Topology and Function in Protein Interaction Networks

Elena Nabieva

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
By the Department of
Computer Science

September 2007

© Copyright by Elena Nabieva, 2007. All rights reserved.

Abstract

A key problem in biology is understanding the work of proteins. While protein sequences have been mostly determined for many organisms, the functions of these proteins and how they work together to accomplish them are much less understood. An important source of information for addressing these questions is protein interaction data. Protein interactions, which, taken together, can be represented as networks or graphs, have been determined on a large scale for several organisms. In this work, we study the relationship between protein function and interaction network topology, focusing on protein-protein physical interaction networks. We address both the task of assigning function to individual proteins and the more global question of the organizational principles underlying these networks.

In the first part of this thesis, we explore the use of physical interaction networks for predicting protein function. We begin by discussing which topological properties of interaction networks should be taken into account by network-based function prediction algorithms, using as illustrations some earlier approaches to this problem. Then, using these desiderata as guidelines, we introduce an original network-flow based algorithm for predicting protein function. This algorithm, FunctionalFlow, takes advantage of both network topology and some measure of locality, and, as a result, has improved performance over previous methods. Finally, we show that performance can be improved substantially as we consider multiple data sources and introduce edge weights to reflect data reliability.

In the second part of this thesis, we take a different view at the topology-function relationship and use known information about protein molecular function to attempt to uncover the organizational principles of physical interaction networks. We examine the networks from the perspective of "pathway schemas," or recurring patterns of interaction among different types of proteins. Proteins in these schemas tend

to act as functional units within diverse biological processes. We discuss computational methods for automatically uncovering statistically overrepresented schemas in protein-protein interaction maps and touch upon the comparative-interactomics aspects of this problem. Coming back to the task of improving our understanding of protein function, we conclude by demonstrating how overrepresented schemas can suggest new insights into the biological function of proteins.

Acknowledgments

Chapter 2 is joint work with Kam Jim, Amit Agarwal, Bernard Chazelle, and Mona Singh. It appeared in the proceedings of ISMB 2005 [68].

Chapter 3 is joint work with Eric Banks, Bernard Chazelle, and Mona Singh. I would also like to thank Moses Charikar and David Blei for helpful discussions about this project.

I would like to thank the members of my thesis committee, especially the readers Bernard Chazelle and Olga Troyanskaya, as well as the non-readers Ned Wingreen, Szymon Rusinkiewicz, and Tom Funkhouser.

I was supported by Princeton University and the following grants: NSF Pecase MCB-0093399, DARPA MDA972-00-1–31, NIH P01-CA-041086, NIH R01-GM-076275, NSF CCF-0542187.

I would like to thank the members of the Singh group, past and present, for discussions about this work as it was being developed: Elena Zaslavsky, Carl Kingsford, Jessica Fong, Robert Osada, Nick Jacobson, Tony Capra, Anton Persikov, Jimin Song, Zia Khan, and Alex Ochoa. Eric Banks gets my special appreciation for our collaboration on the schema finding project.

Mona Singh has been a wonderful advisor and mentor, and I am very fortunate to have been her student.

Finally, I would like to thank my family: my parents Tatiana and Rashit, my grandparents Galena and Alexander, and my sister Svetlana, for all their support and encouragement.

Contents

Chapter 1

Introduction

This thesis is largely concerned with proteins, the workhorse molecules of living organisms. Genome sequencing efforts have made available protein (as well as non-coding) sequences for a large number of organisms, yet in many cases not much more is known. Broadly speaking, in this thesis we address two questions concerning proteins that are fundamental to our understanding of biology: what do proteins do and how do they work together to do it. We attempt to answer these questions via analysis of high-throughput experimental data sets, and in particular, large-scale protein interaction data.

High-throughput biology provides a radically new look at proteins within their cellular context. Using a modern analogy, the high-throughput view can be likened to taking satellite images, whereas traditional experiments on individual proteins play the role of street-level photographs of individual houses or city blocks. The genome-wide view can bring to light a protein that has not yet come into the sight of a traditional experiment, like the obscure house behind a tall fence on the outskirts of a city which the photographer never got to. High-throughput data makes it possible to ask questions about such uncharacterized proteins—which may have been difficult

to trace experimentally or which simply belong to pathways or organisms that have not been sufficiently studied. Additionally, and more importantly, analysis of high-throughput data can begin to reveal answers to these questions and can provide a glimpse at what these proteins do and how they do it. In addition, the relatively unbiased nature of proteome-wide experiments and the sheer scale of the resulting data make it possible to ask questions about the organizational principles that govern life; these types of questions would be difficult or impossible to ask with traditional pathway-oriented data. As a result, it is now feasible to try to uncover the modularity of cellular organization, the interplay between different pathways, or other organizing principles of protein networks—just as a satellite image of a city makes it possible to see its layout and organization.[1]

At the same time, the advent of high-throughput technology has changed the relationship between data and knowledge in biology. Whereas traditional experiments usually concern particular pathways or molecules in an organism and are intended to answer a certain question or evaluate a particular hypothesis, high-throughput biology often produces data which are yet to be placed within biological context. The large scale of high-throughput data requires considerable computational efforts to elucidate it. In this work, we focus on one family of genome-wide data sets: protein interactions.

1.1 Introduction to protein interaction networks

"Protein interaction" is a fairly broad term that is used to refer to a wide range of relationships between proteins. Protein interactions can describe concrete relation-

[1]Perhaps a comparison to satellite images is too optimistic at this point. As we discuss later, due to experimental noise, it is more accurate to draw the analogy with first-generation aerial photographs, shot in poor lighting with a weak lens.

ships, such as physical interactions between proteins or between proteins and other molecules, or regulatory relationships, where one protein regulates the transcription of the gene coding for the other protein, or phosphorylation relationships, where one protein phosphorylates the other. Other types of interactions describe more abstract relationships between proteins, such as "genetic" interactions which describe the relationship between two genes with respect to the well-being of the organism (e.g., synthetic lethality), or coexpression interactions, in which two proteins are said to be interacting if they have similar expression patterns (under some set of conditions). Finally, protein interactions may represent functional associations; here two proteins are said to interact if they take part in the same biological process in the cell.

Protein interactions can be naturally represented as a graph, or a *protein interaction network*, in which vertices correspond to proteins and edges connect interacting proteins. If the underlying interactions are intrepreted as being symmetric, as is the case of protein physical interactions or coexpression interactions, the graph is undirected. If there is a clear directionality to the interactions, as in the case of regulatory or phosphorylation interactions, the graph is directed. Hybrid networks can be built by combining different interaction networks. Together with metabolic networks, which include metabolites in addition to proteins, the combination of various protein networks, many of which can be determined on a genomic scale, is a powerful source of information about the workings of cells. Analysis of these networks can provide hints to the organization of the cell, can help elucidate protein function and can provide an understanding of the interplay between proteins. After all, proteins do not perform their cellular roles in isolation, but do so in collaboration with other proteins. Here, we focus on physical protein-protein interaction data, although many of our techniques can be readily extended to other types of binary experimental interactions.

