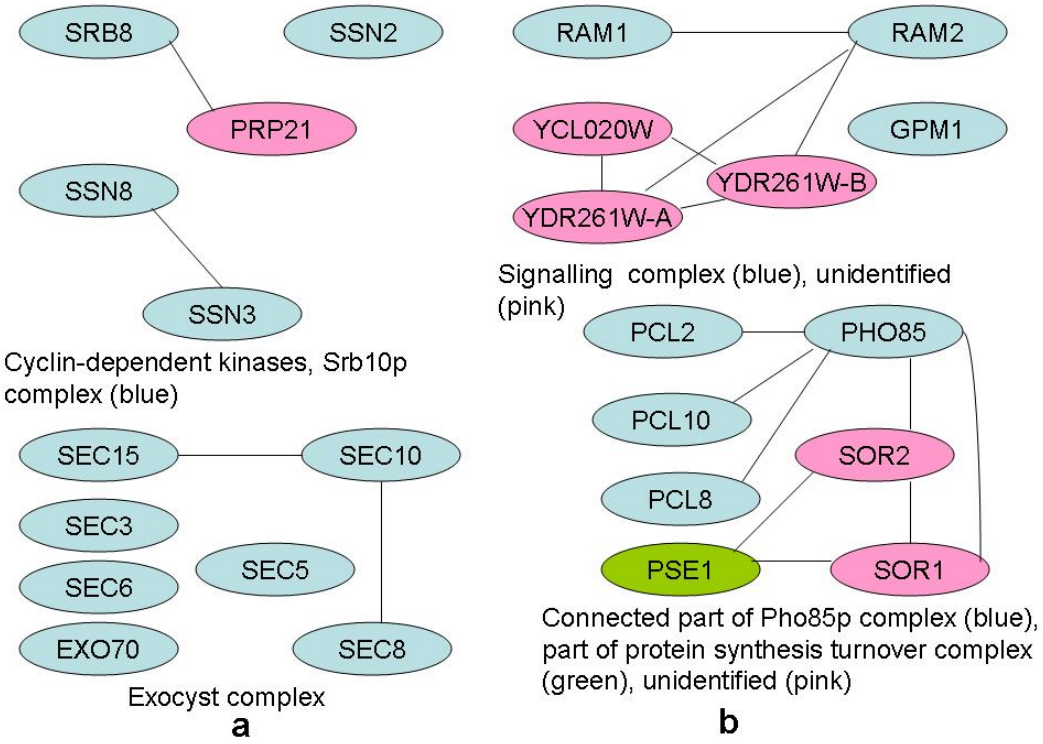


Figure 2.1: **(a)** Protein complexes that have low Y2H connectivity. **(b)** Protein complexes with "fused" out-of-complex proteins.

same protein complex. In fact, there are 788 protein complexes (from BIOGRID and MIPS) with at least 3 nodes. Of those, 463 do not have a single Y2H edge in the complex, 129 have only one Y2H edge, and 71 have two edges. There are only 125 complexes which contain at least three Y2H edges in the complex and can *potentially* have a minimum level of connectivity necessary to be identified by a connectivity-based computational method.

The majority of graph-based methods for extracting protein complex information look for densely connected, clique-like regions of the PPI network [4,5,7,8]. However, the problem of noise in yeast two-hybrid experiments required these methods to supplement pair-wise interaction data with other biological markers, such as co-expression [6], functional annotation [5], small-scale immunoprecipitation [8] , microarrays [13], or inter-specie PPI data for conserved protein complexes [4, 14].

If a protein complex corresponds to a clique-like subgraph in the Y2H PPI graph, then increased FP and FN rates might at least interfere with and at most preclude the search for such structures. For example, as shown in Figure 2.1 **b** high false positive rate in the Y2H data can produce areas of "false" density or increase the connectivity between protein complexes making them impossible to be identified in the network. At the same time, high false negative rate can disguise clique-like protein complexes. Nonetheless, note that even if proteins from a complex lose a few edges, they should maintain their association and still be connected by enough paths in the network.

If we take a *path* between two proteins as evidence that they are in the same complex, then the number of *edge-disjoint paths* is related to the degree of confidence we have of complex co-membership (observe that an edge is also a path,

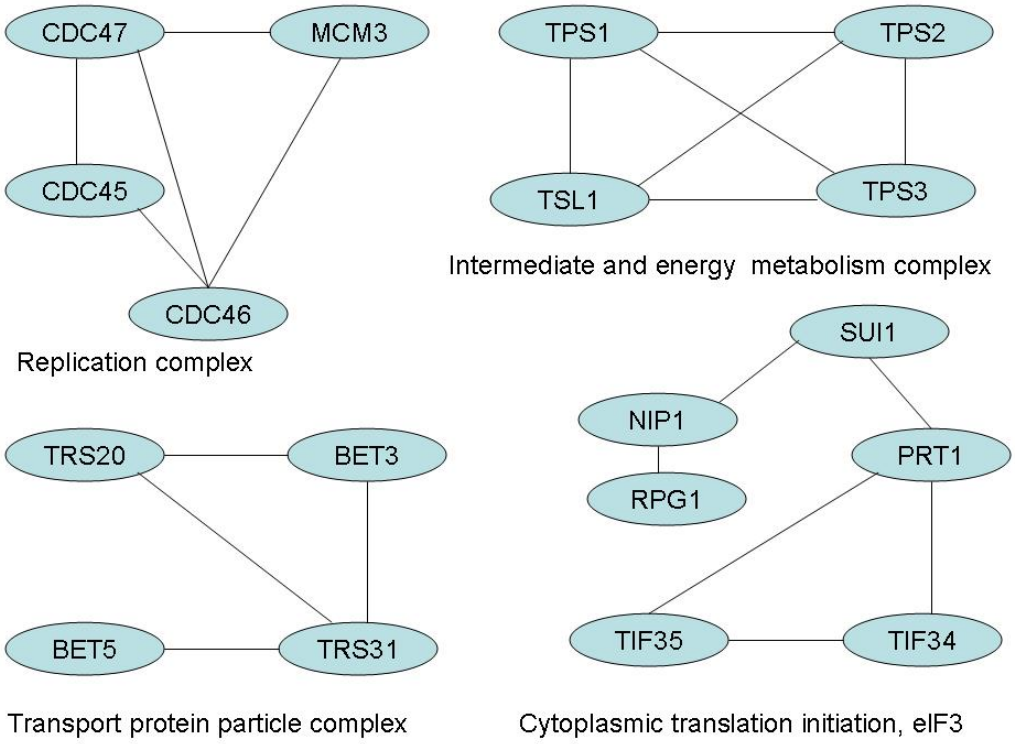


Figure 2.2: Examples of protein complexes that contain 2-edge connected subgraphs.

but a path is not limited to an edge). We say that a pair of nodes with only one edge-disjoint path between them is *weakly linked*, and proteins with at least two edge-disjoint paths are *strongly linked*. The high false-positive rate of Y2H experiments suggests that we should limit ourselves to finding strongly-linked complexes. Moreover, all proteins sharing one edge-disjoint path would simply belong to the same connected component.

Consequently, we examined the Yeast Y2H PPI network with respect to the number of edge-disjoint paths for pairs of proteins that belong to the same protein complex (in-complex group) versus pairs of protein that do not belong to the same protein complex (non-complex group), thus covering all possible protein pairs. In Figure 2.3, we show an example of a distribution of edge-disjoint paths in each group: it is more common for non-complex group to share just one edge-disjoint path, compared to the in-complex group. At the same time, in-complex group shows a clear evidence of sharing two and more edge-disjoint paths compared to the non-complex group. Overall, the proportion of protein pairs sharing one path versus sharing more than one path for the in-complex group is 1.059 and for non-complex group is 2.868, emphasizing the importance of the greater number of edge-disjoint paths for proteins from the same complex. For pairs of proteins that *do not share* an edge the same dynamics is observed: the above proportion is 1.081 for in-complex group and 2.873 for non-complex group.

Our ultimate goal is to find the number of edge-disjoint path between all pairs of nodes in the network and then combine this information to search for groups of proteins sharing more edge-disjoint paths among themselves than they share with the rest of the network. In our study, we consider all edges in the network as being unweighted and undirected. In such a network, the number of edge-disjoint paths

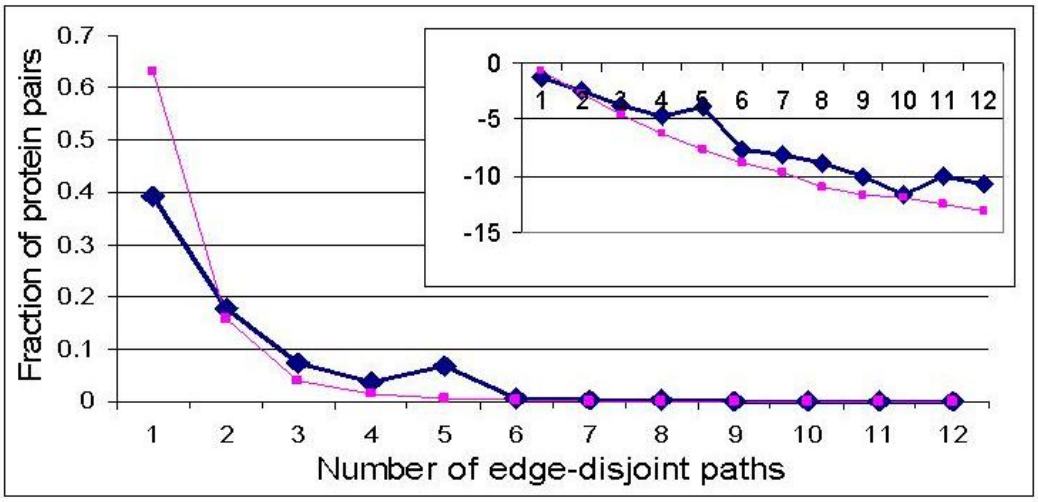


Figure 2.3: Distribution of edge-disjoint paths for protein pairs that belong to the same protein complex (shown in blue) vs pairs that do not belong to the same protein complex (shown in pink). Embedded chart shows the same distribution in log_2 scale.

between a pair of nodes corresponds to the value of the *maximum flow* between that pair. However, there is no need to consider all $\binom{n}{2}$ pairs of nodes in the network, since the number of edge-disjoint paths (or maximum flow value) for all pairs of nodes can be calculated in only $n-1$ steps and succinctly represented in a *Gomory-Hu tree* [15], as detailed below. We use the Gomory-Hu tree to partition the graph into components within which proteins share more edge-disjoint paths (flow) within themselves than with the rest of the network. By rapid elimination of low-weighted edges of the Gomory-Hu tree, we are able to identify clusters of proteins which represent protein complexes and other functional modules.

2.2 Gomory-Hu tree: Background

The reader familiar with Gomory-Hu tree and maxflow/min-cut fundamentals [15–19] may skip directly to section 2.3.

Let $G = (P, E)$ be a *protein-protein interaction network*, where P is a set of proteins and there is an unweighted undirected edge $e_{\alpha,\beta} \in E$ iff there exists Y2H interaction between proteins p_α and p_β. In our case, the absence of weights on edges is equivalent to assigning a weight of "1" to each existing edge $w(e) = 1$, $e \in E$.

A cut in $G = (P, E)$ is defined by a partition of P into two disjoint sets P^1 and P^2 and consists of all edges $E' \in E$ which have one vertex in P^1 and one vertex in P^2. The weight of the cut is defined as $W(E') = \sum_{e \in E'} w(e)$. In the case of unweighted and undirected graph, $W(E')$ would correspond to the number of edges $e \in E'$ contributing to a cut.

The problem addressed in this work is to find a minimum cost cut (or, equivalently, a cut consisting of a minimum number of edges) separating two nodes p_α and p_β. This problem is the dual of a maximum flow problem, which is solvable in polynomial time.

The generalization of min-cut problem, which is NP-hard, is a so-called *Multiway cut*. In the Multiway cut problem, we are given a set of terminals $Q = \{q_1, q_2, \ldots, q_k\} \in P$. The objective is to find a minimum-weight set of edges $E' in E$ whose removal separates each pair of terminals. Although the Multiway cut problem is equivalent to minimum cut and thus polynomial time solvable when $k = 2$, it becomes NP-hard for any fixed $k > 2$.

On the other hand, a maximum flow problem tries to maximize the flow that

can be pushed from p_α and p_β in the network G. Once some capacity is pushed, it is subtracted from the edges' weights. In other words, if all edges have weight 1, then the max flow of 1 can be pushed along one path from p_α to p_β. If there are more paths from p_α to p_β, then it is possible to push more flow etc. In an undirected unweighted graph a maximum flow between p_α and p_β is equivalent to the number of edge-disjoint paths between p_α and p_β in G.

Max flow problem is known to be a dual of min cut problem. In fact, the maximum flow between p_α and p_β equals to the minimum cut that separates them. The max flow/min cut theorem states that in a flow network, the maximum amount of flow passing from p_α to p_β is equal to the maximum capacity that needs to be removed from the network so that no flow can pass from p_α to p_β.

We are interested in max flow/min cut values for all pairs of nodes in G. In fact, we can consider all $\binom{n}{2}$ pairs of nodes and collect max-flow (min-cut) values for all pairs in a $|P| \times |P|$ table. However, it is possible to calculate max-flow/min-cut only $|P| - 1$ times by contracting a Gomory-Hu tree, also called a Flow Equivalent Network to G.

A *Gomory-Hu tree* for $G = (P, E)$ is a tree T on the same set of vertices P. The edges of T are not necessarily in the edge set E and have a new weight function W' associated with them. In addition, each edge $e \in T$ which partitions T into components S and $T - S$ is said to represent the cut associated with separating S and $T - S$ in G.

The conditions for a Gomory-Hu tree follow:

- For every pair of vertices p_α and p_β, the weight of minimum cut between p_α and p_β is the same in both G and T.

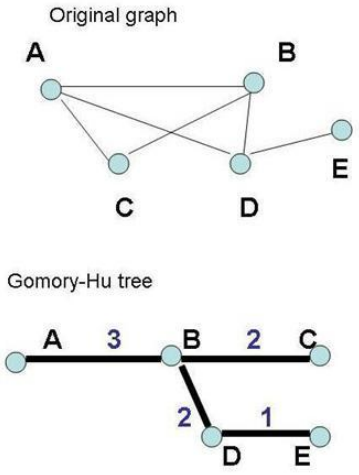
- For each edge $e \in T$, $w'(e)$ is the weight of the cut represented by e in G.

In order to construct a Gomory-Hu tree for a graph G, we first choose two nodes p_α and p_β and calculate a minimum cut (maximum flow) between them. This would divide P into two sets P^1 containing p_α and P^2 containing p_β. We might think of P^1 and P^2 as two "super-nodes" in T with an edge $e_{\alpha,\beta}$ between them representing the minimum cut (maximum flow) we just calculated. Then, the next pair of nodes is chosen so that both of them belong to either P^1 or to P^2, and so on. Performing this step $|P| - 1$ times results in a tree in which each node in T represents a single vertex from G. The algorithm ends when the $|P| - 1$ links are constructed. This final tree satisfies the properties of the Gomory-Hu tree.

2.3 Methods

We begin by computing a Gomory-Hu tree for each connected component of the PPI graph. If we consider a graph with n nodes, a table of $\binom{n}{2}$ pairwise max-flow values is a cumbersome way to represent the connectivity information of a graph. Gomory and Hu [15] noticed that only $n - 1$ distinct max-flow values are possible in any graph, and these can be represented in a so-called *Gomory-Hu Tree*. A Gomory-Hu tree is a weighted tree that spans the nodes of a graph such that the max-flow between any two nodes in the graph is the same as the max-flow between those nodes in the tree. That is, the max-flow from p_α to p_β in the network has value equal to the minimum edge on the path between p_α and p_β in the Gomory-Hu tree, as shown in Figure 2.4.

To compute max-flow value, we may use the Ford-Fulkerson method [19]. In particular, the best known deterministic max-flow algorithm for the undirected



Max-flow values for each pair of nodes

	A	B	C	D	E
A					
B	3				
C	2	2			
D	2	2	2		
E	1	1	1	1	

Figure 2.4: Gomory-Hu tree and its matrix representation. Nodes of the tree represent proteins and weighted edges of the tree represent max-flow values between proteins. The max-flow value between any pair of nodes in the Gomory-Hu tree corresponds to max-flow value between this pair of nodes in the graph.

unweighted graph is one proposed by Matula [20] and Nagamochi and Ibaraki [21] that runs in $O(|P||E|)$ steps (where $|P|$ is the number of nodes/proteins and $|E|$ is the number of edges). Thus, the time complexity of our algorithm is bounded by $O(|P|^2|E|)$ in the worst time.

First we remove minimum-weighted edges from the Gomory-Hu tree. Removing an edge induces a bipartition between the nodes of the tree. Thus an edge in the Gomory-Hu tree corresponds to an edge-cut in the PPI graph. After such elimination we *recompute* a Gomory-Hu tree for each induced connected component, since the forest obtained by removing edges (of weight > 1) from the Gomory-Hu tree is no longer the Gomory-Hu trees of the partitions. Consider the example shown in the Figure 2.5 **B**. Nodes E and H have a max-flow 4 between them in the Gomory-Hu tree. However, after eliminating edges of weight 2 from the tree, the max-flow value between nodes E and H in the residual network is no

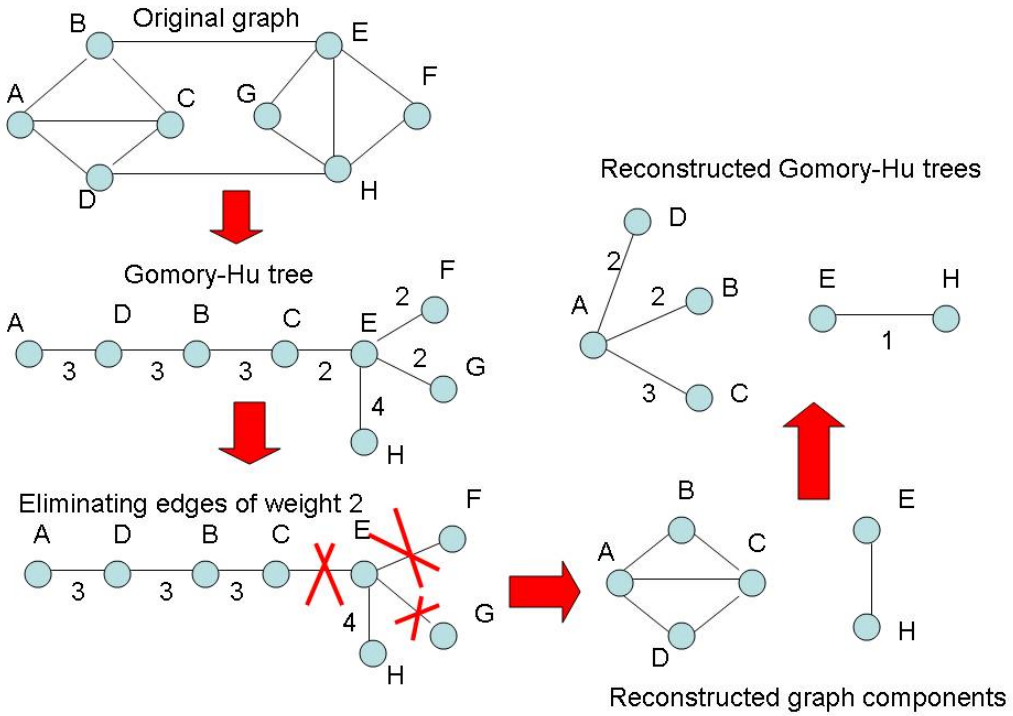


Figure 2.5: Cutting the small-weight edges of a Gomory-Hu tree to induce a partition on the nodes.

longer 4, but 1. We proceed recursively, by eliminating least weighted edges and recomputing Gomory-Hu tree for each induced connected component until there are no more edges to eliminate.

We call the set of nodes in each connected component of the Gomory-Hu forest a **cluster**. We eliminate singleton nodes at each phase and we say that a cluster with a single node *disappears*. With every elimination phase, each Gomory-Hu tree becomes smaller, splitting clusters or reducing their size until each cluster disappears. Clusters found this way are then subjected to further selection according to criteria of statistical significance, as described in 2.4.2.