1.2 Our focus: Topology and function in physical interaction networks

In this thesis, we analyze aspects of the relationship between protein interactions and protein function. Before we proceed, therefore, we need to discuss what we mean by protein function.

The word "function" in relation to proteins has a fairly broad meaning. Usually, one considers two views of protein function. Following the terminology of the Gene Ontology [2], we will call them molecular function, which describes the biochemical activity of a protein, and biological process, which specifies a more abstract notion of the role the protein plays in the cell or the pathway in which it participates. These views of protein function are largely orthogonal: proteins with the same molecular function can take part in different pathways, and a pathway is built of proteins of various molecular functions. From the perspective of function prediction, molecular functions, which correspond to the intrinsic features of the protein, are often predicted based on sequence or structural similarity to proteins of known function, whereas biological processes, being fundamentally collaborative, are often predicted based on a protein's functional interaction partners (e.g., the protein it interacts with physically). In this thesis, we will look at both views of protein function.

Not surprisingly, these roughly orthogonal aspects of protein function obey different general principles with respect to protein interaction: interacting proteins tend to participate in the same biological process, whereas the life of the cell relies on the interaction of proteins that often have different molecular function. In the first part of this thesis, we focus on the biological processes in which proteins participate, and develop algorithms for analyzing protein physical interaction graphs in order to predict protein function. We exploit and extend the principle of *guilt by association*,

which is the basis of most of the work of this kind. This portion of our work helps elucidate how network topology should be used in making predictions about the biological processes of proteins. In the second part of the thesis, we look at the molecular features of proteins, and focus specifically on the interaction between proteins that may have different molecular features. We introduce algorithms for automatically inferring what types of proteins or groups of proteins tend to interact with each other, and in what topology, in order to accomplish diverse biological processes. Whereas the first part of the thesis uses protein interaction graphs to predict biological processes of individual proteins, the second part of the thesis attempts to uncover how the cell is organized with respect to the molecular function of proteins. The first part of the thesis uses network topology to predict biological process function and the second part of the thesis uncovers over-represented topologies between proteins of particular molecular function.

1.3 Determination of protein interactions

Protein interactions can be roughly divided into those that are determined experimentally and those that are created computationally. In the former case, the results of an experiment can be quite naturally interpreted as interactions between pairs of proteins. In this case, the bulk of the work in determining the protein interactions is done by the experimentalist. Physical [41], [107], genetic (e.g., synthetic lethality [103]), regulatory [79], and phosphorylation [43] interactions are among members of this category.

Since this thesis is largely focused on protein-protein physical interactions, we briefly describe the high-throughput technologies used to determine them. Experimental techniques for determining physical protein-protein interactions have been

dominated by the yeast two-hybrid method and complex pull-down methods. In the yeast two-hybrid, one protein is fused to an activator domain of a transcription factor for a yeast reporter gene, and the other protein is fused to its DNA-binding domain; the expression of the reporter is evidence of interaction [9]. In the pull-down methods, a bait protein is used to identify other proteins (preys) that are co-complexed with it [36, 80]. A variant of the two-hybrid method which uses a split ubiquitin insted of the transcription factor [47] has been used more recently on a large scale to probe membrane protein interactions missed by the two-hybrid method [64]. As a result, interactions covering a large portion of the interactomes of several organisms, including yeast, fruit fly, worm C. elegans, and human have been compiled [97].

On the other hand, computational interactions, as the name suggests, are determined by a computational biologist, perhaps based on other kinds of experimental data. Thus, relationship of coexpression is established on the basis of numerical data about levels of gene expression under various conditions and/or over a period of time. The usual approach to converting this numerical data into binary interactions is to measure the similarity between the expression profiles of all pairs of proteins and to consider as coexpressed (i.e., interacting) those pairs for which the coexpression measure is sufficiently high. It is up to the computational biologist to decide on the measure of similarity (e.g., Pearson correlation coefficient [22], mutual information [3], or others), on the ways of combining the expression profiles over different experiments or time series (several of which are discussed in [38]), and on other aspects of the task, such as the desired semantics of interaction (discussed below). Other types of computational interactions have been determined via coevolution [73], conservation of gene order [13], gene fusion events [23], and the tendency of proteins to co-occur in scientific literature [45]. An area of active and fruitful research is establishing functional interactions by computational integration of data of different types, which usually

include both experimental and basic computational interactions [54, 105, 114].

Of course, the distinction between experimental and computational interactions is somewhat blurred, since even experimental interactions require some computational processing—in the very least, a decision on where to draw the cutoff in the strength of the experimental evidence for interaction, such as the binding affinity of a transcription factor for its binding site. In addition, computational effort is used to consruct experimental interaction networks in a way that deals with experimental noise, which will be discussed shortly.

Finally, there are efforts to use computational techniques to *predict* experimental interactions. These include, for example predicting physical protein-protein interactions using sequence features including domains [31], sometimes focusing on particular types of interactions (e.g., coiled-coil interactions [25]), or other types of evidence such as gene fusion in other organisms or the existence of orthologs interacting in another genome (see [91] for a review). These studies are meant to augment the corresponding experimental efforts; ideally, computationally predicted interactions can be treated in the same way as experimental ones.

1.4 Challenges of interaction network analysis

There are certain challenges that one faces when studying protein interaction networks; most of them have to do with features or weaknesses of experimental techniques used for determining the interactions. These include dealing with experimental noise and incompleteness, understanding the dynamics of the interactions, and interpreting the "meaning" of a type of interaction data (i.e., its semantics). We focus our discussion primarily on physical protein-protein interaction data in this section.

1.4.1 Quality of experimental data

High-throughput experiments tend to be both noisy and incomplete, especially since many of them involve new (at the time of the experiment) technology. For example, it has been estimated that at least half the interactions reported by the early high-throughput two-hybrid screens in yeast are spurious [110].

One way to deal with noise in physical interaction data is to assign edge weights that reflect the reliability of the data underlying the interaction. Usually, a different type of data is used to measure the reliability of the interaction; for example, expression has been used to evaluate physical protein interaction data [16], and, in our work on function prediction in Chapter 2, we use biological process information for the same purpose. In addition, it is common practice in physical interaction network analysis, which we follow, to exclude proteins that have a large number of interaction partners, as they may be "promiscuously" interacting in the experiment.

Besides being noisy, physical interaction maps are incomplete. Determination of interaction partners for all proteins in an organism have not always been attempted. Moreover, the experiments are also believed to have high false-negative rates, even for interactions that are detectable by the experimental technique. For example, the overlap between two early high-throughput screens using the yeast two-hybrid technology [41, 107] is only 16.8%-20.4% of the two experiments' core data [41], which should be attributed both to false positives and false negatives in the results of these experiments. Moreover, certain types of interactions may be missed because of the nature of the experimental techniques. Interactions that are conditioned on post-translational modification, for example, are likely to be severely underestimated. Similarly, certain types of proteins, such as those that are integral to the membrane pose particular challenge to interaction assays. Of course, the development of new experimental techniques, the improvement of existing ones, and the execution of new

experiments should lead to improvement both in the coverage of interaction data and in its accuracy (if one, for example, chooses to ignore the earlier less reliable data sets). For example, a large-scale screen specifically designed to capture the interactions between membrane proteins has been performed in yeast [64].