The formal description of the algorithm follows: Let $G = (P, E)$ be a *protein-*

protein interaction network, where P is a set of proteins and there is an edge $e_{\alpha,\beta} \in E$ iff there exists Y2H interaction between proteins p_α and p_β. Let G^i be the graph obtained after i phases of our algorithm. We designate the jth connected component of G^i by $G^{i,j}$. Consider any weighted forest $T^i = (P, E_T, W_T)$ spanning the protein set P, where $W_T^{'}$ represents max-flow values on the edges produced by the Gomory-Hu calculations on each connected component. Then T_k^i is the forest obtained from T^i by eliminating all edges of weight at most k. As before, $T_k^{i,j}$ is the jth connected component of T_k^i. Let us define the procedure that takes X as an input and produces Y as an output as $X \longrightarrow Y$. Let $i = 0$ and $k = min(w_{\alpha,\beta} : e_{\alpha,\beta} \in E_{T^i})$:

1. $\{G^{i,1}, \ldots G^{i,x}\} \longrightarrow \{T^{i,1}, \ldots, T^{i,x}\}$. /* During the ith phase, a Gomory-Hu tree is computed for each connected component in G^i. */

2. Let k_i be the minimum weight of an edge in T^i. $\{T^{i,1}, \ldots, T^{i,x}\} \longrightarrow \{T_k^{i,1}, \ldots, T_k^{i,x'}\}$. /* We eliminate all minimum-weight edges from the Gomory-Hu trees. */

3. *Output:* $P_{T^{i,j}}$ for $j = 1, \ldots, x'$.

4. $\{T_k^{i,1}, \ldots, T_k^{i,x'}\} \longrightarrow \{G^{i+1,1}, \ldots, G^{i+1,x'}\}$. /* Eliminating an edge from a Gomory-Hu tree corresponds to eliminating all edges of some cut in the graph. */

5. $i = i + 1$.

6. go to step 1, unless $E_{G^i} = \emptyset$.

2.4 Results

2.4.1 Eliminating high degree nodes

To minimize the number of nodes and interactions that would give statistically insignificant clusters, the common practice is to limit the number of non-selective interactions (possibly false positives) and eliminate "excessive-degree" nodes from the Y2H PPI graph, as for example exercised in [8]. Such nodes usually induce the agglomerates (Figure 2.1, :**b**) which obscure protein complexes.

Degree	6	7	8	9	10	11	12	13	14	15	16	17
Final clusters	2	3	5	7	7	8	11	10	11	14	11	11
MIPS clusters	1	2	3	3	3	3	3	5	4	4	4	3
$P, \%$	50	67	60	43	43	37	27	50	36	29	36	27
$Q, \%$	17	33	50	50	50	50	50	83	67	67	67	50

Table 2.1: Results on the training set of the Yeast Y2H PPI network: eliminating nodes with the degrees higher than the threshold $d = 6, \ldots, 17$.

To learn the degree threshold, to determine which nodes and adjacent edges should be eliminated, we select a so-called "training set", which corresponds to about $\frac{1}{4}$ of the network (the remaining network is called a "testing set"). To choose a training set, we start with a random protein in the graph and accumulate the desired number of nodes by breadth-first search. During the learning stage, we eliminate nodes and their outgoing edges with respect to various degree thresholds, as depicted in the Table 2.1. For example, if we choose a degree threshold to be $d = 16$, then we eliminate all nodes (with corresponding edges) of degree higher than 16.

We define several performance measures, based on a quality of produced clusters, to evaluate the node/edge elimination effects. First, we calculate the percent coverage, P, of how many final clusters correspond to MIPS/BIOGRID protein complexes. Additionally, we introduce a new measure of protein complex coverage, i.e., 2-edge connectedness. In particular, the graph is 2-edge connected if there are at least two edge-disjoint paths between every pair of nodes in the graph, with some examples shown in Figure 2.2.

We found that from 125 MIPS/BIOGRID protein complexes with at least three Y2H edges , 74 (60 %) are fully or partially 2-edge connected. However, we observed these protein complexes often overlap with each other or are the subsets

of each other, producing data redundancy that can negatively influence the analysis. In fact, there are 33 not overlapping non-redundant 2-edge connected protein complexes, which we use as an additional measure in our further statistical analysis.

In particular, we measure Q, the recall rate for 2-edge connected MIPS/BIOGRID complexes, that are in the training set. For example, in our training set initially there exist six 2-edge connected non-overlapping protein complexes and Q is the fraction (in %) of these six that we identify in each run. We evaluate our training phase with respect to both criteria: the highest Q values are observed with $d = 13 - 16$. Among those, $d = 13$ executes highest P, which we consider our threshold and eliminate all nodes of degree higher than 13 from our dataset (85 nodes or 2.16 % of total network nodes).

2.4.2 Statistical Significance of Clusters

As the algorithm proceeds, many clusters of different sizes are generated. The final part of our algorithm is to estimate statistical significance of produced clusters and decide which correspond to protein complexes and functional modules.

To measure the statistical significance of the cluster, we need to account for the probability of finding such cluster in a random graph. To generate random graphs, we use the Maslov-Sneppen procedure [22], which shuffles the edges of the original Y2H PPI network so that the number of interactions for each protein in the network is preserved.

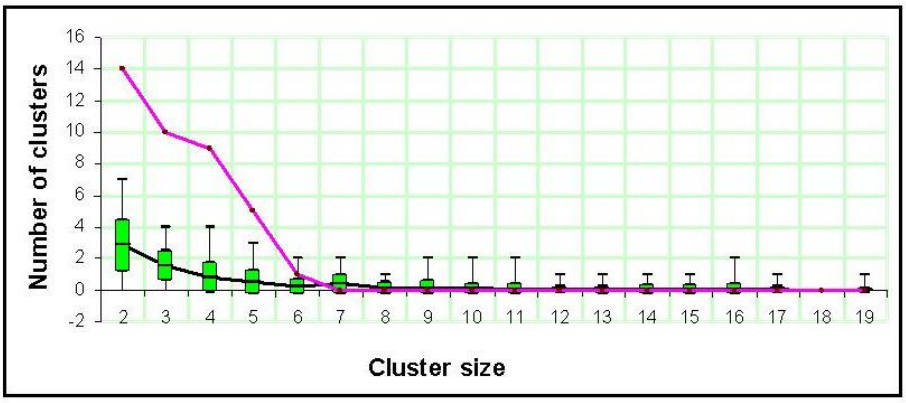


Figure 2.6: Number of clusters as a function of cluster size in the whole Yeast Y2H PPI network (red line) and in the random graphs (black line). On the lower line: rectangles represent standard deviation, with max and min as up/down bars.

Size-based p-value:

First, we calculate p-value with respect to different cluster sizes. The Figure 2.6 shows enrichment in the number of clusters of sizes 2 to 6 in the original Y2H PPI graph, as compared to results on 100 random graphs. Clusters of size 7 and higher, in contrast, appear more often at random. For each cluster of size s, we calculate p-value as a probability of finding a cluster of size s at random, fit to a normal distribution. Clusters of size $2, 3, 4$ and 5 showed p-value $p < 1 \times 10^{-4}$, which we consider statistically significant. In contrast, clusters of size 6 showed $p = 0.20$ and therefore can appear at random with reasonably high probability.

Density-based p-value:

However, we cannot draw definite conclusions just based on bases of cluster size or even on the number of edge-disjoint paths inside the cluster alone. In our method, proteins inside clusters should share more edge-disjoint paths among

themselves than with the rest of the network. Therefore, it is important to consider each cluster individually and define a variable which would reflect the above relationship. We define a "cluster-network density", *CND*, in the following way. First, we calculate the average number of edge-disjoint paths in the cluster per pair of proteins, ED_c. Then, we compute the average number of edge disjoint paths from the proteins of this cluster to the proteins in the rest of the network (ignoring proteins from different connected components), ED_r. The CND is calculated as $ED_c - ED_r$ and reflects the difference between the connectivity inside the cluster and connectivity of this cluster with the rest of the network. Our assumption is that the cluster-network density for each cluster does not assume some random value , but is a product of the unique biologically significant relationship. Of course, all clusters from the original Yeast Y2H PPI network produced by our algorithm show CND greater than 0. Here we again consider 100 random graphs generated by the Maslov-Sneppen procedure [22] and calculate p-value (fit to a normal distribution) per cluster produced. Let us define a CND of a cluster in the original PPI network as $CND*$. Then for each cluster of the original Y2H PPI network, p-value reflects the probability that CND at random would be greater or equal to $CND*$. To correct for multiple hypotheses tested , we apply Bonferroni Correction and multiply the calculated p-values by the number of hypothesis tested (in our case, it is equal to the number of observed original clusters). We consider those clusters with corrected p-values less than 1×10^{-4} as statistically significant. It appeared that clusters with p-values $< 10^{-4}$ do not violate the statistically significant sizes shown in Figure 2.6. We report 38 out of 56 clusters as being statistically significant according to the criteria described above. Among clusters with p-value $> 10^{-4}$, two represent protein complexes and two can be

considered as functional modules, see online *Supplementary material* for details.

2.4.3 Evaluation of our method using MIPS dataset of protein complexes

Our **Y2H** PPI network consists of 3930 proteins and 6219 interactions available from BIOGRID. In this study, we consider only manually curated high quality Y2H interactions, omit computationally derived interactions, ignore self interactions, and include only proteins with at least one Y2H interacting partner. After training on $\frac{1}{4}$ of the network, we run the algorithm on the remaining testing set consisting $\frac{3}{4}$ of the network and then combine (take a union of) their results.

We first examine clusters produced by our method against a MIPS database of Yeast protein complexes [12]. We consider a cluster a *match* if *all* of its proteins belong to the same MIPS/BIOGRID protein complex. However, we also evaluate the performance of our method with respect to varying matching thresholds, as discussed in section 3.4. In addition to protein complexes, we define a notion of a *Functional Module*, similarly to [8], as a group of proteins that participate in the same cellular process in the same cellular location, however not necessarily at the same time. In order for a cluster to be identified as a Functional module, its proteins should reside in the same cellular location and should share similar/relevant functions (Gene Ontology classification). Even stronger supporting evidence for Functional modules includes co-expression and literature co-citation (we analyzed our clusters against co-expression and co-citation evidence, using tables and criteria provided by [23] for pairwise log-odds scores). Since in many cases we cannot say with certainty whether proteins enter a process at different

times or at the same time, clusters from this category are strong candidates for protein complex predictions.

We present results for both the $\frac{1}{4}$ and $\frac{3}{4}$ fractions of the network in the Table 2.2. Among 38 clusters with p-value $< 10^{-4}$, there are 20 MIPS/BIOGRID protein complexes (yielding precision rate of 53%) and 18 functional modules. Five of the functional modules are supported by co-expression evidence or co-citation from the literature, thus making a strongly grounded predictions for new protein complexes, as shown in *Supplementary material*. An example of such a protein complex prediction is a cluster of 3 proteins, YBL078C(ATG8) YHR171W(ATG7) YNR007C(ATG3), where YBL078C(ATG8) is a protein essential for autophagy, YHR171W(ATG7) is a component of the autophagic system, and YNR007C(ATG3) is essential for autophagocytosis. All proteins reside in cytoplasm and are significantly co-expressed, making a strong prediction for a protein complex. We discovered a cluster of 5 proteins (listed as Functional Module): YDL140C(RPO21) YOR116C(RPO31) YOR341W(RPA190) YBR154C(RPB5) YOR224C(RPB8), four of which correspond to the 550.1.213 [12] protein complex responsible for transcription, DNA maintenance and chromatin structure. We propose to list an extra protein YDL140C(RPO21), which is located in the nucleus and regulates DNA-binding and transcription, as an additional potential member of this complex. Among 18 functional modules, two clusters had weaker evidence of forming a functional association primarily because of lack of information about the functional annotation or cellular location of participating proteins.

Additionally, several functional modules consist of proteins that have not been classified before. We anticipate making functional predictions for these proteins

based on functional annotation of other proteins in the cluster. For example, an unclassified protein YLR257W that has had no known function assigned so far, exhibits strong evidence of being a participant in transcription initiation.

For completeness, we also studied a recall rate, which corresponds to the proportion of 33 2-edge connected complexes covered. As we select our training and testing sets, five 2-edge connected protein complexes become broken (one part appears in training set while another part in testing set) and could not possibly be identified in any of the sets. As a result, there are only 28 2-edge connected MIPS/BIOGRID graphs in both sets, 20 of which we identify (yielding recall rate of 61%, or 72% if 28 non-broken 2-edge connected complexes are considered). Additionally, we characterize the performance of our method by a parameter M that is a fraction of proteins in the matched MIPS/BIOGRID cluster over proteins in the 2-edge connected part of the corresponding complex. In particular, 18 out of 20 clusters show $M = 1$.

In general, among the 33 2-edge connected complexes, there are 17 triangles, 13 4-node graphs (2 fully-connected graphs with 6 links, 10 graphs with 5 links, and 1 graph with 4 links), 2 graphs of 5 nodes (one fully-connected graph and one graph with 7 edges), and one fully-connected graph of 7 nodes. Thus, the clusters that cover 2-edge connected protein complexes are partially bounded by the above sizes. We expect these sizes to increase however as the Y2H data becomes more complete.

Sets	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4} \bigcup \frac{3}{4}$
Total clusters with $p < 10^{-4}$	10	28	38
Clusters that cover MIPS complexes	5	15	20
FM with co-location and (co-expression or co-citation	1	4	5
FM with co-location	4	7	11
FM with limited information	0	2	2

Table 2.2: Final clusters of testing and training sets in the Yeast Y2H PPI network.

2.5 Protein Complexes of Other Species

We have applied our method to other species such as Mouse, Human, and Worm. The Y2H interactions and protein complex information were obtained from the BIND database.

2.5.1 Human Protein Complexes

We applied our method to Human Y2H PPI network containing 2699 proteins and 3360 interactions. Our experiments identified four out of five 2-edge connected protein complexes. Clusters that did not cover protein complexes were tested against compatible location evidence (cellular location information for every Human protein was taken from swiss-prot [24] database), such that all of them responded positively.

Additionally, we subjected identified clusters to broader biological tests, described as hsPPIP [25]. The method in [25] uses the idea that Protein Protein Interaction Predictions are similar in different organisms. It takes Human proteins/genes of interest as input. These genes are compared to Yeast proteins/genes for possible orthologs. If such orthologs in Yeast participate in a protein complex, then the probability that Human orthologs will form a complex is greater than zero

CATEGORIES	
final clusters	31
cover BIOGRID complexes	3
$0 < PPIP < 17$	10
$18 < PPIP < 49$	3
$50 < PPIP < 89$	3
$90 < PPIP \leq 100$	3
NON-ZERO PPIP total	19
$> 50\%$ *PPIP total*	6
zero-PPIP +compatible location	9

Table 2.3: Results on Human Y2H PPI network.

(non-zero). This probability depends on the degree of orthology and the possibility of Yeast orthologous proteins to build a protein complex. Thus, when any of our clusters respond positively, they can lend some support to the hypothesis that the cluster is a complex, whereas when the prediction is low, they yield no information.

Table 2.3 shows results of an experiment on the Human Y2H PPI network with corresponding hsPPIP probabilities (higher probabilities are better). All Human PPI clusters are listed at

http://research.rutgers.edu/$\sim$ amitrofa/predictions.html.

2.5.2 Worm Protein Complexes

The Worm Y2H PPI network consists of 3154 proteins connected by 4921 edges. All identified clusters were tested against compatible cellular location evidence, taken from [24]. Unfortunately, the subcellular location information for Worm is very limited, to such an extent that we were not able to obtain this information for more than one protein in a cluster. The network, however, contains one 2-

edge connected protein complex, which we identify. All 29 predicted protein complexes can be found at
http://research.rutgers.edu/$\sim$ amitrofa/predictions.html.

2.5.3 Mouse Protein Complexes

In Mouse Y2H PPI network, containing 723 proteins and 630 edges, we identified 17 clusters, 13 of which responded to the compatible location evidence. All Mouse complex predictions are listed at
http://research.rutgers.edu/$\sim$ amitrofa/predictions.html.

2.6 Robustness via Statistical Analysis

2.6.1 Experiment design

We examine the robustness of our algorithm by its ability to recover protein complexes as we vary the number of FP and FN in a randomly constructed network. Since we think of protein complexes as highly connected "clique-like" structures in the PPI network [4, 5, 7, 8], we build our random *test graph* in the following way: we introduce complete subgraphs of size from 10 to 2 and singletons (following the power-law distribution: 10 subgraphs of size 10, 20 subgraphs of size 9, etc, 300 singletons), similarly to the approach described in [10]. These groups of nodes are our initial complexes (clustering C). At each step we delete some number of randomly chosen edges of the network (introducing false negatives) and add edges (false positives), which mirrors current Y2H PPI networks, with relatively high levels of false positive and false negative interactions. We charac-

terize these graphs in terms of parameters (k, n), which denote the fraction of the total edges modified (k = added, n = deleted). We denote clustering obtained on the modified network as C'. Our main goal is to determine how well clustering C' approximates the true clustering C (we use terms "clustering C" and "complexes" interchangeably in this section). For this purpose, we operate two statistical measures, Separation [10] and Variation of Information (VI) [9], both defined in section 2.8,

2.6.2 Robustness results

With respect to robustness, we compare our method with the technique reported as the most robust clustering on PPI networks in [10], namely, Markov Clustering (MCL) [26]. The method in [26] is based on simulation of (stochastic) flow in graphs assuming the presence of many edges between the members of the cluster. It is assumed that higher-length (longer) paths between two arbitrary nodes in the cluster is high compared to nodes from different clusters. In other words, if we impose a random walks on the graph, it will *infrequently* go from one natural cluster to another. The algorithm simulates random walks in the graph by alternation of two operators called expansion (computing random walks of higher length) and inflation (boosting the probabilities of intra-cluster walks and demoting inter-cluster walks). Eventually, iterating expansion and inflation results in the separation of the graph into different segments, interpreted as clusters.

Since the protein complexes in current PPI networks suffer from low connectivity, it is more important to examine the robustness of the algorithms against increasing FN rates. We present some results in Figure 2.7, which show that our

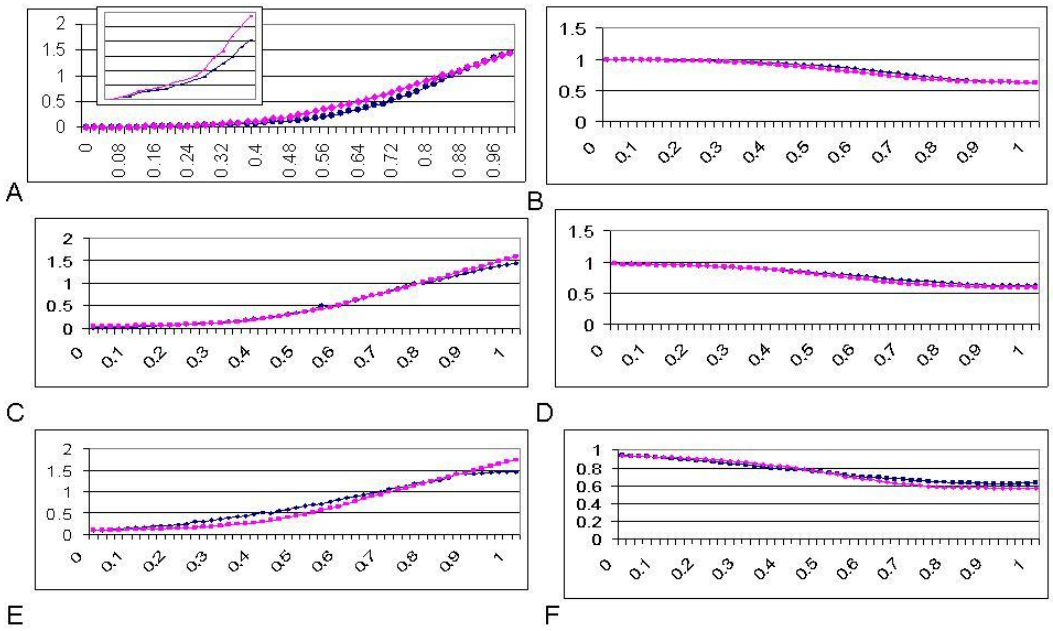


Figure 2.7: Each curve represents the value of VI (left panels) or Separation (right panels) **(A-B)** edge removal from the test graph. **(C-D)** edge removal from an altered test graph with 5% of randomly added edges. **(E-F)** edge removal from an altered test graph with 10% of randomly added edges. Lower VI and higher Separation are preferred.

approach is as much as or more robust compared to Markov clustering when examined against increased FP (by 5% and by 10%, which are most likely to exist in Y2H PPI networks) and varying FN rates. Both methods show smooth curves toward increased FN rates.

2.7 Comparative Analysis and Discussion

In this section, we compare our method to those previously described in the literature and most widely used, such as MCL [26], RNSC [5], and Spirin and Mirny method [8]. The comparison is made on the same dataset of Y2H interactions obtained from BIOGRID [11], where we considered only manually curated high

quality interactions.

For completeness, we first briefly summarise the methods in [5, 8, 26] and present their original results [1].

King et al. [5] develop the Restricted Neighborhood Search clustering algorithm, RNSC, using a cost function (i.e., cost-based local search). After generating clusters, proteins are selectively chosen from clusters using a filtering model (based on cluster density and functional homogeneity). In their study, King et al. identified 33 clusters, 22 of which matched known MIPS protein complexes by at least 90% of cluster proteins. Among them, six clusters match protein complexes by 100% of cluster proteins (20% precision) and four clusters turned out to be 2-edge connected (12% recall). We identify 35 new clusters, among which there are 17 new protein complexes, not covered by the method of King et. al.

On the other hand, Spirin and Mirny [8] look for heavily connected, clique-like groups of nodes in the PPI network. They use the union of four different clustering methods to identify 67 total clusters, 30 of which correspond to protein complexes (45% precision) and 18 contain 2-edge connected subgraphs (55% recall). However, in addition to two-hybrid interactions, they used information from other hypothesis-driven studies of protein interactions such as coprecipitation, which is excluded in our method to avoid the selection bias inherent in such small samples. We identify four new protein complexes and 13 new functional modules, not listed by the method of Spirin and Mirny.

The MCL clustering [26], described in detail in section 2.6, has not been applied to search for protein complexes yet, but only to cluster proteins based

[1] Our method identified 38 protein clusters, 20 of which correspond to protein complexes (53% precision and 61% recall).

on their sequence similarity. However, it represents an important alternative, since this is a widely used clustering technique and is regarded as a highly robust against increasing false positive and false negative rates in PPI networks, as shown in [10]; consequently, we decided to conduct as well provide an extensive comparison of our method against MCL.

For a statistically significant and biologically sound comparison, we examined the performance of our method and other computational techniques on the same dataset with respect to several statistical measures, such as clustering-wise positive predictive value (PPV), precision on various matching thresholds, and average percent coverage per cluster (ACC).

In Figure 2.8, we show how various methods change their precision values as we vary the cluster matching threshold. Precision reflects the proportion of clusters that match known complexes in at least "m" % of their proteins, with respect to the total number of clusters. For example, in our method, the fraction of clusters that match known protein complexes in at least 100% of their proteins is 0.53 (corresponding to 53%) and of those matching in at least 50% of their proteins is 0.67 (corresponding to 67%).

In Figure 2.9 we present an average percent coverage per cluster, as an averaged proportion of members of cluster j which belong to complex i, with respect to the cluster size, as formally defined in section 2.8.

Finally, Figure 2.10 shows clustering-wise positive predictive value, as a proportion of members of cluster j which belong to complex i, with respect to the total number of members of this cluster assigned to all complexes, as detailed in section 2.8. Positive predictive value is expected to be lower compared to other statistical measures since MIPS/BIOGRID protein complexes very often are re-

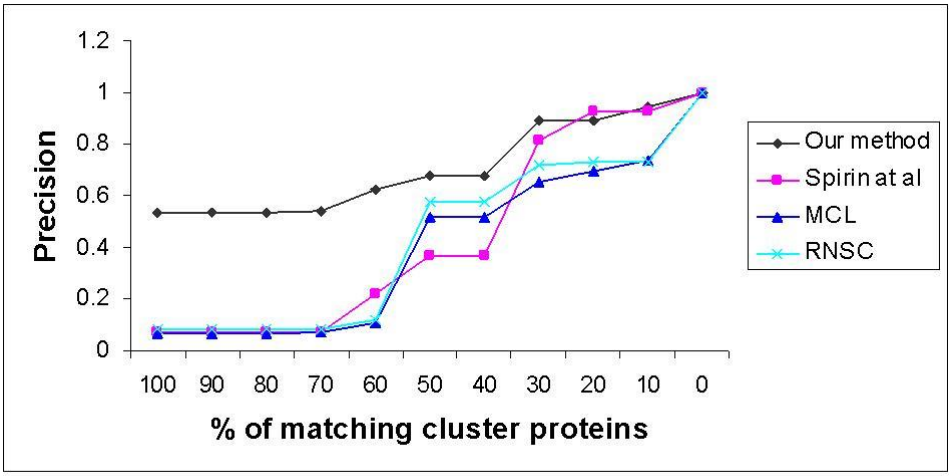


Figure 2.8: Precision at various cluster matching thresholds. A point in a graph corresponds to a fraction of clusters (y axis) that match known protein complexes in at least "m"% of their proteins (x axis).

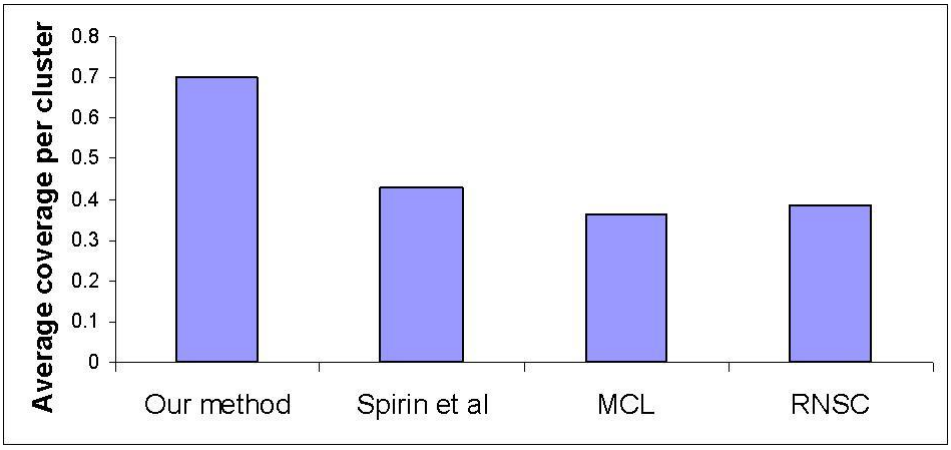


Figure 2.9: Comparative analysis of our algorithm to other methods by using average coverage per cluster.

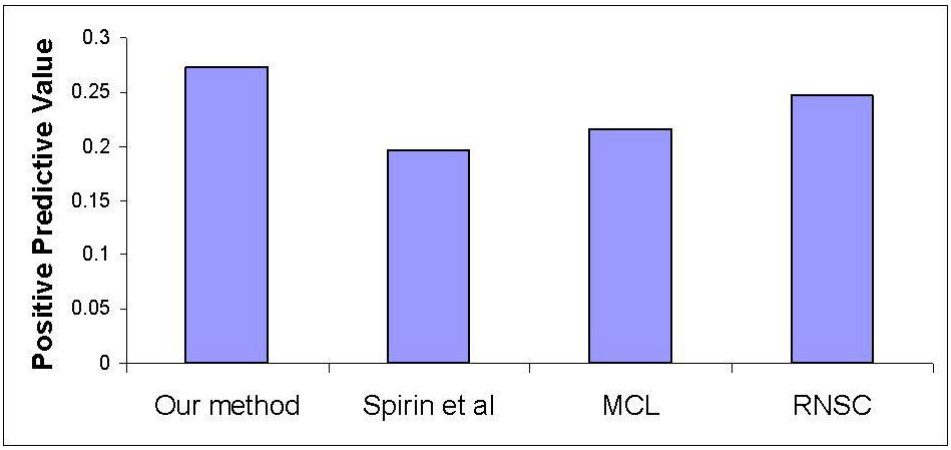


Figure 2.10: Comparative analysis of our algorithm to other methods by using positive predictive value.

dundant or show large overlaps. Our method shows better performance, compared to other computational techniques, in all three measures thus indicating clusters of higher quality and biological significance.

One promising future direction for our method would be to assign a confidence score for each Y2H interaction (i.e. conservation of the interaction across species). It is possible to define a distance-based measure between proteins and use a Diffusion Map for spectral clustering, as in [27]. However, this method is very computationally expensive and hard to scale to large datasets. We plan to explore an efficient implementation of a continuous approach of diffusion maps with discrete approach of Gomory-Hu trees.

An interesting min-cut clustering approach of Tarjan et al [28], which was applied to find communities in web and citation networks, introduced an artificial sink node connected to all other nodes. We plan to expand our understanding of competitive bounds for communities' sizes addressed in [28]. Another interesting approach described by Newman in [29] (later extended in [30]) describes graph

decomposition based on edge betweenness, defined as the number of shortest paths which go through an edge. Hartuv et al. in [31] present an algorithm based on min cut idea, which shows an improved time complexity and generates clusters with diameter 2 (two vertices are either adjacent or share one or more common neighbors). We do not require nodes in the cluster to be adjacent or to necessarily share a neighbor; however, they can be connected by much longer edge-disjoint paths.

2.8 Statistical measures

2.8.1 Separation

Separation [10] is the statistical measure equivalent to the product of the complex elements found in the cluster and the cluster elements found in the complex. Having k complexes (from clustering C) and k' clusters (from clustering C'), we define the contingency table as a $k \times k'$ matrix M where row i corresponds to the i^{th} complex and column j corresponds to the j^{th} cluster. The value of a cell $M_{i,j}$ indicates the number of proteins in common between complex i and cluster j.

From these values, we derive *relative frequencies* with respect to the marginal sums, either per row $F_{row(i,j)}$ or per column $F_{col(i,j)}$.

$$F_{row(i,j)} = \frac{M_{i,j}}{\sum_{j=1}^{k'} M_{i,j}}$$

$$F_{col(i,j)} = \frac{M_{i,j}}{\sum_{i=1}^{k} M_{i,j}}$$

The frequency per column $F_{col(i,j)}$ is equivalent to the well known Positive Predictive Value, $PPV_{i,j}$. The *Separation* between complex i and cluster j, $Sep_{i,j}$, is defined as a product of a column-wise and row-wise relative frequency.

$$Sep_{i,j} = F_{col(i,j)} \times F_{row(i,j)}$$

The value of *Separation* is between 0 and 1. The perfect value of *Separation* $Sep_{i,j} = 1$ indicates a perfect match between complex i and cluster j, i.e. when a cluster comprises all of complex's proteins and nothing more. Interestingly, *Separation* penalizes cases where proteins of a given complex are assigned to multiple clusters, by using row sums rather than complex sizes.

A complex-wise separation $Sep_{co(i)}$ is calculated as the sum of separation values for a given complex i.

$$Sep_{co(i)} = \sum_{j=1}^{k'} Sep_{i,j}$$

A cluster-wise separation $Sep_{cl(j)}$ is calculated as a the sum of separation values for a given cluster j.

$$Sep_{cl(j)} = \sum_{i=1}^{k} Sep_{i,j}$$

A complex-wise separation over all complexes Sep_{co} is calculated as the average of $Sep_{co(i)}$,

$$Sep_{co} = \frac{\sum_{i=1}^{k} Sep_{co(i)}}{k}$$

Respectively, the cluster-wise separation over all clusters is

$$Sep_{cl} = \frac{\sum_{j=1}^{k'} Sep_{cl(j)}}{k'}$$

We then compute and operate the *geometrical separation* Sep as the geometric mean of

$$Sep = \sqrt{Sep_{co} \times Sep_{cl}}$$

2.8.2 Variation of information

Variation of information another useful metric based information-theoretic criterion that measures how much information is lost or gained in going form clustering C to C'. In our case, C corresponds to the initial complexes. If we let n be the number of nodes and K be the total number of clusters, with n_k being a size of cluster C_k, then the uncertainty (or entropy) of the clustering C is defined as

$$H(C) = -\sum_{k=1}^{K} P(k) \log(P(k)),$$

where $P(k) = n_k/n$. The joint distribution that a point belongs to cluster C_k in C and to cluster $C'_{k'}$ in C' is

$$P(k, k') = \frac{|C_k \bigcap C'_{k'}|}{n}$$

Then the mutual information between clustering C and C' is

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log(\frac{P(k, k')}{P(k)P'(k')})$$

And finally, the Variation of Information is defined as

$$VI(C, C') = H(C) - I(C, C') + H(C') - I(C, C')$$

Higher Variation of Information corresponds to bigger deviation from the original clustering C.

2.8.3 Average cluster coverage

Operating the contingency table M, for each cluster j, we find its best complex matching complex i. Then the number of matched proteins is divided by the cluster size, basically indicating the % of cluster proteins which are matched.

$$ACC_j = max_{i=1}^{k} \frac{M_{i,j}}{|cl_j|}$$

These values for each cluster are then averaged (summed and divided by the total number of clusters).

$$ACC = \frac{\sum_{j=1}^{k'} ACC_j}{k'}$$

2.8.4 Positive predictive value

Operating the contingency table M defined above, Positive predictive value is defined as

$$PPV_{i,j} = \frac{M_{i,j}}{\sum_{i=1}^{k} M_{i,j}} = \frac{M_{i,j}}{M_{.j}}$$

$M_{.j}$ is a marginal sum of a column j. The cluster-wise positive predictive value PPV_{cl_j} reflects the reliability with which cluster j predicts that a protein belongs to its best matching complex.

$$PPV_{cl_j} = max_{i=1}^{k} PPV_{i,j}$$

To characterize the general PPV of a clustering result as a whole, we present a clustering-wise PPV as a the weighted average of a PPV_{cl_j} over all clusters.

$$PPV = \frac{\sum_{j=1}^{k'} M_{.j} PPV_{cl_j}}{\sum_{j=1}^{k'} M_{.j}}$$

2.9 Conclusions

We have presented an efficient algorithm for identifying protein complexes through efficient manipulation of the Gomory-Hu tree of the PPI Y2H network. Our method is shown to be robust against high FP and FN rates and capable of producing clusters of high quality when compared to other approaches. Additionally, the algorithm shows a good recall for identifying existing MIPS protein complexes with sufficient data coverage. Identified Functional modules are strong candidates for complex predictions and constitute reliable material for experimental research.

2.10 Web Resources and Supplementary material

Supplementary and output data are available from

`http://www.cims.nyu.edu/~antonina/predictions.html`

Chapter 3

Prediction of Protein Functions with Gene Ontology and Inter-Species Protein Homology Data

3.1 Introduction

Accurate computational prediction of protein functions increasingly relies on network-inspired models for the protein function transfer. This task can become challenging for proteins isolated in their own network or those with poor or uncharacterized neighborhoods. Here, we present a novel probabilistic chain-graph based approach for predicting protein functions that builds on connecting networks of two (or more) different species by links of high inter-species sequence homology. In this way, proteins are able to "exchange" functional information with their neighbors-homologs from a different species. The knowledge of inter-species relationships, such as the sequence homology, can become crucial in cases

of limited information from other sources of data, including the protein-protein interactions or cellular locations of proteins. We further enhance our model to account for the Gene Ontology dependencies by linking multiple but related functional ontology categories within and across multiple species. The resulting networks are of significantly higher complexity than most traditional protein network models. We comprehensively benchmark our method by applying it to two largest protein networks, the Yeast and the Fly. The joint Fly-Yeast network provides substantial improvements in precision, accuracy, and false positive rate over networks that consider either of the sources in isolation. At the same time, the new model retains the computational efficiency similar to that of the simpler networks.

3.1.1 Motivation

In protein-protein networks, each node represents a protein and edges between nodes represent different types of functional associations, such as protein-protein interactions, sequence similarity, co-expression patterns, and others. Majority of computational methods for protein classification rely on the property that close neighbors in a protein-protein network typically share a function [1,32–38]. These methods assign the function (or functions) to a protein of interest based on the annotations of its neighbors. Such approaches have shown success in cases where proteins have multiple, mostly annotated neighbors. However, the methods display much less success on proteins with insufficient neighborhoods: those proteins isolated in their own network or the ones surrounded by poorly annotated neighbors.

In this work we propose a novel approach to protein function prediction,

which overcomes these limitations and incorporates inter-species evolutionary information with multi-functional Gene Ontology (GO) dependencies. The fundamental conceptual innovation of our method is to connect protein-protein networks of two (or more) different, but related species, into a single computational model. Through the edges of high homology, proteins are able to expand their learning neighborhood and acquire additional functional information from their neighbors-homologs of a different species network.

Our new approach relies on a chain-graph probabilistic approach to integrate multiple sources of information: protein-protein interactions, multi-functional ontology information, intra-species sequence similarity, and inter-species homology which captures evolutionary relationships between species. In connecting networks, we rely on the fact that proteins of different species, which share high sequence similarity, are likely to share similar protein classification. In most cases such proteins, orthologs, had established functions before the speciation event. Thus, high similarity of sequences between species is likely to lead to shared functions. Even though the resulting large chain-graphs can suffer from increased time and space complexity of the models, compounded by the added complexity of the multi-species network, we show that the combined models often lead to efficient implementations and significant improvements in predictive accuracy not observed in isolated networks or other competing approaches.

The rest of the chapter is organized as follows. In Section 3.1.2 we first present an overview of the closely related network approaches to protein function prediction. We then introduce, in Section 3.2, a chain-graph based probabilistic network model that combines both the GO structure and the information from protein-protein networks of multiple species. Section 3.3 demonstrates the effectiveness

of the proposed approach when applied to large Fly and Yeast networks, at different granularities of the GO. We finally discuss the new results in Section 3.4 and relate them to the performance of related state-of-the-art probabilistic network models.

3.1.2 Related Work

Proteins are involved in many if not all biological processes, such as energy and RNA metabolism, signal transduction, and translation initiation. However, for a large portion of proteins, their biological function remains unknown or incomplete. Thus, constructing efficient and reliable models for predicting protein functions has thus become the task of immense importance.

A critical factor that impacts performance of network models is the choice of functional association between proteins. The most established methods for protein function prediction are based on sequence similarity (e.g., a BLAST score). A large set of methods relies on the fact that similar proteins are likely to share common functions, subcellular location or protein-protein interactions (PPIs). Such similarity-based methods include sequence homology, similarity in short signaling motifs, amino acid composition and expression data [35, 36, 38–41].

Using PPI data to ascertain protein function within a network has been studied extensively. For example, methods in [32, 42, 43] used the PPI to define a Markov Random Field over the entire set of proteins. These methods are based on the notion that interacting neighbors in PPI networks should share a function [32–34].

One promising computational approach to protein function prediction utilizes the family of probabilistic graphical models, such as belief networks, to infer

functions over sets of partially annotated proteins [32, 42, 43]. Using only a partial knowledge of functional annotations, probabilistic inference is employed to discover other proteins' unknown functions by passing on and accumulating uncertain information over large sets of associated proteins while taking into account different strengths of associations.

Several related studies used various probabilistic frameworks to infer functions of proteins [44–48]. For example, the method in [46] used multiple Support Vector Machines for the classification of protein predictions using protein sequences of several organisms for training. GOtcha approach developed in [48] and method in [47] search for similar sequences, using the scoring scheme for GO annotations, based on degree of similarity of the original query and frequency of occurrence of GO in different sequences. Shin at el [45] proposed graph sharpening as a way to eliminate undesirable edges from sequence and 3D similarity graphs, and showed that graph sharpening together with data integration produced improvement in protein function prediction. Tsuda at el [44] proposed automated method to choose/weigh best networks (out of PPI, genetic interactions, protein complex, Pfam domain structure, gene expression networks) for each protein class, using Support Vector Machines.

More recently, the approach of incorporating Gene Ontology structure into probabilistic graphical models [35] has shown promising results for predicting protein functions while outperforming approaches that do not take advantage of dependencies among different functional terms. The approach described in [35] considers multiple functional categories in the Gene Ontology (GO) simultaneously. In their model each protein is represented by its own annotation space - the GO structure. In this case, the information is passed within the ontology structure

as well as between neighboring proteins, leading to an added ability of the model to explain potentially uncertain single term predictions.

Multiple approaches have proven that incorporating heterogeneous data to predict protein function can improve the overall predictive power of automated protein/gene annotation systems, as for example shown in [1, 35, 37]. Integrating multiple sources of information is particularly important as each type of data captures only one aspect of cellular activity—PPI data suggest a physical interaction between proteins, sequence similarity captures relationships on a level of orthologs (inter-species relationship) or paralogs (intra-species relationship), and gene ontology defines term-specific dependencies.