While noise in physical interaction data is certainly an issue, the methods we develop in this thesis are performed on the "global" scale. Such analysis alleviates some of the data quality problems, since it does not rely on any single interaction. For example, when global interaction data is used, the flow of information (e.g., "functional flow" of Chapter 2) may circumvent the missing edge(s). Similarly, we judge the significance of recurring interaction patterns (in Chapter 3) via comparison to randomized networks; this should reduce the effect of false interactions, since they would be indistinguishable from "random."

1.4.2 Dynamics of interaction networks

One salient feature (that may also be known as a bug) of many protein interaction networks in their present state is that they give a static view of the interactome—in other words, our "aerial photograph" is taken with a very long exposure. Experiments for determining protein interactions may take place outside of the cell which is being studied and under conditions that may not reflect the conditions in the cell when the interaction takes place. Yeast two-hybrid experiments for determining physical interactions are an important example. The outcome of such experiments is thus information that an interaction may take place, but not the conditions under which it takes place—not to mention, in the case of multicellular organisms, the type of cell in which it would occur. Conversely, many interactions may be missed because the experimental conditions under which they are assayed are different from what is required for the proteins to interact. In contrast, gene expression is usually

studied in a condition-specific manner; however, it is a task for the computational biologist constructing coexpression interactions to make sure that the condition of the experiment is reflected in the coexpression interactions.

In response to this weakness in physical interaction datasets, there have been several recent computational studies that endeavored to incorporate information about interaction dynamics into the study of interaction networks. Among them are studies that looked at physical interactions in condition-dependent manner, using gene expression [33] to get a glimpse at interaction dynamics or GO biological process annotations [77] for "biological context" in which the interactions may take place. Such analysis has revealed, for example, that proteins that appear as hubs (highly connected proteins) in the static interaction network, can be split into those that interact with all their neighbors simultaneously and those that bind their partners at different times or in different locations [33]. Other studies have looked at the dynamics of complex formation [14] and of regulatory networks [58], once again, using expression data for information of when proteins may be active.

So far, the analysis of network dynamics has largely focused on topological properties of networks under different conditions. As this approach to network analysis matures, dynamical views should become prominent in other types study of interaction networks.

In this thesis, we largely view protein interaction networks as a static picture. However, some of our work in network analysis reveals patterns that are likely to be indicative of protein pathway dynamics (Chapter 3). Deciphering such patterns is an intriguing topic for follow-up research.

1.4.3 Semantics

Another issue with interaction networks that needs awareness is the semantic interpretation of interactions. The meaning of interactions is perhaps of particular concern for researchers who construct computational interactions. Since these researchers have a degree of control over defining what constitutes an interaction, they may try to define interactions in a way that suits their purpose. For example, much research on protein interaction network analysis concerns using interaction information beyond a protein's immediate interaction partners to predict protein function (in the first part of this thesis, we address this question in the context of physical interaction networks). However, an alternative approach to this problem might be to design a network of computational interactions in which relevant functional information is nearly guaranteed to be contained in the interaction partners of a protein; in such networks, long-range relationships would be "short-circuited." Alternatively, one may be faced with the opposite task of constructing a sparser computational interaction network which contains only direct interactions. An example of such task is constructing interaction networks based on coexpression data. Similarity of expression profiles of two proteins, A and C, may indicate both direct relationship between them or an indirect relationship that is mediated by a third protein B (or group of proteins). To address this effect, [3] have used a heuristic approach to construct networks of expression-based interactions that are likely to be direct.

Although physical interaction data is more readily interpretable, the semantics of the physical interactions is not always obvious, as different experimental techniques give different views of interaction. Two-hybrid techniques, for example, detect pairs of proteins that are likely to be directly interacting, whereas pull-down experiments reveal instead complex co-membership with the bait protein. The actual topology of interactions between the bait and the prey proteins and especially among the prey

proteins is not immediate from the pull-down data. Usually, interactions found in pull-down experiments are represented according to either the "spoke" model in which a bait protein are connected to every prey protein or the "matrix" model in which interactions are assumed between all pairs of proteins that participate in a pull-down. There have been computational studies meant to shed light on the interpretation of pull-down, such as a method to to determine complex co-membership based on two-hybrid and pull-down data [86]. Even if an interaction is believed to be direct, it can have different interpretations: some types of physical interactions play structural roles, some form complexes that perform various functions, and some are involved in transmitting information through the cell. Furthermore, interactions may be more-or-less permanent (e.g., between members of stable complexes), or transient, as is often the case between signaling molecules.

Understanding of the semantics of protein interactions is important for design and application of computational methods for network analysis. For example, the network-based function prediction algorithm we develop in Chapter 2, FunctionalFlow, was designed for physical protein-protein interaction networks, with the awareness of two properties of these networks: first, that each interaction is informative by itself and cannot be derived from other interactions, and that the network may contain highly connected subgraphs which correspond to protein complexes. The assumption that each interaction is a non-reducible piece of evidence for association between proteins, leads to insights about the features of network structure that should be taken into account in design of network-based function-prediction algorithms (Section 2.2). At the same time, FunctionalFlow can be readily applied to other networks of undirected binary experimental interactions; we demonstrate that including information about genetic interactions improves predictive performance. For the same reason, FunctionalFlow would not perform well on computational net-

works with "transitive closure" of interactions, such as a coexpression network in which proteins need only to have sufficiently similar expression profiles to be considered interacting. However, if the same network is filtered to focus on direct interctions using the approach of [3], then the inclusion of the filtered network improves the predictive performance.

FunctionalFlow's accomodation of protein complexes in the networks permits us not to dwell on the interpretation of pull-down data; however, we give more attention to this problem in the research presented in in Chapter 3, where we look for overrepresented patterns of interaction. Indirect interactions that may arise from complex comembership would only obscure our findings; therefore, we specifically filtered the interaction data to include only interactions that are likely to be direct.

1.5 Our contributions

Protein interaction networks offer hope of understanding the workings of a living organism, the organization of pathways and the interplay between them. In this thesis, we study "what do proteins do and how do they interact to do it". More specifically, we study the interplay between protein function and topology, first focusing on the use of network topology for protein biological process prediction, and then looking at the organizational principles of interaction networks while taking into the account biological features of the proteins.