Many learning approaches rely on information available from neighbors in a protein network [1,32,37]. However, there may exist proteins with *no* edges connecting them to other proteins in their own networks. For example, considering Yeast and Fly networks, Yeast protein YPL262W has no edges of high sequence similarity to other proteins in its own Yeast network, but it is connected to two Fly proteins (CG6140-PA, CG4095-PA) through high similarity edges. On the other hand, Fly protein CG4866-PA and Yeast protein YHR148W do not share any sequence similarity with proteins in their own networks, but are connected through a highly homologous edge with each other. In a single species network it is often the case that proteins are surrounded only by proteins whose functional information is absent or very limited. In such cases, using information from multiple species becomes crucial: neighborhoods of many proteins are expanded by connecting them to proteins of high sequence similarity in a different species' network. Through such multi-species networks sufficient information may be accumulated to improve the accuracy of protein functional prediction.

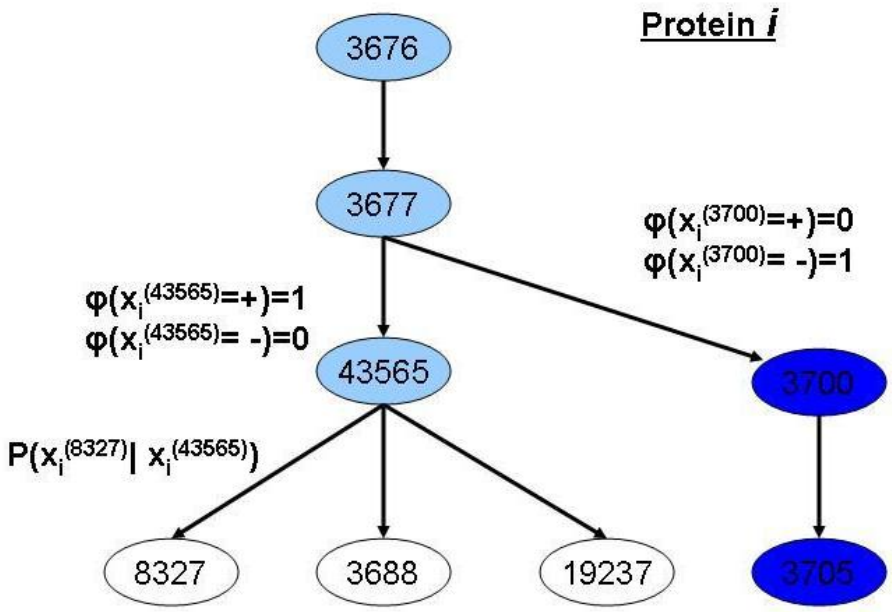


Figure 3.1: The hypothetical protein is positively annotated (light blue color) to GO term 43565 and, thus, also positively annotated to its parent - GO term 3677 , and further up the tree to the parent's parent, term 3676. The term 3700, with the darker blue shade, indicates the negative annotation of the protein to this term. Its child, term 3705, inherits this negative annotation. The protein is unknown at the three unshaded (white) terms.

3.2 Methods

3.2.1 Single Species Network

In our work, we employ the idea of probabilistic chain graphs with incorporated Gene Ontology dependencies [35] to build protein network for each species (such as Yeast and Fly).

In this method, each protein is represented not by a single node, but by a replicate of a Gene Ontology (or subontology), as depicted in Figure 3.1. Gene Ontology (GO) is a directed acyclic graph which describes a parent-children relationship among functional terms. The child term either *"IS_A"* special case of the parent or is a *"PART_OF"* the parent's process or its component. Every protein has its own annotation space corresponding to each of the functional terms in the Gene Ontology. The annotations can, in turn be, assigned positive, negative or unknown states.

Because the relationships between children and parents are directional, if a protein is positively annotated to a child, it is also, by definition, positively annotated to a parent. However, the reverse relationship does not hold. At the same time, if a protein is negatively annotated to a parent term, it will be negatively annotated to all the children terms.

From the above definition it becomes clear that the probability that the child term is negative, given that the parent term is negative, is one. In the presence of multiple parents, a negative state of any parent immediately yields a negative state for child. This step leaves the only probabilities that remain to be estimated as those that define the likelihood of a child being positively/negatively annotated when its parent is (or all parents are) positive.

$$P\left(\left\{x_i^{(c)}\right\}_{c\in GO, i\in\mathcal{I}}\right) = \frac{1}{Z}\prod_{c\in GO}\prod_{i\in\mathcal{G}^{MRF}}\phi(x_i^{(c)}) \prod_{(i,j)\in\mathcal{G}^{MRF}}\psi_{within}(x_i^{(c)}, x_j^{(c)}|\theta_{i,j}^{MRF})\prod_{i\in\mathcal{I}}P(x_i^{(c)}|Pa(x_i^{(c)}), \theta_c^{GO}), \tag{3.1}$$

By defining such probabilistic dependencies for the Gene Ontology terms (conditional probability distribution of all child terms given their parent terms), we create a Bayesian network (BN) representation for each protein, as represented in Figure 3.1.

We encode the ability of our model to transfer functions among similar proteins using a probabilistic graphical representation of a Markov Random Field (MRF) [49], similarly to [32, 42, 43]. In our work we consider two measures of similarity within each species network: sequence similarity determined through normalized BLAST scores and protein-protein interactions. The notion of similarity between proteins in this case is not directional, unlike the case of Gene Ontology.

For each measure of similarity we define a potential function, which corresponds to the probability of joint annotation of two proteins at a term, given that the proteins are similar. The sequence similarity-based potential for proteins i and j at term c is defined as

$$\begin{aligned}\psi(+,+) &= \psi(-,-) = s_{i,j,c}^{\text{within}} \\ \psi(+,-) &= \psi(-,+) = 1 - s_{i,j,c}^{\text{within}}\end{aligned}$$

where $s_{i,j,c}^{\text{within}}$ is a pairwise normalized BLAST score (we only consider normalized BLAST scores above 0.5). In this case, $s_{i,j,c}^{\text{within}} = s_{i,j}^{\text{within}}$ for all terms c.

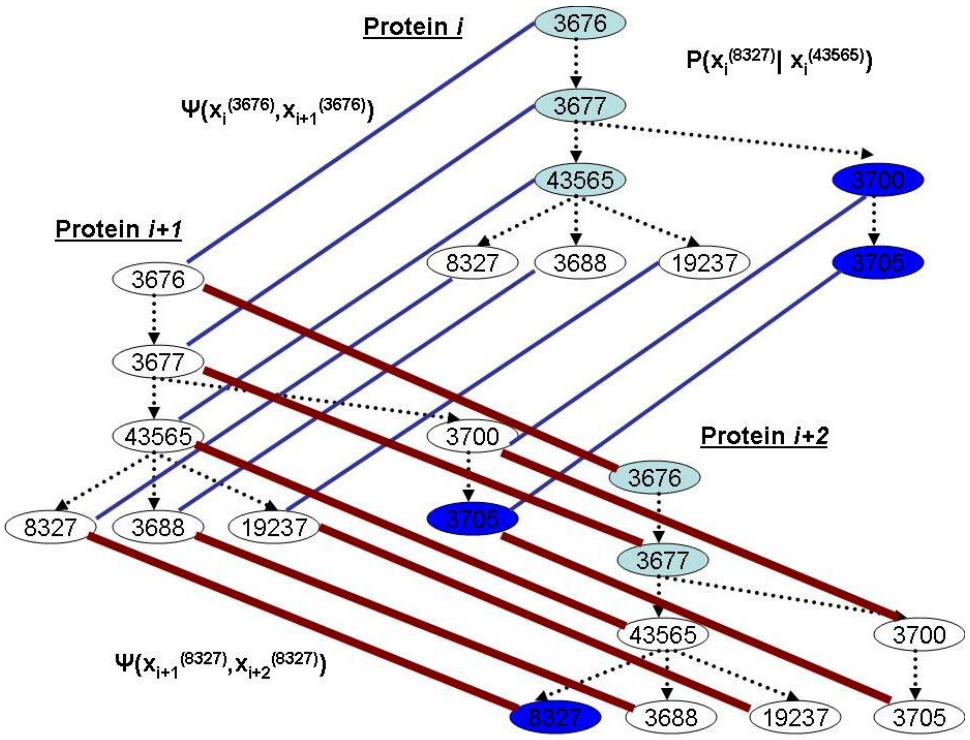


Figure 3.2: A chain graph model with three proteins. Each protein is represented by GO subontology of size eight, with different annotations present at each protein. Some model elements, P and potential function ψ, are shown.

Similarly, the PPI-based potential is defined in a term-specific way as shown below

$$\begin{aligned}
\psi(+,+) &= P(+,+|interaction), \\
\psi(-,-) &= P(-,-|interaction), \\
\psi(+,-) &= P(+,-|interaction) \\
\psi(-,+) &= P(-,+|interaction),
\end{aligned}$$

where the quantities are estimated using relative frequency counts from the training data.

$$P\left(\left\{x_i^{(c)}\right\}_{c\in GO, i\in\mathcal{I}_{\text{sp 1}}\cup\mathcal{I}_{\text{sp 2}}}\right) = \frac{1}{Z}\prod_{c\in GO}\prod_{i\in\mathcal{G}^{MRF(\text{sp 1})}}\phi(x_i^{(c)})\prod_{i\in\mathcal{G}^{MRF(\text{sp 2})}}\phi(x_i^{(c)}) \tag{3.2}$$

$$\prod_{(i,j)\in\mathcal{G}^{MRF(\text{sp 1})}}\psi_{within}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{sp 1})})\prod_{(i,j)\in\mathcal{G}^{MRF(\text{sp 2})}}\psi_{within}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{sp 2})})$$

$$\prod_{(i,j)\in\mathcal{G}^{MRF(\text{sp 1}\cap\text{sp 2})}}\psi_{between}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{sp 1}\cap\text{sp 2})})\prod_{i\in\mathcal{I}}P(x_i^{(c)}|Pa(x_i^{(c)}),\theta_c^{GO})$$

If both the similarity measure and the PPI occurred between a pair of proteins, the total potential ψ is defined as a product of the similarity-based potential and the PPI-based potential [35].

In the model, each protein i can have the evidential function ϕ at each term c, defined as follows. Let $x_i^{(c)}$ be the positive or negative annotation of a protein i to a particular term c. Then the evidential function models our knowledge of particular term annotations: a positively annotated protein at term c is modeled with $\phi(x_i^{(c)})$ defined as $\phi(+)=1, \phi(-)=0$. Similarly, when a protein is negatively annotated at c, the zero and one values are interchanged so that $\phi(+)=0, \phi(-)=1$. For proteins with no annotation the evidence ϕ is set to 0.5.

Our final model is embodied in a chain graph [50], a hybrid between a Bayesian Network (BN) and a MRF, see Figure 3.2. Operating all of the above parameters, the single-species model (of either Fly or Yeast, in our case) can now define a joint Gibbs distribution of functional term annotations over a set of proteins, as defined in Equation (1),where Z is the normalizing constant and $Pa(x_i^{(c)})$ is a parent (parents) of the GO term c in the protein x_i.

Once the network (chain graph) is built, the information is passed from annotated proteins through undirected links to their neighbors. At the same time the in-

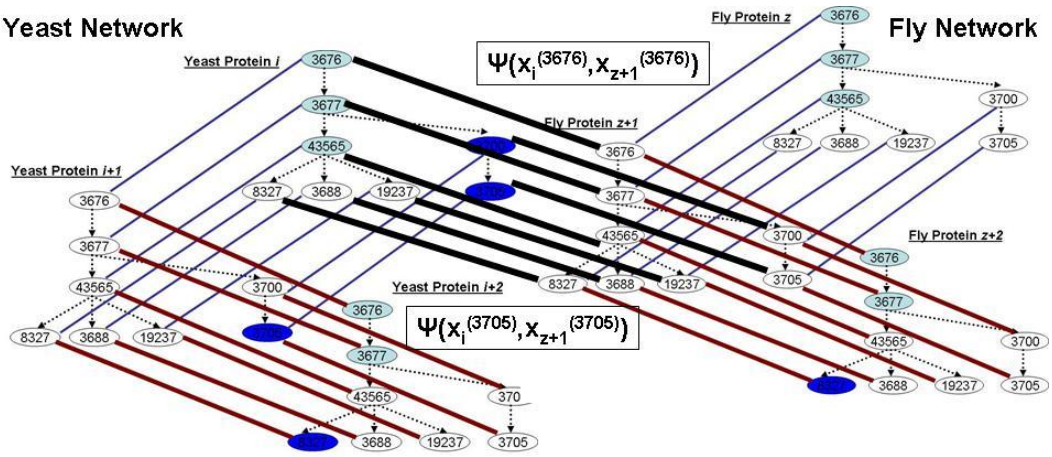


Figure 3.3: Yeast and Fly networks joint by the similarity edges between Yeast's protein *i* and Fly's protein *z+1*. The edges between all GO terms of these proteins are in dark bold, with ψ shown.

formation flows within each protein's Bayesian network along the directed links, according to the conditional probabilistic relationships among different terms. In this fashion the annotation information is accumulated both via the similarity MRF and the ontology BN. For each term of a protein, a set of neighbors is defined by the local connectivity: for example, in the Figure 3.2 the neighbors of 3688 in the protein $i+1$ are $x_{i+1}^{(43565)}, x_i^{(3688)}, x_{i+2}^{(3688)}$.

The flow of information is modeled using a message-passing mechanism for chain graphs, similar to that described in [35]. Messages are passed until the state of convergence is reached; we define it as state at which all normalized messages change by less than 10^{-4} between successive iterations. We employ the "down" message-passing schedule: messages are initiated from the annotated term nodes, sent to all of their neighbors, then to the neighbors of their neighbors, and so on, until all nodes have sent their messages out.

At convergence, the posterior probabilities of membership in the classes de-

fined by GO are calculated at the target proteins, and predictions are made based on those probabilities. We compare the beliefs, obtained thus, to a preselected threshold. Prediction decisions are based on 0.8 decision threshold, as suggested in [32, 35].

3.2.2 Multi-species network

Our next step is to join networks of two (or more) species by edges of high sequence similarity into one computational model. In particular, an edge is introduced between homologous proteins in two species if their normalized BLAST score is above 0.5 (the similarity is high). On the other hand, inter-species edges are not introduced when the score is below 0.5 (the similarity is low), since dissimilar proteins may or may not be involved in the same biological process. Moreover, most of the protein pairs would share some low similarity, which would obscure the network with potentially irrelevant low-similarity edges.

More formally, in a two-species setting, we define a similarity measure between protein i in Yeast network and protein j in Fly network, at term c, as $s_{i,j,c}^{\text{between}}$, a normalized pairwise BLAST score. Consequently, the potential function for homologs between different species is defined as

$$\begin{aligned} \psi(+,+) &= \psi(-,-) = s_{i,j,c}^{\text{between}} \\ \psi(+,-) &= \psi(-,+) = 1 - s_{i,j,c}^{\text{between}} \end{aligned}$$

Similarly to a single-species model, we consider $s_{i,j,c}^{\text{between}} = s_{i,j}^{\text{between}}$ for all terms c of the Gene Ontology, as illustrated in Figure 3.2. While this assumption may be open to debate, it is shown to lead to improved annotation performance.

Considering heterogeneous values of similarity $s_{i,j}^{\text{between}}$ at each term c may lead to additional improvements, at a cost of a more complex and demanding parameter estimation process.

The combined model for joint Fly-Yeast (referred to as "sp 1" and "sp 2") network now defines a joint Gibbs distribution of functional term annotations over a set of all proteins in the chain graph, detailed in Equation (2). Here, Z is the normalizing constant, ψ_{within} is a similarity measure within one species network, $\psi_{between}$ is a similarity measure between the networks, and finally, $Pa(x_i^{(c)})$ is a parent (parents) of the GO term c in the protein x_i.

After the joint network is built, a belief propagation algorithm is used to make predictions at all ontology terms in both species. We consider a state of convergence and decision thresholds to be defined similarly to a single-species network.

Adding inter-species homology information into the learning model has unique advantages and shows significant improvements in protein function prediction. The model is specifically beneficial for proteins isolated in their own networks (having no interacting neighbors) or for proteins which are surrounded by poorly annotated neighbors. In a multi-species setting, the neighborhood of such proteins is expanded so that they can learn their functional annotations from their homologs in the different species.

3.3 Experiments and Results

3.3.1 Experiment design

We apply our method to two largest protein networks of Yeast and Fly as well as to a joint Yeast-Fly network. Predictive performance of our models is evaluated in a 5-cross validation setting. The test set consists of a random 20% of annotated proteins, that maintains the same proportion of negatively and positively annotated proteins as the remaining 80% of the data used for training the model. For each randomly chosen test protein, *all* of its annotations are left out—the Gene Ontology structure remains in place but the functions at all terms are now listed as unknown. In the case of a joint Fly-Yeast network, we eliminate annotations of 20% of annotated proteins from *each* network. In the testing phase, upon convergence of the message-passing process, predictions at terms whose annotations were left out are tested against the known eliminated annotations.

For each tested network, we conduct a total of ten experimental rounds using the random splitting process. In each round, we compared results of runs on single networks (without joining) to that of the joint network. Individual and joint networks are trained and evaluated on the same training/testing data.

For the measure of intra- and inter-species similarity we used normalized BLAST scores, defined as a BLAST score divided by self score of query (i.e. BLAST score of the homologue divided by the BLAST score of the protein against itself), ranging from 0 to 1. We obtained sequence and annotation data from Saccharomyces genome Database [51] for Yeast (February 2 and April 11, 2009 release) and FlyBase [52] for Fly (April 27, 2009 release). Protein-protein interaction data were obtained from BIOGRID [53] database (April 27, 2009 re-

lease). We considered only manual (higher quality) annotations, since computational predictions have been noted to present a conflicting evidence. To expand the applicability of our method, we considered *all* reported in the above sources Yeast and Fly proteins (as opposed to considering only proteins with specific evidence, such as protein-protein interactions). This approach resulted in a combined set of 12199 Fly and 6008 Yeast proteins that were used to construct our joint belief networks.

Gene ontology structure was downloaded from the Gene Ontology database [54]. When reading Gene Ontology annotations, we consider two fundamental GO assumptions: GO hierarchy is expanded up for positively annotated proteins (if a protein is positively annotated by a term, then it is also positively annotated by all of its parents/ancestors) and is expanded down for negatively annotated proteins (if a protein is negatively annotated by a term, then it is negatively annotated by all of its children/descendants). We construct a negative set relying on co-annotation (co-occurence) statistics of GO annotations in the data (further maintaining two fundamental GO assumptions). In particular, a protein is considered negatively annotated by a specific GO term if this term has never been observed to co-occur with a known function for a given protein, given the training data.

Our example of gene ontology was taken from molecular function subtree of GO hierarchy [1], as depicted in Figure 3.1. As previously investigated in [32, 33, 35, 37, 38] among others, PPI networks have strong predictive power for molecular function categories of Gene Ontology, especially in combination with other

[1]The original GO subontology covered eight terms: nucleic acid binding (3676), DNA binding (3677), sequence-specific DNA binding (43565), methyl-CpG binding (8327), DNA replication origin binding (3688), centromeric DNA binding (19237), transcription factor activity (3700), and RNA polymerase II transcription factor activity, enhancer binding (3705).

sources of evidence (such as intra- and inter- species homology). Previously PPI and intra-species sequence homology *together* showed significant improvements in predicting molecular functions of proteins, as for example shown in [35,37,38]. Most importantly, the use of the proposed *inter-species homology* may render our computational method, a core concept of this work, broadly applicable to all three ontologies: molecular function, biological process and even the cellular component.

Our method can be applied to the entire gene ontology, at the expense of time and space complexity. However, in practice, biologists and clinicians are interested in *specific*, relatively small, subontologies, targeted in our study. For instance, vaccine and drug targets are usually the proteins that perform very specific functions, represented by the leaves of a specific Gene subontology.

Using our algorithm, 583 Fly and 236 Yeast proteins are annotated to one or more terms of the selected subontology (among those 110 Fly and 29 Yeast proteins were assigned some negative annotations). Other proteins are unannotated to a given subontology and are used as intermediate points for information passage.

3.3.2 Results

For our model, we operate several performance measures, such as: precision, recall, accuracy, false positive rates, and F1 defined as: $recall = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$, $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $fpr = \frac{FP}{TN+FP}$, $F1 = \frac{2*precision*recall}{precision+recall}$, respectively.