We begin in Chapter 2 by looking at the problem of biological process prediction based on physical interaction data. Since the function of many proteins is unknown or known poorly, understanding protein function is an important challenge of modern biology. In this task, some of the proteins have known functional annotations, whereas others do not and need to be assigned biological process information. Many

methods have been devised to address this problem, and the majority of them are based on the principle of guilt by association, which predicts that interacting proteins tend to participate in the same process. The simplest approach to this problem involves assigning to a protein the process in which the majority of its interaction partners participate [87]. However, it may be desirable to use more global information about the interaction network, such as larger neighborhoods around the protein in question, either because the immediate neighbors are poorly or not at all annotated (this is especially relevant for organisms that have not yet been sufficiently studied) and thus do not provide enough information, or simply because one believes that global information is valuable for function prediction. In Chapter 2 of this thesis, we examine some physical interaction network-based methods for function prediction. In particular, we look at the importance of different features of the interaction networks, such as topology or distance in the graph, using these methods for illustration. We then propose a novel function prediction algorithm FunctionalFlow, based on graph flow, which incorporates these features and as a result achieves superior performance. Interestingly, we find that if a protein has sufficiently many annotated interaction partners, it is best to use just the local neighborhood for function prediction; however, for a large number of proteins for which the immediate neighborhood "signal" is not as strong—which is the case for many newly sequenced organisms with few functional annotations—FunctionalFlow brings improved functional prediction. We conclude this chapter by addressing the issue of noise in physical interaction networks and proposing a simple edge-weighting scheme that reflects the reliability of underlying interactions, and showing that this scheme improves the performance of the methods considered, and finally showing how we can gain greater improvement in performance by adding synthetic lethality interactions.

Then, in Chapter 3, we turn to elucidating the organizational principles of protein

interaction networks. This has been an area of fruitful research, which has previously focused on topological properties of the networks, such as their degree distribution [46], on finding dense subgraphs which correspond to protein complexes [94], on overrepresented graph substructures [65], and so on. This type of analysis has largely ignored the features of proteins and treated the graph as unlabeled. We, on the other hand, turn our attention to patterns of interaction between proteins having various molecular features. We propose a bottom-up view of protein networks that focuses on the "building blocks" of which they are constructed. Towards such analysis, we introduce *pathway schemas* as a means of describing recurring patterns of interactions that tend to act as functional units. Pathway schemas are defined by descriptions of proteins (i.e., protein features) and the interactions among them (i.e., a specific topology). Simple pathway schemas associated with signaling can consist, for example, of a kinase interacting with another kinase, or a GTPase interacting with both a GTPase activating protein and a GEF protein which reverts it. Building on work on finding overrepresented subgraphs or motifs, we use a collection of randomized graphs to find overrepresented pathway schemas. We present a statistical framework for uncovering schemas that are overrepresented in the protein interaction network compared to a collection of similar random networks and develop an algorithm for automatically uncovering such schemas. We present our results for four small topologies and allow proteins to be described via Pfam domains [4]. We uncover many pathway schemas that are over-represented in protein-protein interaction networks compared to randomized graphs having the same properties. They include both well-known interacting units as well as putative novel structures. In the end, we come back to the problem of biological process prediction, and show how the patterns we uncover can help predict the cellular role of uncharacterized proteins and protein families. Overall, our work suggests that pathway schemas are a powerful new paradigm for

modularizing cellular networks.

Thesis organization

In Chapter 2, we discuss using protein interaction data to predict biological processes of proteins. In Chapter 3, we look at the organizational principles of the interaction networks and introduce pathway schema analysis which incorporates interaction topology and protein features. In Chapter 4, we conclude and discuss future directions for research.

Chapter 2

Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps

2.1 Introduction

A major challenge in the post-genomic era is to determine protein function at the proteomic scale. Even the best-studied model organisms contain a large number of proteins whose functions are currently unknown. For example, about one-third of the proteins in the baker's yeast *Saccharomyces cerevisiae* remain uncharacterized. Traditionally, computational methods to assign protein function have relied largely on sequence homology. The recent emergence of high-throughput experimental datasets has led to a number of alternative, non-homology based methods for functional annotation including biological process annotation. These methods have generally ex-

ploited the concept of guilt by association, where proteins are functionally linked through either experimental or computational means.

Large-scale experiments have linked proteins that physically interact [29, 30, 36, 41, 57, 78, 107], that are synthetic lethals [103, 104] and that are coexpressed [21] or coregulated [34, 55]. In addition, computational techniques linking pairs of proteins include phylogenetic profiles [28, 73], gene clusters [70], conserved gene neighbors [13] and gene fusion analysis [23, 59]. Perhaps not surprisingly, integrating the information from several sources provides the best method for linking proteins functionally [44, 54, 60, 105, 109].

It has been postulated that analysis of the resulting protein networks should help the understanding of protein function (for a recent review, see [89]). In this chapter, we focus on the problem of predicting protein function by analyzing proteins as components within protein interaction networks.

Physical interaction network-based protein function prediction has been an area of active research. Several groups have attempted to partition interaction networks into functional modules that correspond to sets of proteins that are part of the same cellular function or take part in the same protein complex. Then, an uncharacterized protein can be classified according to the functional annotation of the characterized members of the same cluster (e.g., taking the most frequent annotation of the cluster members). Other groups have used machine learning techniques to address the problem. Many of these approaches can be extended to the data integration framework, which usually results in improved performance. We review some of these methods in section 2.1.1.

The research described here is most closely related to the attempts to classify proteins according to functional annotations of their network neighbors. Schwikowski *et al.* [87] use physical interaction data for baker's yeast, and predict the biological

process for each protein by considering its neighboring interactions and taking the three most frequent annotations. While such a simple majority vote approach, which we refer to as Majority, has clear predictive value, it takes only limited advantage of the underlying graph structure of the network. For example, in the interaction network given in Figure 2.1, Majority would assign functions to proteins d and f, but not to protein e, even though our intuition might indicate that protein e has the same function as proteins d and f; there are several examples in the yeast proteome similar to this one [87]. Naturally, one wishes to generalize this principle to consider functional linkages beyond the immediate neighbors in the interaction graph, both to provide a systematic framework for analyzing the entirety of physical interaction data for a given proteome and to make predictions for proteins with no annotated interaction partners.

Hishigaki *et al.* [35] extend Majority by predicting a protein's function by looking at all proteins within a particular radius and finding over-represented functional annotations. However, this approach, which we refer to as Neighborhood, does not consider any aspect of network topology within the local neighborhood. For example, Figure 2.2 shows two interaction networks that are treated equivalently when considering a radius of 2 and annotating protein a; however, in the first case, there is a single link that connects protein a to the annotated proteins, and in the second case, there are several independent paths between a and the annotated proteins, and moreover, two of these proteins are directly adjacent to a.

Two papers [48, 108] subsequent to [35, 87] exploit the global topological structure of the interaction network by annotating proteins so as to minimize the number of times different annotations are associated with neighboring proteins. [48] additionally consider the case where edges in physical interaction networks are weighted using gene-expression data. We refer to this overall approach as GenMultiCut, as it is a

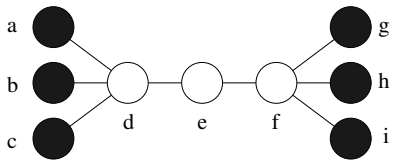

Figure 2.1: A protein interaction graph. Nodes represent proteins and edges represent interactions between proteins. For example, protein d interacts with proteins a, b, c and e. Proteins a, b, c, g, h and i (shown in black) are known to take part in the same biological process, and proteins d, e and f are unannotated.

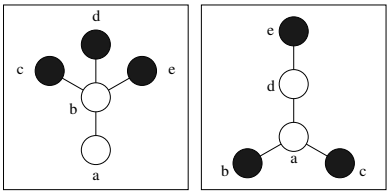

Figure 2.2: Two protein interaction graphs that are treated identically by Neighborhood with radius 2 when annotating protein a. Dark colored nodes correspond to proteins that are known to take part in the same process.

generalization of the well-studied multiway k-cut problem in computer science. While GenMultiCut takes into account more global properties of interaction maps, it does not reward local proximity in the graph. For example, if only two proteins have annotations in a particular network, all other proteins will be labeled by one of these annotations, regardless of the size of the network.

To overcome the weaknesses of previous methods, we introduce an algorithm, FunctionalFlow, for annotating protein function in interaction networks. FunctionalFlow uses the idea of network flow, which is dual to the notion of graph cut (e.g., see [11]). Each protein of known functional annotation is treated as a 'source' of 'functional flow' which is then propagated to unannotated nodes, using the edges in the interaction graph as a conduit. This propagation is governed by simple local rules. By considering a formulation based on flow, we can incorporate a distance effect. That is, the effect of each annotated protein on any other protein decreases with increasing distance between them. In addition, network connectivity is exploited, as each edge has a 'capacity' and multiple paths between two proteins result possibly in

more flow between them. After simulating the spread of this functional flow for a fixed number of time steps (so that flow from a source is restricted to a local neighborhood around it) we obtain the 'functional score' for each protein. This score corresponds to the amount of flow for that function the protein has received over the course of simulation. In contrast to Majority, FunctionalFlow considers functional annotations from proteins that are not immediate neighbors, and thus can annotate proteins that have no neighbors with known annotations. In contrast to Neighborhood, FunctionalFlow considers the underlying topology of the graph, and the multiple edge-disjoint interaction paths between two proteins give additional evidence for common function. Finally, in contrast to GenMultiCut, FunctionalFlow takes into account network locality.

We compare the performance of FunctionalFlow with Majority, Neighborhood and GenMultiCut. In the process, we reformulate the computational problem given by the objective function of [108] and [48] as an integer linear program (ILP), and as opposed to the previous two studies, we find optimal (not heuristic) solutions to the problem using ILP. Since we find optimal solutions, we directly test the utility of the GenMultiCut objective function. In addition, we show how to obtain multiple optimal solutions using ILP, and show that this is one way to incorporate the idea of distance implicitly within the GenMultiCut framework. In cross-validation testing on the yeast physical interaction network, we show that FunctionalFlow outperforms Neighborhood and GenMultiCut, and has better performance than Majority in predicting the function of proteins with few (or no) annotated protein neighbors. We estimated that at the time of writing this paper, there were, in the yeast proteome, ~ 1200 such unannotated proteins where FunctionalFlow would make improved predictions over Majority. This number (at the time of writing this paper) was 2400 for fruit fly, and the fraction of such proteins should be much higher for less characterized

proteomes. Finally, we propose a simple weighting scheme that captures the variation in reliability of the experimental data that form the basis of the interaction network, and show that this scheme results in improved performance for all methods.

Overall, we demonstrate that network analysis algorithms such as FunctionalFlow provide an effective new line of attack in determining protein function. Moreover, we show empirically that network analysis algorithms for function prediction obtain the best performance when incorporating overall network topology, network distance and edges weighted by a reliability parameter estimated from multiple data sources. The FunctionalFlow method we introduce incorporates these features and outperforms previously published methods. While all of our cross-validation testing has been on baker's yeast, FunctionalFlow is likely to be especially useful in characterizing less-studied proteomes.

2.1.1 Further related work

The locality effect in FunctionalFlow is similar in some ways to the locally constrained diffusion kernel developed by [106]. However, the flow in the FunctionalFlow algorithm is limited by capacities on edges, and in the context of our method, this prevents all the proteins that have the same annotation but have largely overlapping paths to protein a from exerting too much influence on a. Moreover, [106] use the diffusion kernel with support vector machines, whereas FunctionalFlow is not a learning method and does not require any training data to be used.

Alternate machine learning approaches for predicting protein function using physical interaction data include those based on Markov Random Fields (MRFs) [18, 56]. In a way, MRFs are a probabilistic extension to the multicut formulation. MRFs are probabilistic graphical models in which the probability of labeling the vertices is computed based on *local potentials* which include the vertex potentials (reflecting,

e.g., the *a priori* probability that the vertex is assigned a given label) and a clique potential which reflects the tendency of members of a clique of size up to n to have a particular combination of labels. (MRFs are so named because they must obey the Markov property which states that the labeling of a vertex is independent of the labelings of its non-neighbors given the labeling on the neighbors.) In the application to the function prediction problem, the clique potential has been computed just over the edges; thus, an MRF with a uniform distribution for vertex potentials and clique potentials that are 0 for assigning different functions to the endpoints of an edge and 1 for assigning the same function is just the multicut problem [17].

Another family of methods endeavor to assign proteins to clusters and then transfer the dominant (e.g., most common) annotation of the cluster members to the uncharacterized members. Proteins in experimentally and computationally determined interaction graphs have been grouped together based on shared interactions [8, 54, 85, 99, 111], or the similarity between shortest path vectors to all other proteins in the network [82]. Other methods use the interaction networks more directly and focus on identifying densely connected components, or clusters or "communities" in the interaction network (e.g. [50, 71, 74, 94] and others). Not all methods in this group are designed specifically for protein interaction network and/or the protein function prediction problem, but the clusters thus discovered tend to be functionally coherent. Several of these clustering methods have the added effect of providing some information about network structure (e.g., identifying protein complexes).

Many of the network analysis methods discussed so far, which are based exclusively on physical interaction data, have counterparts in integrated networks. First, one can extend an algorithm developed for the physical network to a hybrid network that is constructed by combining several experimental networks. The challenge lies in the way the networks are combined. [17] extend the MRF method to a network

with physical and genetic interaction edges by defining the potential function on the hybrid network as a combination of the edge potential functions on the individual experimental networks. Similarly, in the end of this chapter, we apply the FunctionalFlow to a network constructed by combining the physical and genetic interaction networks into a single network with a simple joint weighting scheme. Kernel-based methods likewise have a data-integration application which combines kernel functions for different types of data [53].

A family of data-integration methds that are related to clustering look for modules consisting of interacting proteins that show similar behavior. Some look for groups of interacting proteins that are differentially expressed under certain conditions, such as [39] or [88]; the latter is another Markov Random Fields-based method, with vertex potentials reflecting differential expression. Other methods combine multiple kinds of genomic data in bicluster analysis [101].

An orthogonal area of research focuses on building truly integrated functional networks that combine diverse data [44, 54, 60, 67, 105, 109]. These integrated functional networks often do not need sophisticated graph algorithms to analyze, since the relevant functional information is contained in the immediate neighborhood of the target protein, and algorithms based on local properties suffice (although some of them [67] have built-in clustering analysis).

2.2 Materials and methods

Physical interaction network. We construct the protein-protein physical interaction network using the protein interaction data set compiled by GRID [7]. The resulting network is a simple undirected graph $G = (V, E)$, where there is a vertex or node $v \in V$ for each protein, and an edge between nodes u and v if the corresponding

proteins are known to interact physically (as determined by one or more experiments). Initially, we consider a graph with unit-weighted edges, and then consider weighting the edges by our "confidence" in the edge (see below). The weight of the edge between u and v is denoted by $w_{u,v}$. For all reported results, we consider only the proteins making up the largest connected component of the physical interaction map (4495 proteins and 12531 physical interaction links).