The calculations are done separately for the Yeast network, the Fly network

size	network	precision	recall	accuracy	FP rate	F1 measure
8	Fly	100	99.78	99.86	0	99.89
	Fly, JN	100	99.78	99.86	0	99.89
12	Fly	99.00	99.40	99.25	0.90	99.23
	Fly, JN	99.36	99.33	99.37	0.60	99.34
16	Fly	98.44	98.93	98.75	1.25	98.68
	Fly, JN	99.20	98.25	98.95	0.625	98.72

Table 3.1: Average precision,recall,accuracy, false positive rate, and F1 over 10 runs for **Fly** species in isolated Fly and joint Fly-Yeast networks (percentage wise) for subontologies of various sizes. JN stands for joint Yeast-Fly network.

and the joint Fly-Yeast network. In the joint network, we separately calculate the performance of Fly and Yeast species and compare them to those in isolated networks.

In this work, we consider GO subontologies of different sizes. The main focus is on the GO subontology of size 8, similarly to our previous work in [55]. We expand our model to subontologies of bigger sizes: 150% the size of the original subontology (size 12) and 200% the size of the original ontology (size 16), shown in Figure 3.4. A typical run of the model with the 8-sized ontology on the joint Fly-Yeast network (on 3.6 GHz CPU with 8GB memory machine) takes approximately 28 minutes (with four iterations of message passing). In comparison, corresponding runs on individual species networks take 59 minutes for Fly and 35 minutes for Yeast.

While the difference in running times may at first appear to go against intuition, faster convergence rates in a Joint Network can be attributed to the presence of "denser" sources of evidence in networks of multiple species compared to that of the isolated runs.

Table 3.1 shows the average precision, recall, accuracy and false positive rate

size	network	precision	recall	accuracy	FP rate	F1 measure
8	Yeast	89.52	97.66	91.13	29.32	93.41
	Yeast, JN	100	96.17	97.27	0	98.05
12	Yeast	94.98	97.05	95.27	7.24	95.96
	Yeast, JN	98.33	96.94	97.33	2.18	97.63
16	Yeast	95.06	96.31	95.54	4.9	95.64
	Yeast, JN	99.01	95.6	97.7	0.465	97.26

Table 3.2: Average precision,recall,accuracy, false positive rate, and F1 over 10 runs for **Yeast** species in isolated Yeast and joint Fly-Yeast networks (percentage wise) for subontologies of various sizes. JN stands for joint Yeast-Fly network.

for Fly: in isolated Fly network, and in joint Fly-Yeast network, for subontologies of various sizes. Table 3.2 shows corresponding measures for Yeast.

The overall performance of Fly and Yeast networks is highly improved (compared to the results presented in our previous work [55]), which is most likely due to the more reliable sequence similarity scores, expanded protein coverage, and more general definition of a negative set.

The joint Fly-Yeast network significantly improves precision, accuracy, and FP rate while only slightly suffering from lowered recall, as shown in Table 3.1, for Fly, and Table 3.2, for Yeast. We stress the importance of F1 measure, a harmonic mean of precision and recall, and notice its consistent significant improvement in the joint network, even for larger subontologies. This improved result indicates that despite the larger size and more complex structure, considering networks of multiple species jointly continues to offer important benefits to the prediction process.

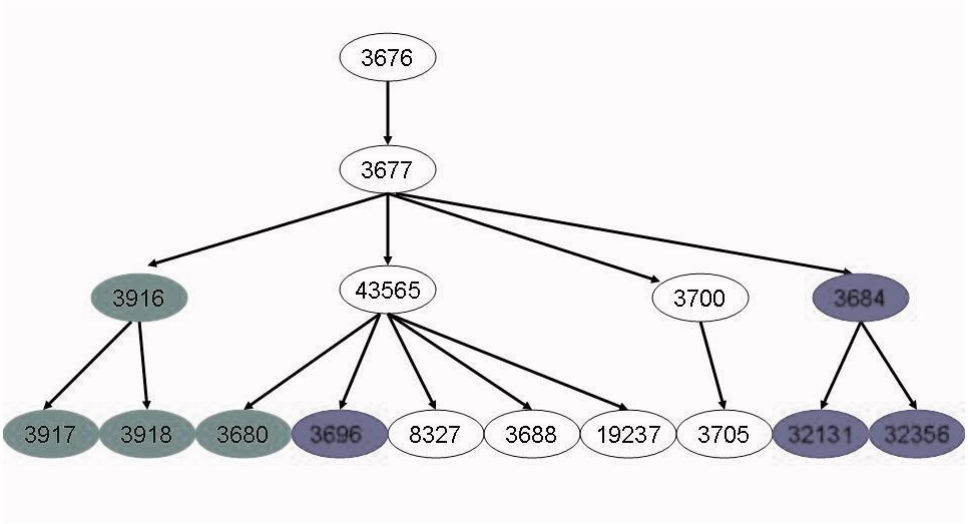


Figure 3.4: Expanded subontologies of size 12 (added nodes are shown in gray) and 16 (added nodes are shown in black).

3.3.3 Statistical analysis

Statistical analysis of significance of the aforementioned performance scores was done using the t-test and the Wilcoxon Signed-Ranks Test [56]. The tests were conducted separately for each species and each performance measure: single Fly network is compared with the performance on the Fly in the joint Fly-Yeast network; and single Yeast network is compared with the performance of the Yeast in the joint Fly-Yeast network. For comparison to be sound, the evaluations on single and joint networks were done using the same random samples (splits for testing and training sets).

t-statistics per species

We present p-values calculated from t-statistics (degree of freedom$= 9$) to evaluate statistical significance of our findings in Table 3.3. We consider p-value to be statistically significant if it is less than 0.05. In general, Yeast shows more sub-

stantial improvements compared to Fly, which could indicate the higher quality of Fly data and better neighborhoods for the majority of Fly proteins.

Wilcoxon signed-ranks test

To remove the possible effects of outliers on the computed t-test statistics random samples can compensate for overall bad performance) we applied the Wilcoxon Signed-Ranks Test. Wilcoxon Signed-Ranks Test is a non-parametric alternative to the t-test, which assumes commensurability of differences in a qualitative way: greater differences count more. In many cases, this test is safer than the t-test since it does not assume a normal distribution.

Let $d_q = E_{c_q^1} - E_{c_q^2}$ be the difference between the performance scores of the approaches on the q-th out of the 10 random trials. Each difference is considered at its absolute value and the values are ranked. In the case of ties between differences, the average score among them is assigned. We use R^+ to denote the sum of ranks for the samples on which the Joint method outperforms the individual network approach; R^- is the sum of ranks when the individual methods "win":

$$\begin{aligned} R^+ &= \sum_{d_q>0} rank(d_q) + \frac{1}{2}\sum_{d_q=0} rank(d_q) \\ R^- &= \sum_{d_q<0} rank(d_q) + \frac{1}{2}\sum_{d_q=0} rank(d_q) \end{aligned}$$

The z-statistic can be calculated as

$$z = \left(T - \frac{1}{4}N(N+1)\right) / \sqrt{\frac{1}{24}N(N+1)(2N+1)},$$

where $T = min(R^+, R^-)$. and$N = 10$ is the number of samples. With $\alpha = 0.05$,

size	network	precision	recall	accuracy	FP rate	F1
8	Fly, t-test	*	*	*	*	*
	Fly, WSR	*	*	*	*	*
	Yeast, t-test	$3 * 10^{-6}$	-	$2.1 * 10^{-5}$	$< 10^{-6}$	$4 * 10^{-6}$
	Yeast, WSR	0.0027	-	0.0046	0.0027	0.0029
12	Fly, t-test	0.22	-	0.35	0.20	0.35
	Fly, WSR	0.024	0.078	0.024	0.024	0.012
	Yeast, t-test	0.016	-	0.019	0.018	0.018
	Yeast, WSR	0.014	0.42	0.014	0.061	0.0053
16	Fly, t-test	0.11	-	0.33	0.10	0.47
	Fly, WSR	0.016	0.003	0.003	0.016	0.11
	Yeast, t-test	0.021	-	0.026	0.011	0.08
	Yeast, WSR	0.016	0.016	0.016	0.016	0.017

Table 3.3: p-statistics from t-test and Wilcoxon Signed-Ranks Test: p-values with respect to precision, recall, accuracy, false positive rate, and F1 as a measure of statistically significant improvements of a joint network performance, for subontologies of various sizes. "*" stands for "cannot be improved".

the null hypothesis will be rejected if $z < -1.96$. We calculate the corresponding p-values from the determined z-values.

The Wilcoxon test similarly confirms significant improvements in performances on the Joint network when compared to individual Yeast and Fly networks, as shown in Table 3.3. In fact, Wilcoxon test "catches" statistically significant improvements where t-test presents no evidence, such as for subontologies of size 12 and 16.

3.4 Comparative analysis

3.4.1 Gene Ontology vs single-term predictions

As a baseline test, we compare our methodology (with GO dependencies) to runs without GO in place, where the whole network of proteins is tested on a single

networks	GO	precision	recall	accuracy	FP rate
Fly	w/o GO	45.57	48.7	74.25	49.05
	GO	100	99.78	99.86	0
Fly | JN	w/o GO	49.5	53.78	54.94	32.13
	GO	100	99.78	99.86	0
Yeast	w/o GO	-	0	43.79	0
	GO	89.52	97.66	91.13	29.32
Yeast | JN	w/o GO	34.76	70.52	72.10	54.1
	GO	100	96.17	97.27	0
JN overall	w/o GO	44.36	59.63	60.9	39.81
	GO	100	98.70	98.98	0

Table 3.4: Comparison of results for the network with GO and without GO

ontology term (single protein function). As before, we perform 5-fold cross validation by choosing random 20% of annotated proteins as a testing set over 10 trials of the program. The results shown in Table 3.4 indicate the superiority of the network with built-in Gene Ontology over the single-term network even in the case of multiple species networks.

It is worth highliting that the model with gene ontology in place makes a true positive prediction where the model without it commits a false negative error. This result is not surprising as there is only one term with one protein annotated to it. In general, similar to [35], incorporating the ontology structure, along with the dependencies among its functional terms, considerably improves performance over that of traditional models that consider each term in isolation.

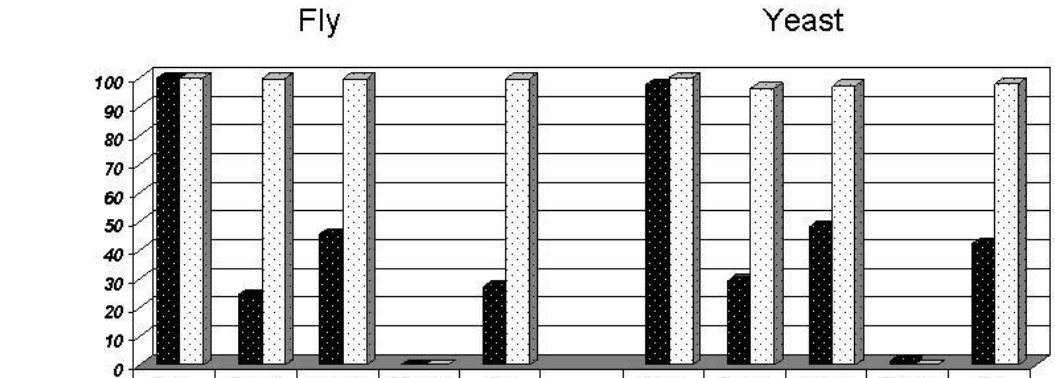
3.4.2 Comparison with other methods

In this section we comprehensively compare our method to the most widely used group of techniques, such as in Nariai et. al. [1], which are based on Bayesian probabilistic approach. In such methodologies, proteins are embed-

ded into protein-protein network so that each protein is represented by a node and similarity measures between proteins (such as protein-protein interactions, sequence similarity, etc.) are represented by edges. In the model, each protein learns its functional annotation based on the number and character of his neighbors in the protein network, particularly *total* number of neighbors and number of *annotated* (to the GO term of interest) neighbors. This information is then embedded into a probabilistic Bayesian framework, which consequently assigns a probability to a protein of interest as positively or negatively annotated to a specific GO term [1]. Since fundamentals of Bayesian probabilistic approach are at the heart of the overwhelming majority of methods currently used for protein function prediction, we compare ourselves against this computational technique.

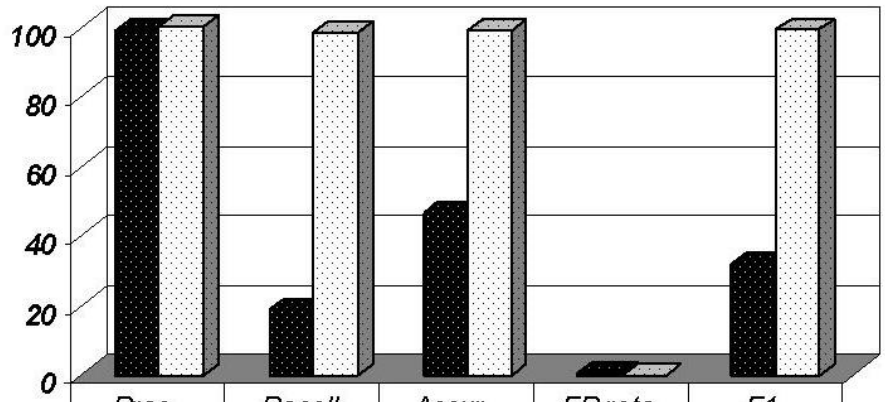
To achieve the most accurate comparative results, we use the same 10 training/testing sets as in our own experimental studies in a 5-fold cross validation setting. Similarly to our setting, both PPI and Sequence similarity (determined by normalized BLAST cores) are used to build protein interaction networks.

We present results as performance of Yeast and Fly species in the joint network (Figure 3.5), as well as overall performance measures (Figure 3.6) in the joint network. We show that our method outperforms the Bayesian probabilistic approach of Nariai et al [1] in all statistical measures, such as F1 rate, precision, recall, and accuracy, for all validation sets considered. Interestingly, Fly species achieves precision of 1 and FP rate of 0 even in the method of Nariai et. al. (the same is observed in our method for the subontology of size 8), which might be indicative of a higher quality of data used to build Fly protein network and a presence of a good learning neighborhood for the majority of Fly proteins.



	Fly					Yeast				
	Prec.	Recall	Accur.	FP rate	F1	Prec.	Recall	Accur.	FP rate	F1
Bayesian	100	24.36	45.67	0	27	97.69	29.44	48.02	1.41	42.2
Our method	100	99.78	99.86	0	99.89	100	96.17	97.27	0	98.05

Figure 3.5: Comparison of our method to the Bayesian probabilistic approach of Nariai et. al. [1]: performance of Fly and Yeast species in a joint Fly-Yeast network.



	Prec.	Recall	Accur.	FP rate	F1
Bayesian	98.89	18.75	46.5	0.299	31.88
Our approach	100	98.7	98.9	0	99.3

Figure 3.6: Comparison of our method to the Bayesian probabilistic approach of Nariai et. al. [1]: overall performance of a joint network.

3.5 A Model Checking Interpretation

Our expanded Gene Ontology approach can also be interpreted as a special case of a new broader framework of "probabilistic graphical model checking." The framework resembles classical model checking algorithms [57] implemented through message passing in a statistical graphical model. This connection becomes explicit when a Gene subontology for a protein (Figure 3.1) is viewed as a family of properties encoded through logical propositions and connectives. Also modal operators and quantifiers may be added, if further generalizations are desired. These properties can be embedded and propagated in a general graphical structure with certain logical implications—all interpreted in a three-valued logic: True (positive), False (negative) and Unknown. For example, in the currently used Gene subontology, the positive information about a child implied positive information about a parent; and negative information about a parent implied negative information about child. Additionally, we define a probability for a child being positive/negative given that a parent is positive, which defines a probabilistic framework for the model. Thus, if we view our graphical model as not strictly related to a GO subontology, but to a more general framework such as this, we can define any set of properties on the elements of this graphical structure, introduce time frames, or imply hierarchical relationships for this graph. Once we define relationships/properties, we can then propagate these properties in the entire model (which in our application, corresponds to message passing).

For specific species, our framework connects subontologies of all proteins by edges. In the language of model checking on graphical models, subontology network for each species can be viewed as an initial labeling of "possible worlds"

with certain relationships/properties. By connecting networks of two different species we thus connect two neighboring "possible worlds" and try to gain some additional information from their distances (measured by orthology or PPI). Theoretically, if the two possible worlds are adjacent, they are expected to satisfy similar properties. Considering both "worlds" simultaneously will lead to algorithms with high fidelity and improved efficiency. Our approach suggests, for propositional and temporal logic, a potentially much broader range of applications including many non-biological problems.

3.6 Conclusions

In this chapter we presented a novel approach that uses inter-species sequence homology to connect networks of two, and possibly more, species together with Gene Ontology dependencies in order to improve the predictive ability needed for protein classification. Joining the networks of two different species shows important advantages over runs on individual networks. While in single species networks proteins may exist that have no annotated partners, they have the potential to acquire annotated interacting partners-homologs in a two-species setting. Additional benefits emerge for species with poorly defined protein functions and/or protein interactions. Additionally, the use of the Gene Ontology enables simultaneous consideration of multiple but related functional categories, opening information paths for further improvements to the model's predictive ability.

Our method readily extends to multiple species settings, and may produce improvements similar to the case of two species. The presence of multiple interacting networks may further enable integration of additional sources of evidence,

thus contributing to increased accuracy in functional predictions.

3.7 Web Resources and Supplementary material

Supplementary and sample input data are available from

`The code (C/C++/Perl) and input files are available from`

`http://www.cims.nyu.edu/~antonina/yeast_fly.html`

Chapter 4

Protein Classification using Malaria Parasite's Temporal Transcriptomic Profiles

4.1 Introduction

Search for vaccine for malaria infections has been under intense study for many years, but it has resisted several different lines of attack attempted by biologists. More than half of Plasmodium proteins still remain uncharacterized and therefore cannot be used in clinical trials. The task is further complicated by the metamorphic life cycle of the parasite, which allows for rapid evolutionary changes and diversity among related strains, thus making precise targeting of the appropriate proteins for vaccination a technical challenge. We propose an automated method for predicting functions for the malaria parasite, which capitalizes on the importance of the intraerythrocytic developmental cycle data and expression changes

during its five phases, as determined computationally by our segmentation algorithm.

Our method combines temporal gene expression profiles with protein-protein interaction data, sequence similarity scores, and metabolic pathway information to produce a set of predicted protein functions that can be used as targets for vaccine development. We use a Bayesian approach, which assigns a probability of having (or not having) a particular function to each protein, given the various sources of evidence. In our method, each data source is represented by either a functional linkage graph or a categorical feature vector.

The methods are tested on *Plasmodium falciparum*, the species responsible for the deadliest malaria infections. The algorithm was able to assign meaningful functions to 628 out of 1439 previously unannotated proteins, which are first-choice candidates for experimental vaccine research. We conclude that analyzing time-course gene expression profiles in separate phases leads to much higher prediction accuracy when compared with Pearson correlation coefficients computed across the time course as a whole. Additionally, we demonstrate that temporal expression profiles alone are able to improve the predictive power of the integrated data.

4.1.1 Background

World-wide, each year, malaria infects approximately 515 million people and kills between one and three million of them. A better understanding of *protein functions* in malaria parasites can have a tremendous effect on approaches aimed at preventing malaria epidemics. This anticipated impact is suggested by the fact

that targets for drug and vaccine design are almost always based on proteins, particularly those involving enzymatic functions. Unfortunately, since many *Plasmodium falciparum* proteins remain uncharacterized, they are mostly ignored by pharmaceutical laboratories and disregarded as potential protein targets in drug and vaccine development. In order to reverse this trend, it is necessary to devise more effective automated bioinformatic tools for protein classification.

Toward this goal, this chapter addresses the issue of predicting protein functions using many sources of data, with an emphasis on the use of time series gene expression data. Unlike most methods, we allow for changes in regulatory patterns, and relationships, over time. The methods are tested on a species of malaria parasite, *P. falciparum*, that accounts for about 15% of infections and 90% of deaths.

In the past, functional annotation of proteins has been addressed by various computational, statistical, and experimental methods. In many cases, it is convenient to provide a graphical representation of protein networks such that each node represents a protein and edges between nodes represent different aspects of their functional association. The choice of functional association is used to determine the predictive power of such a network. One promising computational approach utilizes the family of probabilistic graphical models, such as belief networks, to infer functions over sets of partially annotated proteins [32, 42, 43, 58]. For instance, Bayesian network methods for data integration have been extensively studied [1, 59–61] for predicting protein-protein interactions and protein function similarity for pairs of genes. Additionally, the approach of incorporating the hierarchical structure of the Gene Ontology (GO) into probabilistic graphical models [35,55] has also yielded promising results for predicting protein functions

for gene subontologies of interest.