Functional annotations. Several controlled vocabulary systems exist for describing biological function, including MIPS (Munich Information Center for Protein Sequences) [63] and the Gene Ontology project (GO) [2]. We use the MIPS functional hierarchy, and consider the 72 MIPS biological processes that comprise the second level of hierarchy. Of the 4495 proteins in the largest connected component of the yeast physical interaction map, 2946 have MIPS biological process annotations. We also experimented with GO annotations; the overall conclusions made in this paper are not affected.

Weighting functional linkages. It is well known that the reliability of different data sources vary, even if they are based on the same underlying technology (e.g., see [16, 96, 110]). In the context of network-based algorithms, it is possible to weight edges so as to model the reliability of each interaction. For physical interactions, this reliability is in turn based on the experimental sources that contribute to our knowledge about the existence of the interaction. To determine these values, we separate all experimental sources of physical interaction data into several groups, placing each high-throughput data set into a separate group (five groups corresponding to each of [41, 42]; [27]; [107]; [29]; and [36]), and allocating one group for the family of all specific experiments. For each group of experiments, we compute what fraction of its interactions connect proteins with a known shared function. We assume that the reliabilities of different sources are independent, and thus conclude by estimating the

reliability of an interaction to be the noisy-or of the unreliability of the underlying data sources. That is, if r_i is the reliability of experimental group i, we compute the reliability of the edge by $1-\Pi_i(1-r_i)$, where the product is taken over all experiments i where this interaction is found. This treats each r_i as a probability and assumes independence; this approach is very similar to the one taken by [109].

We also consider augmenting the interaction network by considering genetic interactions from GRID [7]. Almost all of these interactions are synthetic lethals, and the weighting scheme can be immediately extended to this network by treating the new types of interactions as an additional experimental source. Thus, our weighting scheme gives us a way of integrating data of different types in addition to integrating different sources of data of one type.

Cross-validation testing and evaluation. We test performance using n-fold cross-validation. That is, the yeast proteome is divided into n groups, and each group in turn is separated from the original dataset and used for testing. The goal of each method is to predict the annotations of the proteins in the test set using the functional annotations of the remaining proteins. We performed experiments with 2-fold, 3-fold, 5-fold, and 10-fold cross-validation. All our cross-validation testing gives qualitatively similar results. We report our findings using 2-fold cross-validation, as baker's yeast is the most extensively studied organism, and 2-fold cross-validation better represents what one may expect to see in other organisms.

We evaluate the performance of the algorithms by considering, for each protein in the test set, whether the top scoring prediction above some threshold is a known functional annotation (true positive, TP) or not (false positive, FP). In the case of multiple predictions, the TP vs. FP status is tricky. For example, we may choose to count a prediction for a protein as a TP if at least one of the predictions made for it is correct, and as a FP otherwise. However, a method that predicts every protein to

participate in every function would only have TPs in this framework. Alternatively, we could count a protein as a TP if every prediction made for it is correct. This, however, would count as FPs those proteins that get many correct predictions and only one incorrect one. We therefore settle for a compromise approach, in which we count a protein's prediction as a TP if more than half of the predictions made for it are correct, and a FP otherwise. All results will be reported using this interpretation of TP and FP, and we use a variant of Receiver Operating Characteristic (ROC) curves, where we plot the number of TPs as a function of the number of FPs as we vary the scoring threshold.

2.2.1 Algorithms

Majority. As described in [87], for each protein we consider all neighboring proteins and sum up the number of times each annotation occurs. In the case of weighted interaction graphs, we simply extend the method by taking a weighted sum instead. For each protein, the score of a particular function is the corresponding sum.

Neighborhood. As described in [35], for each protein, we consider all other proteins within a radius r, and then for each function, we use a χ^2 test to determine if it is over-represented. For each protein, the score of a particular function is given by the value of the χ^2 test. Neighborhoods of radius one, two and three are considered. This method does not extend naturally to the case of weighted interaction graphs.

GenMultiCut. Two groups of researchers have suggested that functional annotations on interaction networks should be made so as to minimize the number of times different annotations are associated with neighboring proteins [48,108]. [108] use simulated annealing to attempt to minimize this objective function and aggregate results from multiple runs, whereas [48] use a deterministic approximation, and consider the

case where edges are weighted using gene expression information. As mentioned earlier, the formulation in these two studies is similar to the minimum multiway k-cut problem. In multiway k-cut, the task is to partition a graph in such a way that each of k terminal nodes belongs to a different subset of the partition and so that the (weighted) number of edges that are "cut" in the process is minimized. In the more general version of the multiway k-cut problem considered here, the goal is to assign a unique function to all the unannotated nodes so as to minimize the sum of the costs of the edges joining nodes with no function in common.

Our implementation of GenMultiCut

Though minimum multiway k-cut is NP-hard [12], we have found that the particular instances of minimum multiway cut arising here can in practice be solved exactly when stated as an integer linear program (ILP). We introduce a node variable $x_{u,a}$ for each protein u and function a. This variable will be set to 1 if protein u is predicted to have function a. If a protein u has known functional annotations, variable $x_{u,a}$ is fixed as 1 for its known annotations a and as 0 for all other annotations. We also introduce an edge variable $x_{u,v,a}$ for each function a and each pair of adjacent proteins u and v. This variable is set to 1 if both proteins u and v are annotated with function a. Minimizing the weighted number of neighboring proteins with different annotations is the same as maximizing the number with the same annotation, and so we have the

following ILP:

$$
\begin{array}{ll}
\text{maximize} & \sum_{(u,v)\in E, a\in FUNC} x_{u,v,a} w_{u,v} \\
\text{subject to} & \\
\quad \sum_a x_{u,a} = 1 & \text{if } annot(u) = \emptyset \\
\quad x_{u,a} = 1 & \text{if } a \in annot(u) \\
\quad x_{u,a} = 0 & \text{if } a \notin annot(u),\ annot(u) \neq \emptyset \\
\quad x_{u,v,a} \leq x_{u,a} & \text{for } (u,v) \in E \text{ and } a \in FUNC \\
\quad x_{u,v,a} \leq x_{v,a} & \text{for } (u,v) \in E \text{ and } a \in FUNC \\
\quad x_{u,v,a}, x_{u,a} \in \{0,1\} & \text{for all } u,\ v \text{ and } a.
\end{array}
$$

Here, $annot(u)$ is the set of known annotations for protein u, and $FUNC = \cup_u annot(u)$ is the set of all functional annotations. The first constraint specifies that exactly one functional annotation is made for any protein. The second and third constraints ensure that if protein u is annotated with function a, $x_{u,a}$ is set as a constant to 1, and if protein u is annotated but not with function a, $x_{u,a}$ is set as a constant to 0. The third and fourth constraints ensure that a particular function is picked for an edge only if it is also chosen for the corresponding proteins.

Considering multiple GenMultiCut optimal solutions

An important consideration in this framework is the existence of multiple optimal solutions. For example, the network in Figure 3 has seven minimum cuts of value one, and while the GenMultiCut criterion does not favor any one cut over the other, if we find all optimal cuts for this graph, we observe that x_2 is in fact annotated with F_1 more often than with F_2 in the assignments made by these cuts. Thus, a sense of distance to annotated nodes is in fact present in the set of all optimal solutions.

The simulated annealing method of [108] implicitly utilizes this information about

multiple solutions. [108] run simulated annealing 100 times, and predict for each protein the function that is assigned to it most often. If each run does indeed converge to an optimal solution, considering multiple runs amounts to sampling from the space of optimal solutions.

We deliberately attempt to sample from the space of optimal solutions. We explore two approaches for ensuring that multiple solutions are obtained by the solver. In the solution-exclusion approach, we add constraints to the ILP that require that each consecutive solution is different from any previous solution in the value it assigns to at least 5% of the node variables $x_{u,a}$. For the weighted yeast physical interaction graph, the first 50 solutions obtained with this restriction are all optimal. Note that in this approach, each successive solution takes longer to find. In the random weight perturbation approach, we introduce uniform self-weights $w_{u,a}$ for each protein u and function a. These self-weights are then perturbed by adding a very small offset to each, drawn at random from the uniform distribution on (-0.00001, 0.00001). We now modify the objective function in the ILP given above to maximize

$$\sum_{(u,v)\in E, a\in FUNC} x_{u,v,a} w_{u,v} + \sum_{u\in V, a\in FUNC} x_{u,a} w_{u,a}.$$

The perturbation in weights is too small to change the solution to the underlying problem, but it does cause the solver to choose a different optimal solution each time. Both methods perform very similarly in the accuracy of predictions made. For the reported results, we use the latter method for obtaining multiple solutions.

As in [108], we let the score for assigning a function to a protein be the number of times this function is assigned to the protein among the solutions that we found. We ran the ILP 50 times, and thus there are 51 possible scores (0-50) for any function for any protein. One solution to the ILP problem on the yeast interaction network

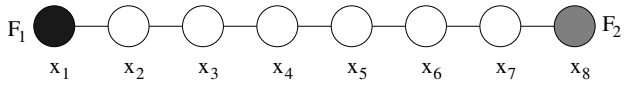

Figure 2.3: Proteins x_1 and x_8 are annotated with functions F_1 and F_2, respectively. There are seven ways to annotate proteins so that there is only one edge that connects proteins with different annotations. However, proteins further away from protein x_1 are less likely to have function F_1 than those closer to x_1. GenMultiCut does not take into account such distance effects.

with annotations for 50% of the proteins cleared can be obtained by AMPL [26] and CPLEX [40] in approximately five minutes when running on a public UNIX machine.

FunctionalFlow. The functional flow algorithm generalizes the principle of "guilt by association" to groups of proteins that may or may not interact with each other physically. We achieve this by treating each protein of known functional annotation as a "source" of "functional flow" for that function. After simulating the spread over time of this functional flow through the neighborhoods surrounding the sources, we obtain the "functional score" for each protein in the neighborhood; this score corresponds to the amount of "flow" for that function that the protein has received over the course of the simulation. The functional flow-based model allows us to incorporate a distance effect; that is, the effect of each annotated protein on any other protein depends on the distance separating these two proteins. Running this process for each biological function in turn, we obtain, for each protein, the score for each function (the score may be 0 if the "flow" for a function did not reach that protein during the simulation). Thereupon, for any protein, we take the functions for which the highest score was obtained as its predicted functions.

More specifically, for each function in turn, we simulate the spread of functional flow by an iterative algorithm using discrete time steps. We associate with each node (protein) a "reservoir" which represents the amount of flow that the node can pass on to its neighbors at the next iteration, and with each edge, a capacity constraint that dictates the amount of flow that can pass through the edge during one iteration.

The capacity of an edge is taken to be its weight. Each iteration of the algorithm updates the reservoirs using simple local rules: a node pushes the flow residing in its reservoir to its neighbors proportionally to the capacities of the respective edges and subject to further constraints that the amount of flow pushed through an edge during an iteration does not exceed the capacity of the edge, and that flow only spreads "downhill" (that is, from proteins with more filled reservoirs to nodes with less filled reservoirs). Finally, at each iteration, an "infinite" amount of flow is pumped into the source protein nodes; thus, the sources always have enough flow in their reservoir to fill the capacity of their outgoing edges.

The functional score is the amount of flow that has entered a protein's reservoir in the course of all iterations. Because flow is "pumped" into the sources at each step, the amount of flow a node receives from each source is greater for nodes that are closer to that source than for nodes that are further away from it. Thus, a source's immediate neighbor in the graph receives d iterations worth of flow from the source, while a node that is two links away from the source receives $d-1$ iterations worth of flow. Similarly, the number of iterations for which the algorithm is run determines the maximum shortest-path distance that can separate a recipient node from a source in order for the flow to propagate from the source to the recipient. For the protein interaction context, a relatively small number is sufficient. We choose $d = 6$, which is half the diameter of the yeast physical interaction network.

More formally, for each protein u in the interaction network, we define a variable $R_t^a(u)$ that corresponds to the amount in the reservoir for function a that node u has at time t. For each edge (u, v) in the interaction network, we define variables $g_t^a(u, v)$ and $g_t^a(v, u)$ that represent the flow of function a at time t from protein u to protein v, and from protein v to protein u. We will run the algorithm for d time steps or iterations. At time zero, we only have reservoirs of function a at annotated nodes:

$$R_0^a(u) = \begin{cases} \infty & \text{if } u \text{ is annotated with } a \\ 0 & \text{otherwise} \end{cases}$$

At each subsequent time step, we recompute the reservoir of each protein by considering the amount of flow that has entered the node and the amount that has left:

$$R_t^a(u) = R_{t-1}^a(u) + \sum_{v:(u,v)\in E} (g_t^a(v,u) - g_t^a(u,v))$$

Initially, at time 0, there is no flow on the edges, and $g_0^a(u,v) = 0$. At each subsequent time step, we have flow proceeding downhill and satisfying the capacity constraints:

$$g_t^a(u,v) = \begin{cases} 0, \text{ if } R_{t-1}^a(u) < R_{t-1}^a(v) \\ min\left(w_{u,v}, R_{t-1}^a(u)\frac{w_{u,v}}{\sum_{(u,y)\in E} w_{u,y}}\right), \text{ otherwise.} \end{cases}$$

Finally, the functional score for node u and function a over d iterations is calculated as the total amount of flow that has entered the node:

$$f_a(u) = \sum_{t=1}^{d} \sum_{v:(u,v)\in E} g_t^a(v,u)$$

2.3 Results and Discussion

2.3.1 Comparison of four basic methods on the unweighted physical interaction map

We compare the performance of Majority, Neighborhood, GenMultiCut and FunctionalFlow on the unweighted yeast physical interaction map, using a 2-fold cross-validation. Figure 2.4 plots as a function of FP the number of TPs each method predicts (i.e., these graphs are obtained by varying the scoring threshold for each