The most established methods for protein function prediction are based on sequence similarity using BLAST [62] analysis, and rely on the fact that similar proteins are likely to share common functions. Such similarity-based methods include sequence homology [35, 39–41, 55, 63], and similarity in short signaling motifs, amino acid composition and expression data [64–69]. At the same time, protein-protein interaction (PPI) data are widely used to infer protein functions. For example, methods described in several recent papers [32,42,43,58] used PPIs to define a Markov random field over the entire set of proteins. In general these methods suggest that interacting neighbors in PPI networks might also share a function [32–34, 70]. Clustering of genome-wide expression patterns has also been used to predict protein function, as described in [1, 71–73].

4.1.2 Protein function prediction in parasites

Saccharomyces cerevisiae (Baker's Yeast) is chosen for many case studies involving protein functions, since it has been extensively studied from multi-omic view-points, and its protein data are also the most complete. The problem of protein function prediction is, however, more difficult in parasites, where genetic and biochemical investigations are much more challenging. For example, it is problematic to isolate a malaria parasite at various stages of its development (e.g., the life-cycle of *P. falciparum* is very rapid, ookinetes are difficult to isolate in large numbers, the liver stage of a parasite's development is hard to study because of technical difficulties). Such obstacles manifest themselves in a paucity of information on the protein properties, interactions, localization and motifs of

Plasmodium species.

When relying on just one source of protein information, it is difficult to devise a reliable probabilistic framework with the ability to automatically predict classifications for proteins of interest. Indeed, combining various types of information was demonstrated to improve the overall predictive power of automated protein/gene annotation systems for *S. cerevisiae*, as shown in [1, 35, 55]. Integrating multiple sources of information is particularly important as each type of data captures only one aspect of cellular activity. For example, PPI data suggest a physical interaction between proteins; sequence similarity captures evolutionary relationships at the level of orthologs; gene expression suggests participation in related biological processes that take place at a certain cell cycle stage; and finally, GO defines term-specific dependencies.

As a result, it motivates one to explore, as in the case of *P. falciparum*, how to combine different sources of information most effectively to infer protein functions. We explore and evaluate a Bayesian probabilistic approach for predicting protein functions in *P. falciparum* by integrating multiple sources of information, namely, protein-protein interactions, sequence similarity, temporal gene expression profiling, metabolic pathway, and GO classifications.

The primary goal of our study is to demonstrate that considering the intraerythrocytic developmental cycle (IDC) phases individually is crucial for protein function prediction in *P. falciparum*. While other data sources (such as sequence homology and protein-protein interactions) describe the static state of *P. falciparum*, time series gene expression data during the IDC reflects the dynamics of the parasite's system, describing rapidly evolving regulatory patterns and expression profiles. In particular, during *P. falciparum*'s IDC, there are distinct periods

of consistent gene regulation, punctuated by instances of reorganization in the regulation pattern. In such a setting, it becomes important to consider each time window (delineating a particular stage) separately. We show that finding these critical timepoints, clustering time-course gene expression data from each stage of the cycle *separately* and then connecting clusters across windows (so that proteins "travel" from one window to the other) produces better results as compared with Pearson coefficient calculations applied to the time-course data as a whole. We assume that if two proteins share expression patterns (i.e., belong to the same cluster) during a period of time, such as the first window or phase, they are likely to share a function. If these proteins also fall into the same cluster in the second window, we would increase our belief in them being similar. Finally, if they belong to the same clusters in all five windows, we would be highly confident that they share related functions.

Additionally, but not less importantly, we illustrate that inclusion of the IDC time-course data improves the predictive power of the Bayesian probabilistic approach even in the integrated setting (when combined with protein-protein interaction, sequence homology, and metabolic pathways data).

Hampered by data-related limitations, we did not expect to make as many accurate predictions as one could for a well-studied organism such as *S. cerevisiae*. However, we were encouraged by being able to propose vaccine-related functions for several *P. falciparum* proteins as these might play a significant role in the next stages of vaccine and drug development, leading to effective control of the disease.

The next part of this process involves trying to understand the underlying causal structure that is governing *P. falciparum*'s gene regulation. That is, now

that we have a set of possible functional annotations, and time course data covering the IDC, we aim to narrow the set of proteins suitable for vaccine exploration by finding those that can be used to affect others. Note that it is likely not as simple as one protein promoting or inhibiting the production of another - there may be arbitrarily complex relationships involving the regulation of multiple genes in concert. Others have recently developed algorithms for causal inference [74], where the relationships are described in a probabilistic temporal logic, allowing arbitrarily complex causes and effects and explicit description of the time between the cause and the effect. Preliminary results of the *P. falciparum* IDC have appeared elsewhere [74]. One of the limitations of this data is the relatively coarse timescale (as compared to other data sets used for causal inference). Rather than exhaustively examining all proteins included in the data, we focused on a smaller set of relationships to test, using our new annotations and processes known to be useful as drug targets.

4.2 Methods

4.2.1 Data

For our analysis, we focused on 2688 *P. falciparum* proteins from the time-course data [75], among which only 1249 proteins possess known biological process annotations.

Protein-protein interaction data

We obtained Y2H (yeast-two-hybrid) data for *P. falciparum* from [76]. This dataset, however, annotates a limited number of protein-protein interactions, because of the confounding effects of the rapid life-cycle of these parasites. The 1130 interactions cover 1312 proteins.

Sequence homology

We started by gathering sequence information for proteins from [76]. Each sequence was queried against the entire *P. falciparum* sequence database [76] using BLAST. We recorded BLAST pairwise p-scores as p_{ij}'s (where i and j index the proteins) and defined a measure of sequence similarity for each pair as $s_{ij} = 1 - p_{ij}$. For our purpose, we defined proteins i and j to be similar (sequence-wise), if their pairwise p-value $p_{ij} < 10^{-4}$. There are 1799 proteins meeting this criteria.

Metabolic pathway data

We used metabolic pathway data from [77]. For example, protein PFA0145c is a part of 'Asparagine and Aspartate metabolism' and 'Protein biosynthesis' pathways. The data consists of 119 metabolic pathway categories for *P. falciparum*. The 3526 data pairs cover 1998 genes.

Temporal gene expression data

Time-course gene expression data covering the 48 hours of the intraerythrocytic developmental cycle of *P. falciparum* was obtained from a study by Bozdech et

al. [75]. While the IDC comprises three main stages (ring, trophozoite, and schizont, separated by two critical transition instants), the work in [78] identified four critical transition instants with major changes in gene regulation, corresponding to the following five developmental periods: End Merozoite/Early Ring stage, Late Ring stage/ Early Trophozoite stage, Trophozoite, Late Trophozoite/ Schizont, and Late Schizont/Merozoite. Each period defines a window of time ranging from 7 to 16 hours. We consider each window separately and process it with k-means clustering.

Gene Ontology data

We used GO terms as the basis of our annotation. In particular, we used the 763 biological process associated GO terms available for *P. falciparum*. For each term we expanded the GO hierarchy "up" (including is-a and part-of relationships) so that if a protein is positively annotated by a GO term, then it is also positively annotated by all of its parents/ancestors. There are 16113 GO biological process associated pairs, which cover 1249 *P. falciparum* proteins. Also, following Nariai et al. [1], we excluded labels that appear fewer than five times among these genes, since these terms did not constitute a sample large enough to make sufficiently predictive contributions. Following suggestions in Nariai et al. [1], we define a negative protein-term association as follows: if the association is not in the positive set (defined above), and a gene is annotated with at least one biological process, and the negative annotation is neither an ancestor nor a descendant of the known function for this protein then it is treated as a negative association.

4.2.2 Data representation

In order to use the available information to its full potential, it is necessary to design a proper data representation that optimally reflects the properties and structure of the data itself. We represent the data using two types of structures: *functional linkage graphs* and *categorical feature vectors*.

A *functional linkage graph* is a network in which each node corresponds to a protein and each edge corresponds to the measure of functional association. Such a network takes into account the number and the nature of interacting partners for each protein. We use this representation for PPI and sequence similarity, since, for these data, interacting partners are more likely to share a function. We encoded PPI and sequence homology data using separate functional linkage graphs. In the case of PPI, the edges represent known protein-protein interactions. In the case of sequence similarity (homology) an edge is added when the pairwise p-score is less than 10^{-4}.

We adopted some ideas of the representation and analysis of functional linkage graphs from Nariai et al. [1]. For each functional linkage graph l and for each GO label t, we define $p_1^{(l)}$ and $p_0^{(l)}$, where $p_1^{(l)}$ is the probability that a protein has label t, given that the interacting partner has label t and $p_0^{(l)}$ is the probability that a protein has label t given that the interacting partner does not have label t. For the *P. falciparum* network, we performed the χ^2 test to show that $p_1^{(l)}$ and $p_0^{(l)}$ were statistically different using a Bonferroni-corrected p-value of $0.001/T$, where T is the number of terms tested from each data set.

Another method of data representation is the *categorical feature vector*, which holds a list of categories where we assign 1 to a protein that belongs to a given

category and 0 otherwise. We used categorical feature vectors for the metabolic pathway data. We define m_r as a random variable associated with a protein so that $m_r = 1$, if a protein participates in metabolic pathway r, and $m_r = 0$, otherwise. A feature vector $\mathbf{m} = (m_1, m_2, \ldots, m_r)^T$ is defined for each protein, where $r = 119$ is the total number of metabolic pathway categories.

Finally, we use categorical feature vectors to represent the gene expression profiles. Gene expression profiles are usually encoded as functional linkage graphs using the Pearson correlation coefficient calculated for all combinations of genes. However, we found that the Pearson coefficient might not reflect the temporal relationships, which are crucial to the *P. falciparum* IDC. Instead, we consider expression data for each phase of the IDC separately. We used the five time points found by [78] and applied k-means clustering to the expression patterns of each time period, as described below. We considered proteins from the same cluster to share the same categorical feature and thus possibly have related functional annotations. Consequently, if proteins fall into the same clusters for all or most of the time periods, they will have similar categorical feature vectors and are more likely to share protein classification.

More formally, we define a random variable d_r^j associated with each protein such that $d_r^j = 1$ if a protein is in cluster r in the time period j, and $d_r^j = 0$, otherwise. A feature vector is then

$$\mathbf{d} = (d_1^1, d_2^1, \ldots, d_q^1, d_1^2, d_2^2, \ldots, d_q^2, \ldots, d_1^w, d_2^w, \ldots, d_q^w)^T,$$

where $q = k$ is the number of clusters produced by k-means clustering and $w = 5$ is the number time windows.

4.2.3 Posterior probability computation

For each protein i and each function t, we computed the posterior probability of the protein having the specified function. We define a variable $L_{i,t}$ which is equal to 1 if i is labeled with t. Our ultimate goal is to calculate the probability of $L_{i,t} = 1$ for all i and t given all the available data sources and network structures. To calculate this probability, we follow the general principles described in Nariai et al. [1] and summarize these principles below.

The graphical data representation emphasizes the importance of the neighbors for each protein. We define $N_i^{(l)}$ as the number of neighbors of protein i in the functional linkage graph l (unannotated neighbors are excluded). Additionally, for the corresponding t, $k_i^{(l)}$ is defined as the number of neighbors of protein i annotated with term t in the functional linkage graph l. In our case, $l = 1$ corresponds to the PPI and $l = 2$ to the sequence similarity network, .

At the same time, $\mathbf{c}_i^{(j)}$ is the feature vector that protein i has for a functional category j. In our case, $\mathbf{c}_i^{(1)}$ is the temporal data gene expression feature vector **d** and $\mathbf{c}_i^{(2)}$ is a metabolic pathway feature vector **m** of a protein i.

We calculate the posterior probability of $L_{i,t} = 1$ given functional linkage graphs and category feature vectors of proteins as follows:

$$P(L_{i,t} = 1|N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}) \tag{4.1}$$

$$= \frac{P(L, N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})}{P(N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})} \tag{4.2}$$

$$= \frac{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|L, N_i^{(1)}, N_i^{(2)})P(L, N_i^{(1)}, N_i^{(2)})}{P(N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})} \tag{4.3}$$

$$= \frac{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|L, N_i^{(1)}, N_i^{(2)})P(L|N_i^{(1)}, N_i^{(2)})}{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|N_i^{(1)}, N_i^{(2)})} \tag{4.4}$$

Assuming that k's and $\mathbf{c}$'s are independent, and that L is independent of the total number of graph neighbors $N_i^{(l)}$, then the numerator becomes:

$$P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|L, N_i^{(1)}, N_i^{(2)})P(L|N_i^{(1)}, N_i^{(2)}) \tag{4.5}$$

$$= \prod_{l=1}^{2} P(k_i^{(l)}|L, N_i^{(1)}, N_i^{(2)}). \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}, |L, N_i^{(1)}, N_i^{(2)}) \times P(L), \tag{4.6}$$

and similarly, the denominator becomes

$$P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|N_i^{(1)}, N_i^{(2)}) \tag{4.7}$$

$$= P(L)P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|L, N_i^{(1)}, N_i^{(2)}) \tag{4.8}$$

$$+ P(\neg L)P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|\neg L, N_i^{(1)}, N_i^{(2)}). \tag{4.9}$$

Assuming further that $k_i^{(l)}$ only depends on $N_i^{(l)}$ and that $\mathbf{c}_i^{(j)}$ does not depend

on linkage graphs,

$$\prod_{l=1}^{2} P(k_i^{(l)}|L, N_i^{(1)}, N_i^{(2)}). \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}, |L, N_i^{(1)}, N_i^{(2)}) \times P(L) \tag{4.10}$$

$$= \prod_{l=1}^{2} P(k_i^{(l)}|L, N_i^{(l)}) \times \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}|L) \times P(L), \tag{4.11}$$

and

$$P(L)P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|L, N_i^{(1)}, N_i^{(2)}) \tag{4.12}$$

$$+ P(\neg L)P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}|\neg L, N_i^{(1)}, N_i^{(2)}) \tag{4.13}$$

$$= P(L) \prod_{l=1}^{2} P(k_i^{(l)}|L, N_i^{(l)}) \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}|L) \tag{4.14}$$

$$+ P(\neg L) \prod_{l=1}^{2} P(k_i^{(l)}|\neg L, N_i^{(l)}) \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}|\neg L). \tag{4.15}$$

Similarly to the other formulations in the literature [1, 32], $P(k_i^{(l)}|L, N_i^{(l)})$ and $P(k_i^{(l)}|\neg L, N_i^{(l)})$ are calculated assuming the binomial distribution. $P(L)$ is the prior probability that gene i is annotated with term t and is calculated as a frequency of term t among genes.

4.3 Experiments and results

For the 5-fold cross-validation study, we created each test set by eliminating all annotations from a random 20% of annotated proteins (250 randomly chosen proteins from the annotated set of 1249). We performed 5 validation runs and report the average of these for the summary statistics. We use the statistical measures

$sensitivity$ and $specificity$, as defined in [79]. We also use the $F1$ measure which represents a weighted harmonic mean of precision and recall and is defined as

$$F1 = \frac{2 \times (precision \times recall)}{precision + recall}$$

Note that $F1$ allows analysis of the performance weighing precision and recall evenly.

4.3.1 Gene expression data of a parasite life-cycle

First, we show and emphasize the importance of gene expression data representation and analysis, especially when applied to parasites. Many parasites, such as malaria parasites, trypanosomes, endoparasites with larval stages (tapeworms, thorny-headed worms, flukes, parasitic roundworms), undergo many changes during their various life-cycle stages as they travel from one host to the other, or from one organ or system to another. Each stage requires utilization of different life functions and possible metamorphosis, which involves up-regulation of necessary genes and/or down-regulation of those not crucial for a specific life-cycle period.

In this study we use the five time windows of the intraerythrocytic developmental cycle (IDC) of *P. falciparum* identified by Kleinberg et al. [78]. This expression data is particularly interesting since the IDC, or blood stage, is the phase responsible for malaria symptoms in humans. This study [78] performs the time series segmentation and clustering of the data concurrently. Their method is formulated in terms of rate distortion theory—it searches for a compressed description of the data (i.e. the fewest clusters of expression profiles, obtained after

an optimal temporal segmentation), while minimizing the distortion introduced by this compression. More formally, this process is characterized by a variational formulation:

$$\mathcal{F}_{min} = I(Z;X) + \beta\langle d(x,z)\rangle, \tag{4.16}$$

where mutual information and average distortion are defined as:

$$I(Z;X) = \sum_{x,z} p(z|x)p(x)log\frac{p(z|x)}{p(z)} \tag{4.17}$$

$$\langle d(x,z)\rangle = \sum_{x,z} p(x)p(z|x)d(x,z), \tag{4.18}$$

and

$$d(x,z) = \sum_{x_1} p(x_1|z)d(x_1,x). \tag{4.19}$$

Then, the set of candidate windows (i.e., enumeration of all possible windowings within constraints on the min and max allowed window sizes) is created, and the data is clustered within each window according to Eq. (4.16). Each window is then scored, based on its length and Eq. (4.16). To find the optimal windowing of the data, they formulate the problem as one of graph search and use a shortest path algorithm to find a combination of windows that jointly provide the lowest cost. For the *P. falciparum* data the study in Kleinberg et al. [78] found the critical time points at 7, 16, 28 and 43 hours, leading to 5 windows, sized non-uniformly. These windows correspond to the three IDC stages and the transitions between them: End Merozoite/Early Ring stage, Late Ring stage/ Early Trophozoite stage, Trophozoite, Late Trophozoite/ Schizont, and Late Schizont/Merozoite.

The clustering by Kleinberg et al. [78] identified 4-5 clusters per window,

corresponding to the three phases of the cycle with an additional one or two clusters per window containing terms regulating the beginning or end of a phase. In order to predict detailed functional annotations, we decided to cluster the data more finely. We use these previously identified windows and clustered the expression profiles within each separately, using the k-means clustering algorithm. We then define d_r^j as a random variable indicating if a protein belongs in the cluster r within window j. The sequence of random variables for each window then constitutes a categorical feature vector **d** of a protein.

We experimented with various values for k and compared results with the linkage graph defined by a Pearson coefficient calculation; we performed this step for all pairs of genes for the entire data set.

In our experiments, due to a high number of negative annotations for the *P. falciparum* dataset, specificity reaches 0.9 immediately after the threshold for posterior probability goes above 0.05. In this case a ROC curve, as shown in Figure 4.1, does not reflect a precise sensitivity-specificity relationship as expected in other cases, obtained with a relatively large amount of data. As a result, it is necessary to use a more sensitive statistical measure that would account for too high or too low statistical values, e.g., a metric computed by taking their harmonic mean. In particular, we aim to maximize the $F1$ statistic, which reflects a relationship of recall to precision, as noted in Figure 4.2. Note that $F1$ will be maximized only if both measures are maximized.

As shown in Figures 4.2, the variation in the number of clusters, k, does not distort the predictive value of the method as for all values of k in this range, the method yields nearly identical ROC and F1 curves. Figure 4.2 also shows the superiority of time-dependent k-means clustering over the Pearson correlation coef-

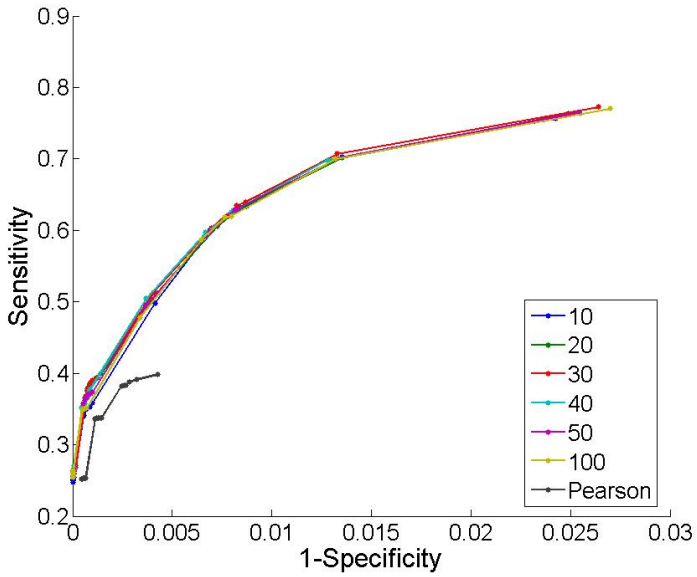


Figure 4.1: The ROC curve of recall experiment by 5-fold cross validation for gene expression data. Numbered legends correspond to k-means clustered datasets.

ficient (in the majority of cases the Pearson curve is completely below the curves for the clustered data). The linkage graph defined by the Pearson correlation coefficient was built using 286620 edges (a protein pair is considered co-expressed if its Pearson coefficient is larger than 0.85 [1]) and covered 2646 proteins.

Since for all values of k both figures showed nearly identical ROC and F1 curves, we fixed it at an arbitrary value, $k = 30$, for the following analysis.

4.3.2 Analysis of prediction accuracy

We compare runs on individual data sources with runs which integrate PPI, sequence similarity, metabolic pathway information, and temporal gene expression data. Our first step is to analyze how well our method predicts known protein-term

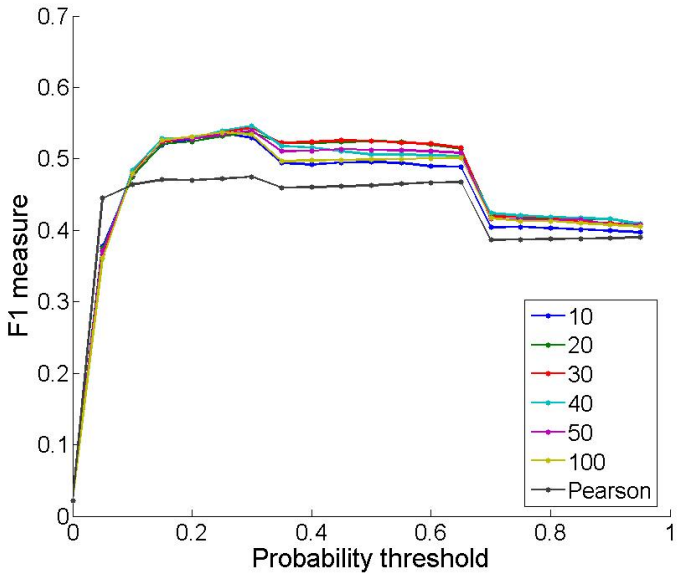


Figure 4.2: The F1 statistics of recall experiment by 5-fold cross validation for gene expression data (posterior probability thresholds range from 0.05 to 0.95, in 0.05 increments). Numbered legends correspond to k-means clustered datasets.

associations, using 5-fold cross validation. We predict that a gene i is annotated with term t if the probability exceeds a specified threshold.

Figures 4.3 and 4.4 summarize the positive impact of data integration (PPI, sequence similarity, metabolic pathway, window-based gene expression clustering) on protein function prediction via ROC and $F1$ measures, respectively. However, since ROC curves are very much influenced by the large number of negative annotations in *P. falciparum* data (similarly to Figure 4.1), specificity reaches 0.9 immediately after the threshold for posterior probability goes above 0.05), this measure is not very sensitive with respect to specificity scores; thus, we prefer the F1 statistic, which uses the harmonic mean of precision and recall. In these figures, we also show the statistics for the data obtained by analysis of gene ex-

pression using Pearson correlation coefficients (showing a clear disadvantage), although it was not a part of the data integration.

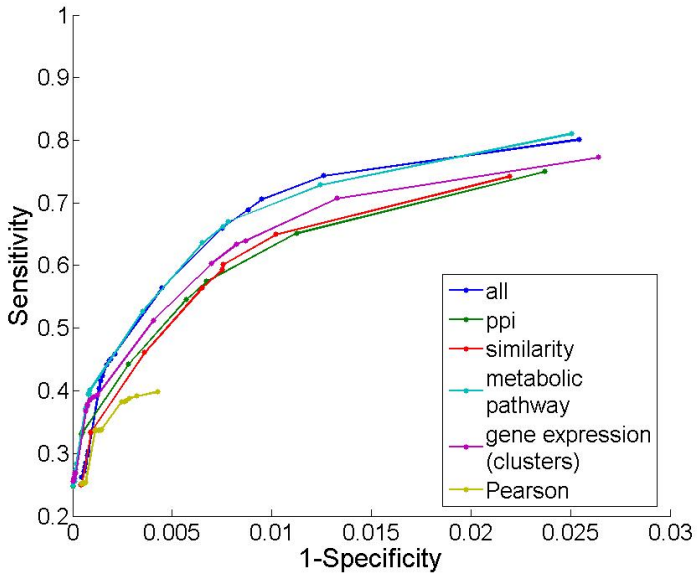


Figure 4.3: The ROC curves for individual data sources and integrated data.

Additionally, we investigated the impact of gene expression data on data integration. In Figures 4.5 and 4.6, we show both ROC and F1 curves, respectively, for fused data (PPI, similarity, and metabolic pathway) alone, then for fused data together with the windowed and clustered gene expression data, and fused data with Pearson coefficient defined data. Clustered temporal gene expression data shows a distinctive positive impact on the overall predictive power of the method; however, Pearson coefficient data has a negative effect on ROC and F1 statistics. Most likely this anomaly is due to a large number of falsely defined associations between co-expressed genes.

Figure 4.7 shows the impact of data integration on the number of TP at two precision levels: 50% and 70%. These two levels of precision are reasonably

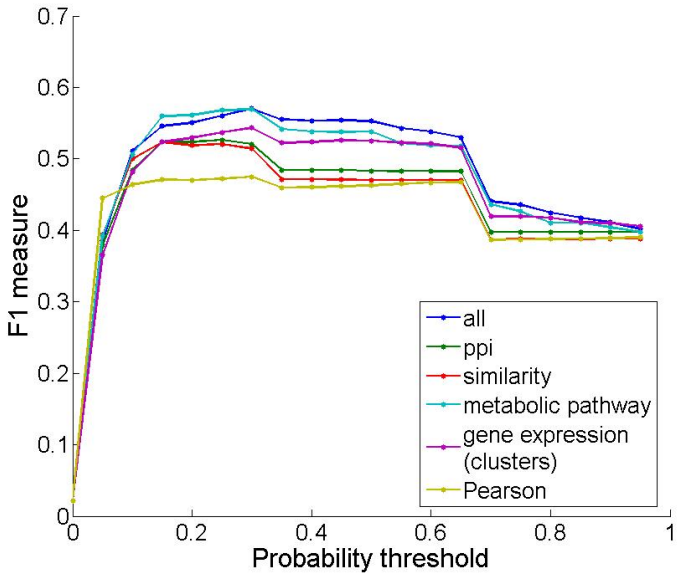


Figure 4.4: The F1 statistics for individual data sources and integrated data (posterior probability thresholds ranges from 0.05 to 0.95, in 0.05 increments).

accurate of the range of possible improvements in our study, and the TP number is calculated when the precision level first hits the specified margin. In Table 4.1, we summarize the improvements of data integration over individual sources and conclude that data integration significantly outperforms individual data sources at 70% precision, which corresponds to 0.35 threshold of posterior probability for function prediction. This probability threshold now can be applied in the second step of our study: attempting to predict functions for the unannotated proteins of *P. falciparum*.

In the second part of our study, we trained our method on all annotated proteins and tried to assign functions to proteins without annotations. By integrating PPI data, sequence similarity, metabolic pathway, and clustered temporal window-based gene expression data we were able to assign probable GO terms

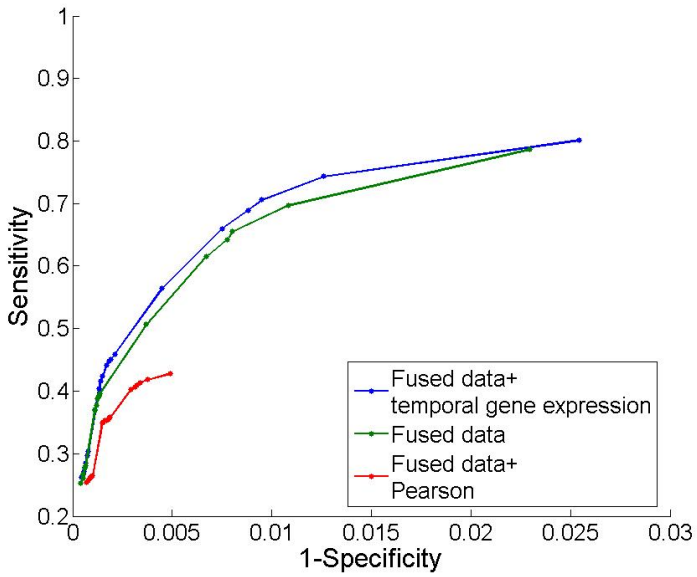


Figure 4.5: The ROC curves for various ways of integrating data:"fused" is defined as ppi+similarity+metabolic pathway.

Data source	50% precision	70% precision
PPI	14%	20%
Sequence similarity	17%	23%
Metabolic pathway	5%	13%
Gene expression (clustering)	11%	10%

Table 4.1: % of improvements of data integration on #TP over individual data sources

to 628 out of 1439 unannotated proteins of *P. falciparum*. We ignored general terms, such as those high up in the GO hierarchy, that appeared more than 300 times. We report more than 2500 gene-GO assignment pairs, which can be viewed at: `http://www.cims.nyu.edu/~antonina/real_output.txt`. The GO terms are reported together with their parents (ancestors) in the GO hierarchy. In Figure 4.8, we present cumulative statistics for the number of predicted functional assignments and probability thresholds they satisfy. As shown

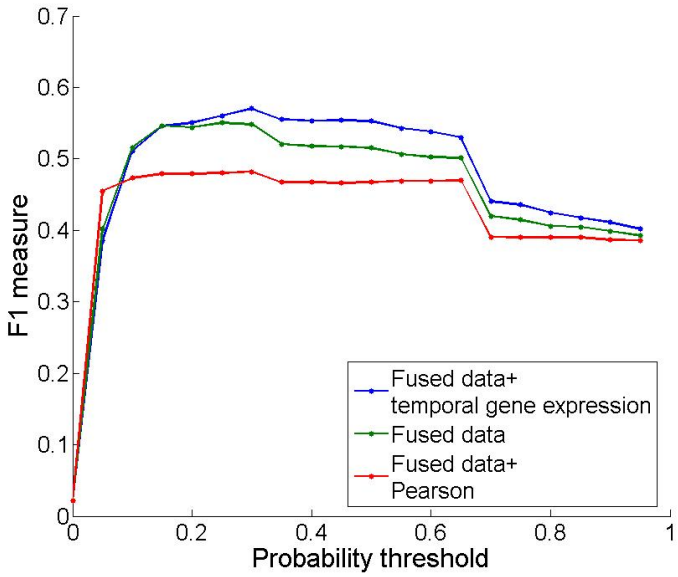


Figure 4.6: The F1 statistics for various ways of integrating data:"fused" is defined as ppi+similarity+metabolic pathway (posterior probability thresholds ranges from 0.05 to 0.95, in 0.05 increments).

in Figure 4.8, by varying the original probability threshold, we can narrow down the possible set of predictions. For example, probability threshold at 0.8 (80%) would correspond to about 500 functional assignments of higher probability.

4.4 Functional predictions for pharmaceutical targeting

The fundamental goal of our study is to assign functions to unannotated *P. falciparum* proteins in order to find possible vaccine and drug targets. For this purpose, we analyzed all predicted functional assignments made by our computational technique to determine if they are related to erythrocytic adhesion and

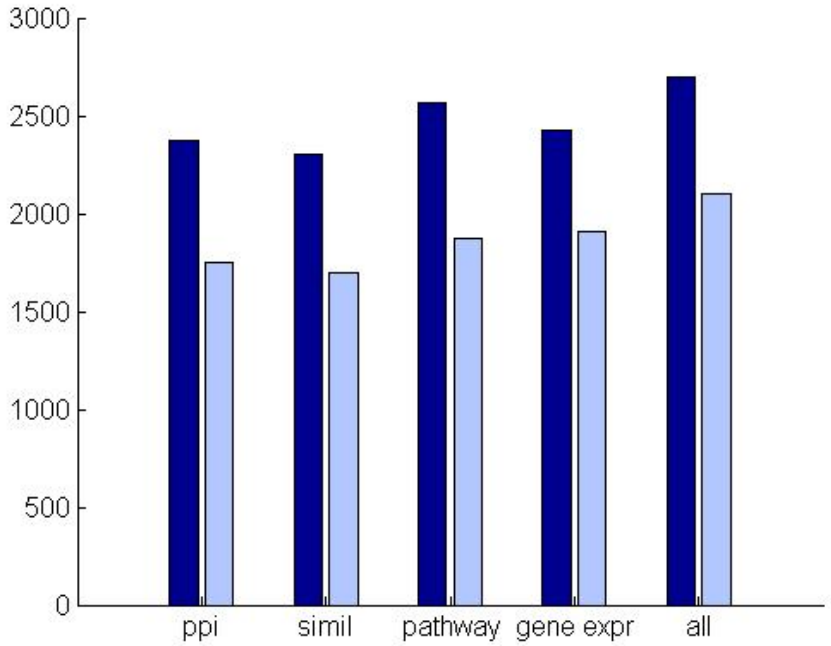


Figure 4.7: Number of True Positive predictions at 50% precision (dark blue) and at 70% precision (light blue).

modification. In particular, we paid close attention to the *P. falciparum* surface proteins responsible for binding of the parasite to human erythrocytes, and to the *P. falciparum* red blood cell (RBC) membrane proteins responsible for the parasite's intraerythrocytic survival and for the adhesion of the RBC to capillary vessels. In our predicted dataset of 628 proteins, 20 are identified as RBC membrane proteins (contributing to 78 functional predictions) and one protein is identified as an erythrocyte binding protein (contributing to two functional predictions).

We further label RBC membrane proteins with one of the address tags: either *Plasmodium export element* (Pexel) or *N-terminal host targeting* (HT) motif. Both of these motifs are responsible for the transport of *P. falciparum* proteins inside erythrocytic cytoplasm, as detailed below.

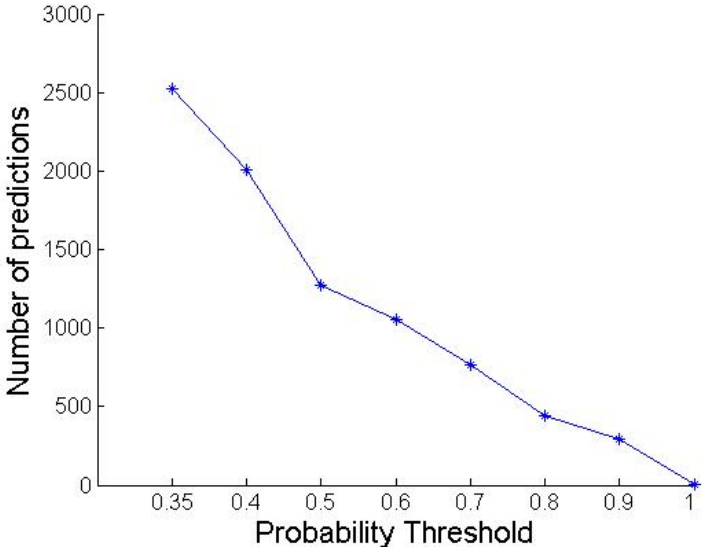


Figure 4.8: Number of possible predictions as a function of probability threshold. Each point corresponds to the number of predicted functional assignments whose probability is greater or equal to the corresponding probability threshold.

During the blood stage of malaria, *P. falciparum* actively penetrates human erythrocytes. In the process of invasion, the parasite initiates the formation of a unique membrane, the parasitophorous vacuole membrane, which surrounds the parasite inside the invaded erythrocyte. The parasitophorous vacuole isolates the parasite and protects it from the host's defenses, such as lysozymal attack.

P. falciparum needs to develop its own strategy in order to survive and feed inside human erythrocytes, since red blood cells lose their nuclei, ability to synthesize new proteins, and vesicular transport system during their formation. Residing inside a red blood cell, *P. falciparum* injects hundreds of its own proteins into erythrocytic cytoplasm [80] to build its living environment. The injected proteins then interact with proteins of the erythrocytic membrane skeleton and induce substantial changes in the morphology and function of the erythrocytic

cell. Such changes include development of various membraneous (tubulovesicular structures and Maurer's clefts) networks from the vacuole to the erythrocyte membrane, which are needed for parasite's nutrient uptake, and protrusion of the erythrocyte membrane in a form of electron-dense elevations called adhesive knobs [81].

To reach the erythrocytic cytoplasm and membrane, *P. falciparum* exported proteins have to traverse a series of physical barriers: parasite membrane, parasitophorous vacuole membrane, and sometimes erythrocytic membrane [81, 82]. First, proteins are exported form the parasite into the vacuolar space following the typical secretion pathway existing in all eucaryotic cells. However, a special mechanism is needed to cross the parasitophorous vacuole membrane and reach the erythrocytic cytoplasm. For the majority of *P. falciparum* proteins, an N-terminus host targeting (HT) motif [82, 83] is required to cross the vacuole membrane.

On the other hand, Pexel is a Plasmodium export element (related to, but distinct from, HT) that is responsible for the transport of soluble *P. falciparum* proteins into the erythrocyte cytoplasm through the parasitophoros vacuole membrane [82].

Exported proteins then interact with the erythrocytic membrane causing its deformation and knobbing. Knobs mediate cytoadherence of infected erythrocytes to capillary blood vessels. In this way, the infected cells hide in an attempt to avoid elimination in the spleen. Such massive accumulation of infected red blood cells in the capillary blood vessels of the brain and kidneys can lead to organ failure and ultimately death. Thus, targeting the parasite's RBC membrane proteins could aid the development of interventions that block the parasite's growth

or limit the severity of the disease.

As reported in the PlasmoDB [77] database, there are 195 RBC membrane proteins containing HT motif and 293 RBC membrane proteins containing Pexel. We predicted functions for 20 RBC membrane proteins containing either of the motifs. The list of RBC membrane proteins (with their predicted GO functions) containing both motifs is shown in Table 4.2 and the list of proteins containing one of the motifs is shown in Table 4.3. Some interesting examples, which could become future pharmaceutical targets, include RBC membrane proteins PFD0495c and PFE0040c assigned the gene ontology term GO:0007155 (cell adhesion) with probability 70% and 99% respectively. Furthermore, close attention should be paid to gene ontology terms responsible for reaction to outside stimulus, as those can play a crucial role in the parasite's survival. For example, RBC membrane protein PFE1605w, assigned GO terms GO:0009628 (response to abiotic stimulus) with probability 80% and GO:0042221 (response to chemical stimulus) with probability 68%, could be a promising drug target.

Finally, there exist 10 *P. falciparum* surface proteins responsible for binding of the parasite to erythrocyte surface ligands, as reported by [77]. Following the establishment of a tight interaction between the parasite and the RBC, entry is initiated by the activation of actin-myosin motor so that the parasite forces the invagination of the erythrocytic membrane with formation of the the parasitophorous vacuole membrane, described earlier. The only surface *Plasmodium* protein, PFE0340c, present in our predicted dataset is assigned GO terms GO:0006511 (ubiquitin-dependent protein catabolism) and GO:0019941 (modification-dependent protein catabolism) both with probability close to 63%.

Protein ID	Probability	GO term
PFD0070c	0.401545231581066	GO:0043412 biopolymer modification
PFD0125c	0.49387370405278	GO:0006412 protein biosynthesis
	0.494801512287335	GO:0009059 macromolecule biosynthesis
	0.513352425272955	GO:0009058 biosynthesis
	0.512411033603626	GO:0044249 cellular biosynthesis
PFD0495c	0.580096975765904	GO:0006412 protein biosynthesis
	0.581846879650176	GO:0009059 macromolecule biosynthesis
	0.603310008429891	GO:0009058 biosynthesis
	0.602215349980034	GO:0044249 cellular biosynthesis
	0.699684816632941	GO:0007155 cell adhesion
PFD1020c	0.7280979853935	GO:0006631 fatty acid metabolism
PFD1170c	0.422344888143171	GO:0044267 cellular protein metabolism
	0.423250325426377	GO:0044260 cellular macromolecule metabolism
	0.432582635313454	GO:0019538 protein metabolism
	0.999999991137921	GO:0006457 protein folding
PFE0040c	0.98843772424871	GO:0007155 cell adhesion
PFE0060w	0.457539670421109	GO:0006468 protein amino acid phosphorylation
	0.521369701293125	GO:0006796 phosphate metabolism
	0.521369701293125	GO:0006793 phosphorus metabolism
	0.437130777000256	GO:0016310 phosphorylation
MAL7P1.170	0.476701149188691	GO:0006810 transport
	0.477943559067398	GO:0051234 establishment of localization
	0.477943559067398	GO:0051179 localization
PF07_0132	0.351905883602301	GO:0019538 protein metabolism
PFI1785w	0.365533239904503	GO:0019538 protein metabolism
	0.999794367636718	GO:0006457 protein folding
PF11_0508	0.375957294327154	GO:0006464 protein modification
PF13_0073	0.739599615802019	GO:0006412 protein biosynthesis
	0.733078093566159	GO:0009059 macromolecule biosynthesis
	0.751850612028445	GO:0009058 biosynthesis
	0.756174623097316	GO:0044249 cellular biosynthesis
PF13_0076	0.358502146993282	GO:0006810 transport
	0.360756570068609	GO:0051234 establishment of localization
	0.360756570068609	GO:0051179 localization
PF13_0275	0.358502146993282	GO:0006810 transport
	0.360756570068609	GO:0051234 establishment of localization
	0.360756570068609	GO:0051179 localization

Table 4.2: RBC membrane proteins possessing HT motif and Pexel, their predicted functions, and corresponding probabilities.

Protein ID	Probability	GO term
Pexel only:		
PFA0225w	0.487145531811038	GO:0043037 translation
	0.461767651699677	GO:0009058 biosynthesis
	0.453487695127242	GO:0044249 cellular biosynthesis
	0.539833659529562	GO:0006082 organic acid metabolism
	0.539833659529562	GO:0019752 carboxylic acid metabolism
	0.386239267057058	GO:0008610 lipid biosynthesis
	0.388721047331319	GO:0006629 lipid metabolism
	0.374685825510083	GO:0044255 cellular lipid metabolism
PFD0080c	0.665278152637513	GO:0006464 protein modification
	0.511094565590855	GO:0043412 biopolymer modification
	0.799046995682858	GO:0006468 protein amino acid phosphorylation
	0.684531881462785	GO:0006796 phosphate metabolism
	0.684531881462785	GO:0006793 phosphorus metabolism
	0.704629372061098	GO:0016310 phosphorylation
PFE1605w	0.802543920254657	GO:0044267 cellular protein metabolism
	0.754472317644877	GO:0044260 cellular macromolecule metabolism
	0.805311582644175	GO:0019538 protein metabolism
	0.72021237129596	GO:0006950 response to stress
	0.70563593694521	GO:0050896 response to stimulus
	0.801087730125495	GO:0009628 response to abiotic stimulus
	0.680269187401183	GO:0042221 response to chemical stimulus
	0.999999999999956	GO:0006457 protein folding
	0.501328353480413	GO:0007155 cell adhesion
MAL7P1.7	0.400917810998465	GO:0006082 organic acid metabolism
	0.400917810998465	GO:0019752 carboxylic acid
	0.999999999989668	GO:0006457 protein folding
PFI1780w	0.653500492148364	GO:0043037 translation
	0.877574807784855	GO:0006412 protein biosynthesis
	0.871609237465261	GO:0009059 macromolecule biosynthesis
	0.874807040307226	GO:0009058 biosynthesis
	0.356761397372017	GO:0044260 cellular macromolecule metabolism
HT motif only:		
PF13_0317	0.781092830960702	GO:0044267 cellular protein metabolism
	0.781906300484652	GO:0044260 cellular macromolecule metabolism
	0.78205791106515	GO:0019538 protein metabolism
	0.373719533733663	GO:0043037 translation
	0.781559322033898	GO:0006412 protein biosynthesis
	0.782192339038305	GO:0009059 macromolecule biosynthesis
	0.818767547332805	GO:0009058 biosynthesis
	0.818207742211449	GO:0044249 cellular biosynthesis

Table 4.3: RBC membrane proteins possessing only Pexel motif or only HT motif, their predicted functions, and corresponding probabilities

4.5 Discussion and conclusions

In this work, we have applied and evaluated a probabilistic approach for predicting protein functions for the malaria parasite *Plasmodium falciparum*. We combined four sources of information using a unified probabilistic framework. PPI and sequence similarity data were presented in the form of functional linkage graphs, since such data imply the importance of the number and GO annotation of the nearest neighbors. Metabolic pathway and temporal gene expression data were encoded using categorical feature vectors, simplifying the search for similar feature patterns among related proteins.

We emphasized the importance of the data representation for parasites, though this might not necessarily apply to non-parasitic organisms. In particular, a malaria parasite's life cycle is affected by change of the host (e.g., mosquito and human), tissues (e.g., salivary glands, blood, gut wall, liver, red blood cells), and possible developmental changes of the parasite itself (e.g., gametocytes, sporozoites, merozoites). Each such change involves different mechanisms for gene regulation and employs many specific life-sustaining genes. Thus, it becomes crucial to analyze gene expression data from each stage separately, as opposed to calculating Pearson correlation coefficients for all pairs regardless of their temporal order. We have demonstrated that the data representation, which takes advantage of the temporal order of gene expression patterns, leads to a clear improvement in statistical significance over function predictions using simple Pearson coefficient calculations.

We show that data integration, previously shown to be beneficial for protein function prediction [1, 35, 55], is crucial when applied to organisms with limited

individual data sources, as in the case of parasites. Even more importantly, the proposed "windowing" of the IDC provides a clear advantage to the data integration, dramatically improving its predictive performance. By embedding various data sources into the probabilistic framework, we have been able to assign functions to 628 previously unannotated *P. falciparum* proteins and expect to find in those some of the most promising candidates for future vaccine trials. Toward this, we have suggested a number of possible RBC membrane proteins that should be explored further.

To extend this study to include ortholog genes, we next tested our method by integrating PPI data of another closely-related malaria parasite *P. vivax* (in particular, we used only PPI data of close orthologs with *P. falciparum*), and were encouraged by the significant improvement in the resulting performance scores and a much improved F1 curve. However, we have omitted further details of these improved results, since the *P. vivax* genomic data await publication and remain publicly unavailable. Once these data are published, we plan to disseminate the improved results through our laboratory website.

We believe that this work will pave the way for more complex automatic annotation algorithms based on model checking with temporal-logic queries—in this picture, one would obtain a succinct Kripke model (a phenomenological model) that summarizes the most important synchronization properties exhibited by a set of temporal data streams; then use these Kripke models to infer properties satisfied in various states (also called possible-worlds) of the model; and finally, associate these properties with functional classes and genes active in these states of the Kripke model. It should also be obvious that, at first, such a method is likely to be employed as a debugging tool for existing ontologies: particularly, to

check if certain ontology terms are being associated incorrectly or inconsistently with a bio-molecule.

4.6 Web Resources and Supplementary material

The list of all predicted functions is available from

$http://www.cims.nyu.edu/\sim antonina/real\_output.txt$

Chapter 5

Conclusion

We live in the era of networks. Living entities interact with each other in more and more complex manners: through social networks; exchanging an exponentially growing traffic of information via World Wide Web networks; and staying alive through complex inter-regulations within biochemical networks. In response, we have begun to see a dramatic growth in quantitative as well as qualitative studies of networks, all aimed at elucidating fundamental concepts of such complex systems. The genomic advances of our time have been many and spectacular: e.g., the DNA double helix structure, gene expression profiles, human genome sequencing, genetic engineering, decoding protein structure, discovery of antibiotics etc. They have opened new horizons for understanding complex, biologically important relationships among a cell's active entities (such as genes regulating other genes, proteins interacting with genes, proteins interacting with small molecules, proteins involved in complexes with other proteins, etc) and given a concrete meaning to biological networks with a complexity and diversity that far exceed anything that the human ingenuity has been able to spawn so far.

This thesis provides a rigorous foundation for the understanding of topology, functionality, and complexity of the protein networks as well as more accurate and efficient algorithms for network analysis. We develop algorithms to understand the temporal aspects of network and to do so at multiple time-scales, while taking into account how information is distributed in the network at a fast time-scale, and how topology is modified over a slow time-scale.

In particular, we demonstrated that re-defining network connectivity (while looking at it from a dynamic, intra-species developmental scale) produces more efficient and robust algorithms for identifying densely connected regions of Y2H PPI network, commonly referred to as protein complexes. On the other hand, evolutionary information, encoded as conserved inter-species homology, provides an essential learning framework for proteins with poor neighborhood in their own network. Finally, we proved that dynamic time-course data is essential for understanding protein functions necessary for identifying vaccine and drug targets in malaria parasite.

To broaden the impact of our research, in addition to developing practical algorithms for network analysis, we publicly distributed computer software implementing these algorithms as well as developed datasets published electronically and shared with other scientific and research groups. Results presented in this study are hoped to be widely used by biological, medical, and pharmaceutical laboratories and play a promising role on drug and vaccine development, personalized medicine, and bioinformatics advances.

Bibliography

[1] N. Nariai, E. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS ONE*, vol. 2, no. 3, 2007.

[2] Y. Ho, A. Gruhler, and A. Heilbut, "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[3] A. Gavin, M. Bosche, and R. Krause, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[4] R. Sharan, T. Ideker, B. Kelley, R. Shamir, and B. Karp, "Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data," in *Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB)* (P. Bourne and D. Gusfield, eds.), pp. 282–289, ACM, 2004.

[5] A. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–53, 2003.

[7] G. Bader and C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 2, 2003.

[8] V. Spirin and L. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the USA*, vol. 100, no. 21, pp. 12123–12128, 2003.

[9] M. Meila, "Comparing Clusterings by the Variation of Information," in *16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop* (B. Schlkopf and M. Warmuth, eds.), pp. 173–187, Lecture Notes in Computer Science, 2003.

[10] S. Brohee and J. Van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 488, 2006.

[11] C. Stark, J. Breitkreutz, B, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: A General Repository for Interaction Datasets," *Nucleic Acids Research*, vol. 34, pp. D535–9, January 2006.

[12] H. Mewes, D. Frishman, U. Guldener, G. Mannhaup, K. Mayer, M. Mokrejs, B. Morgenstein, M. Munsterkotter, S. Rudd, and B. Weil, "MIPS: a database

for genomes and protein sequences," *Nucleic Acid Research*, vol. 30, no. 1, pp. 31–24, 2002.

[13] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein-protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[14] E. Hirsh and R. Sharan, "Identification of conserved protein complexes based on a model of protein network evolution," *Bioinformatics*, vol. 23, no. 2, pp. e170–e176, 2007.

[15] R. Gomory and T. Hu, "Multi-Terminal Network Flows," *Journal of SIAM*, vol. 9, no. 4, pp. 551–570, 1961.

[16] T. Hu and M. Shing, *Combinatorial Algorithms*. Mineola, New York: Dover Publications INC, enlarged 2nd ed. ed., 2002.

[17] D. Barth, P. Berthome, and M. Diallo, "On the analysis of gomory-hu cut trees relationship," *Research report*, 2002.

[18] V. Vazirani, *Approximation Algorithms*. New York, NY: Springer, first ed., 2003.

[19] L. Ford and D. Fulkerson, *Flows in networks*. Princeton, NJ: Princeton University Press, 1973.

[20] D. Matula, "Determining edge connectivity in O(n,m) ," in *28th IEEE Symposium on Foundations of Computer Science*, pp. 249–251, IEEE Computer Society, 1987.

[21] H. Nagamochi and T. Ibaraki, "Computing edge connectivity in multigraphs and capacitated graphs," *SIAM Journal of Discrete Mathematics*, vol. 5, no. 1, pp. 54–66, 1992.

[22] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[23] I. Lee, S. Dae, A. Adai, and E. Marcotte, "A probabilistic functional Network of Yeast genes," *Science*, vol. 306, no. 5701, pp. 1555–1558, 2004.

[24] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its supplement TrEMBL," *Nucleic Acids Research*, vol. 26, no. 1, pp. 38–42, 1998.

[25] `http://973-proteinweb.ustc.edu.cn/hsppip/faq.php`.

[26] J. Enright, S. Van Dongen, and A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[27] G. Lerman and B. E. Shakhnovich, "Defining functional distance using manifold embeddings of gene ontology annotations," *Proceedings of the National Academy of Sciences of the USA*, vol. 104, no. 27, pp. 11334–11339, 2007.

[28] G. W. Flake, R. Tarjan, and K. Tsioutsiouliklis, "Graph Clustering and Minimum Cut Trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2004.

[29] M. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 036104, 2006.

[30] F. Luo, Y. Yang, C.-F. Chen, R. Chang, J. Zhou, and R. Scheuermann, "Modular organization of protein interaction networks," *Bioinformatics*, vol. 23, no. 2, pp. 207–214, 2007.

[31] E. Hartuv and R. Shamir, "A Clustering Algorithm based on Graph Connectivity ," *Information Processing Letters*, vol. 76, no. 4-6, pp. 175–181, 2000.

[32] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19, no. 1, pp. i197–i204, 2003.

[33] U. Karaoz, T. Murali, S. Letovsky, Y. Zheng, C. Ding, C. Cantor, and S. Kasif, "Whole-genome annotation by using evidence integration in functional-linkage networks," *Proceedings of the National Academy of Sciences of the USA*, vol. 101, no. 9, pp. 2888–2893, 2004.

[34] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[35] S. Carroll and V. Pavlovic, "Protein classification using probabilistic chain graphs and the gene ontology structure," *Bioinformatics*, vol. 22, no. 15, pp. 1871–1878, 2006.

[36] B. Engelhardt, M. Jordan, K. Muratore, and S. Brenner, "Protein molecular function prediction by bayesian phylogenomics," *PLoS Computational Biology*, vol. 1, no. 5, p. e45, 2005.

[37] A. Mitrofanova, S. Kleinberg, J. Carlton, S. Kasif, and B. Mishra, "Systems biology via redescription and ontologies (iii): Protein classification using malaria parasite's temporal transcriptomic profiles," in *IEEE International Conference on Bioinformatics and Biomedicine* (X.-W. Chen, X. Hu, and S. Kim, eds.), pp. 278–283, IEEE Computer Society, 2008.

[38] N. Yosef, R. Sharan, and W. Stafford Noble, "Improved network-based identification of protein orthologs," *Bioinformatics*, vol. 24, no. 16, pp. i200–i206, 2008.

[39] J. Liu and B. Rost, "Comparing function and structure between entire proteomes," *Protein Science*, vol. 10, no. 10, pp. 1970–1979, 2001.

[40] J. Whisstock and A. Lesk, "Prediction of protein function from protein sequence and structure," *Quarterly Review of Biophysics*, vol. 36, no. 3, pp. 307–340, 2003.

[41] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant, and R. Apweiler, "The proteome analysis database: a tool for the in silico analysis of whole proteomes," *Nucleic Acids Research*, vol. 31, no. 1, pp. 414–417, 2003.

[42] M. Deng, T. Chen, and F. Sun, "An integrated probabilistic model for functional prediction of proteins," in *Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB)* (W. Miller, ed.), pp. 95–103, ACM, 2003.

[43] M. Deng, Z. Tu, F. Sun, and T. Chen, "Mapping gene ontology to proteins based on protein-protein interaction data," *Bioinformatics*, vol. 20, no. 6, pp. 895–902, 2004.

[44] K. Tsuda, H. Shin, and B. Scholkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. 2, pp. ii59–ii65, 2005.

[45] H. Shin, A. Lisewski, and O. Lichtarge, "Graph sharpening plus graph integration: a synergy that improves protein functional classification," *Bioinformatics*, vol. 23, no. 23, pp. 3217–3224, 2007.

[46] A. Vinayagam, R. Konig, J. Moormann, R. Schubert, F. ad Eils, K.-H. Glatting, and S. Suhai, "Applying support vector machines for gene ontology based gene function prediction," *BMC Bioinformatics*, vol. 5, no. 116, 2004.

[47] T. Hawkins, S. Luban, and S. Kihara, "Enhanced automated function prediction using distantly related sequences and contextual association by pfp," *Protein Science*, vol. 15, no. 6, pp. 1550–1556, 2006.

[48] D. Martin, M. Berriman, and G. Barton, "Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC Bioinformatics*, vol. 5, no. 178, 2004.

[49] S. Geman and D. Geman, "Stochastic relaxation, gibbs distribution and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[50] S. Lauritzen, *Graphical Models*. New York: Oxford University Press, 1996.

[51] `http://www.yeastgenome.org/`.

[52] http://www.flybase.org/.

[53] B. Breitkreutz, C. Stark, and M. Tyers, "The grid: the general repository for interaction datasets," *Genome Biology*, vol. 4, no. 3, p. R23, 2003.

[54] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, "The gene ontology consortium. gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[55] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Integrative protein function transfer using factor graphs and heterogeneous data sources," in *IEEE International Conference on Bioinformatics and Biomedicine* (X.-W. Chen, X. Hu, and S. Kim, eds.), pp. 314–318, IEEE Computer Society, 2008.

[56] J. Demsar, "Statistical comparison of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[57] E. Clarke, O. Grumberg, and D. Peled, *Model Checking*. The MIT Press, 1999.

[58] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[59] I. Lee, S. Date, A. Adai, and E. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, no. 5701, pp. 1555–8, 2004.

[60] L. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," *Genome Research*, vol. 15, no. 7, pp. 945–53, 2005.

[61] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences of the USA*, vol. 100, no. 14, pp. 8348–53, 2003.

[62] S. Altschul, T. Madden, A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[63] J. Liu and B. Rost, "CHOP proteins into structural domain-like fragments," *Proteins: Structure, Function, and Bioinformatics*, vol. 55, no. 3, pp. 678–688, 2004.

[64] K. Nakai and P. Horton, "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends in Biochemical Sciences*, vol. 24, no. 1, pp. 34–36, 1999.

[65] R. Nair, P. Carter, and B. Rost, "NLSdb: database of nuclear localization signals," *Nucleic Acids Researh*, vol. 31, no. 1, pp. 397–399, 2003.

[66] B. Rost, J. Liu, R. Nair, K. Wrzeszczynski, and Y. Ofran, "Automatic prediction of protein function," *Cellular and Molecular Life Sciences*, vol. 60, no. 12, pp. 2637–2650, 2003.

[67] A. Valencia and F. Pazos, "Computational methods for the prediction of protein interactions," *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 368–373, 2002.

[68] M. Galperin and E. Koonin, "Who's your neighbor? new computational approaches for functional genomics," *Nature Biotechnology*, vol. 18, no. 6, pp. 609–613, 2000.

[69] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," *Journal of Molecular Biology*, vol. 301, no. 4, pp. 1059–1075, 2000.

[70] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21, no. 1, pp. i302–i310, 2005.

[71] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing* (R. Altman, A. Dunker, L. Hunter, and T. Klein, eds.), pp. 418–429, World Scientific, 2000.

[72] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the USA*, vol. 95, no. 25, pp. 14863–8, 1998.

[73] X. Zhou, M. Kao, and W. Wong, "Transitive functional annotation by shortest-path analysis of gene expression data," *Proceedings of the National Academy of Sciences of the USA*, vol. 99, no. 20, pp. 12783–8, 2002.

[74] S. Kleinberg and B. Mishra, "The Temporal Logic of Causal Structures," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, (Montreal, Quebec), *to appear*, June 2009.

[75] Z. Bozdech, M. Llins, B. Pulliam, E. Wong, J. Zhu, and J. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum," *PLoS Biology*, vol. 1, no. 1, 2003.

[76] D. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. Hesselberth, L. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, and et al, "A protein interaction network of the malaria parasite Plasmodium falciparum," *Nature*, vol. 438, no. 7064, pp. 103–107, 2005.

[77] A. Bahl, B. Brunk, R. L. Coppel, J. Crabtree, S. J. Diskin, M. J. Fraunholz, G. R. Grant, D. Gupta, R. L. Huestis, J. C. Kissinger, P. Labo, L. Li, S. K. McWeeney, A. J. Milgram, D. S. Roos, J. Schug, and C. J. Stoeckert, "Plasmodb: the plasmodium genome resource: An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished)," *Nucleic Acids Research*, vol. 30, no. 1, pp. 87–90, 2002.

[78] S. Kleinberg, K. Casey, and B. Mishra, "Systems biology via redescription and ontologies (i): finding phase changes with applications to malaria temporal data," *Systems and Synthetic Biology*, vol. 1, no. 4, pp. 197–205, 2007.

[79] D. Altman and J. Bland, "Diagnostic tests. 1: Sensitivity and specificity," *British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.

[80] A. Maier, B. Cooke, A. Cowman, and L. Tilley, "Malaria parasite proteins that remodel the host erythrocyte," *Nature Reviews Microbiology*, vol. 7, pp. 341–354, May 2009.

[81] S. Reiff and B. Striepen, "Malaria: The gatekeeper revealed," *Nature*, vol. 459, pp. 918–919, June 2009.

[82] M. Marti, R. Good, M. Rug, E. Knuepfer, and A. Cowman, "Targeting malaria virulence and remodeling proteins to the host erythrocyte," *Science*, vol. 306, no. 5703, pp. 1930–1933, 2004.

[83] J. MacKenzie, N. Gomez, S. Bhattacharjee, S. Mann, and K. Haldar, "Functions in export during blood stage infection of the rodent malarial parasite plasmodium berghei," *PLoS ONE*, vol. 3, no. 6, p. e2405, 2008.