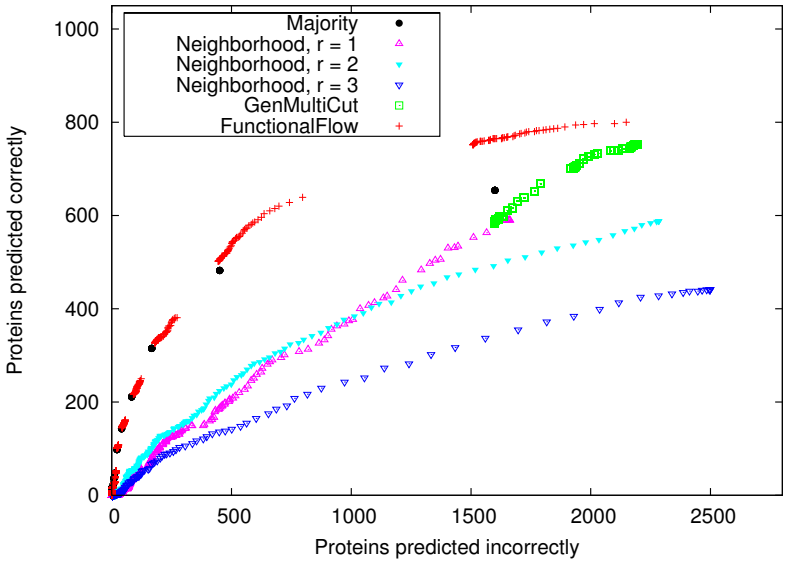

Figure 2.4: ROC analysis of Majority, Neighborhood, GenMultiCut and FunctionalFlow on the yeast unweighted physical interaction map.

of the methods). The FunctionalFlow algorithm identifies more TPs over the entire range of FPs than either GenMultiCut or Neighborhood using radius 1, 2 or 3. FunctionalFlow performs better than Majority when proteins are not directly interacting with at least three proteins of the same function; this is evident from Figure 2.4 since the score for Majority counts up the the most frequent neighboring annotation (e.g. the rightmost point for Majority corresponds to proteins whose highest functional scores are one). Thus, FunctionalFlow is the method of choice when considering proteins that do not interact with many annotated proteins. Even in well-characterized proteomes, such as baker's yeast, there are $\sim$ 1200 proteins that have fewer than three annotated neighbors.

The Neighborhood algorithm with radius 2 performs better than radius 1 only in the high-confidence region (i.e. corresponding to a low FP rate, given in the leftmost portion of the ROC curve). In addition, radius 1 and 2 have better overall performance than radius 3, demonstrating that Neighborhood's strategy of ignoring

topology is not optimal. Moreover, comparing Majority with Neighborhood using radius 1 demonstrates that the χ^2-test is not as effective in scoring as just summing up the number of times a particular annotation occurs in the neighboring proteins.

Since the score for GenMultiCut comes from multiple solutions to the underlying optimization problem, each point in Figure 2.4 for GenMultiCut corresponds to the proteins that are annotated with a particular function the same number of times. For example, the leftmost point for GenMultiCut corresponds to proteins where the top scoring functional prediction is found in each of the 50 solutions found. If we were to find just one optimal GenMultiCut solution, its performance in terms of TPs and FPs is comparable to the rightmost point for GenMultiCut (data not shown).[1] Thus, multiple solutions for GenMultiCut are necessary to identify its most confident predictions, and as pointed out earlier, these multiple solutions capture some notion of locality in the graph.

[108] report in their paper improved performance for GenMultiCut over Majority for proteins with degree > 1. Their measure of success is the fraction of times the top prediction for each protein is correct. Although they do not specify how they deal with multiple top predictions, we note that this measure corresponds to computing TPs and FPs for the rightmost points in Figure 2.4 for each of the methods. Assuming that the top predictions for each protein are treated separately, and that failure to make a prediction for a protein corresponds to an incorrect prediction, the top predictions for proteins with degree > 1 are correct 0.267 of the time for Majority. These values are 0.246 for Neighborhood with radius 1, 0.239 for Neighborhood with radius 2, 0.297 for GenMultiCut and 0.311 for FunctionalFlow. Although we believe ROC curve analysis gives a more complete picture of performance, FunctionalFlow

[1] It is not precisely the rightmost point in Figure 2.4 since this point aggregates solutions from multiple runs.

performs better than the other methods using this measure. Moreover, we tested the performance of all methods clearing a smaller fraction of the annotated proteins. In a 10-fold cross-validation (i.e. where only 10% of the yeast annotations are cleared), GenMultiCut has a slight advantage (25 proteins out of ~ 2500) over FunctionalFlow in the very low-confidence region; all other observations are qualitatively the same as for 2-fold cross-validation.

2.3.2 Reliability and data integration

To evaluate our approach for modeling physical interaction reliability as edge weights, we test the performance of FunctionalFlow using three ways of assigning physical interaction weights. First, we assign each edge a unit weight; this corresponds to the unweighted physical interaction map used above. Second, we assign each experimental source a reliability score of 0.5; this rewards interactions that are found by more than one experiment. Finally, we assign each experimental source the predictive value (estimated in cross-validation) as described in the Materials and Methods section; here, edges obtained from multiple, more reliable experiments are given higher weights. Figure 2.5 shows that rewarding multiple experimental evidence is beneficial, but that the main advantage comes from taking into account the actual reliability values for the different experiments.

Figure 2.6 shows how Majority, GenMultiCut and FunctionalFlow perform on the yeast physical interaction map, where edges are weighted by individual experimental reliability. The baseline performance of Majority on the unweighted physical interaction graph is also shown. There is substantial improvement in predictions using all three methods when incorporating edges weighted by reliability.

We further explored whether the network analysis algorithms would perform well when other types of experimental information are added. As a proof of principle,

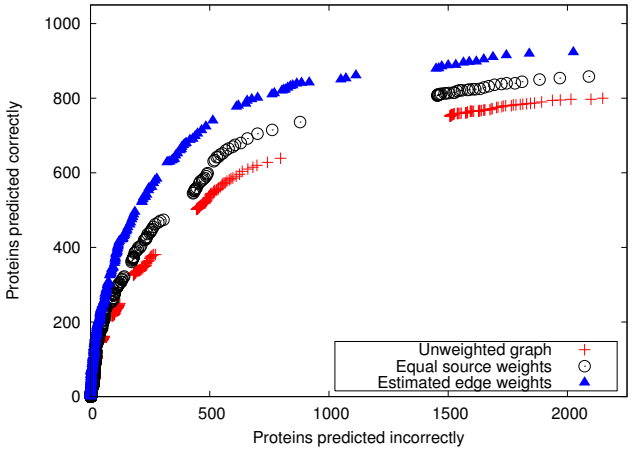

Figure 2.5: The FunctionalFlow algorithm on (1) the unweighted physical interaction map, (2) the physical interaction map with edges weighted using equal reliabilities for each experiment and (3) the physical interaction map with edges weighted by reliabilities estimated individually for each experiment.

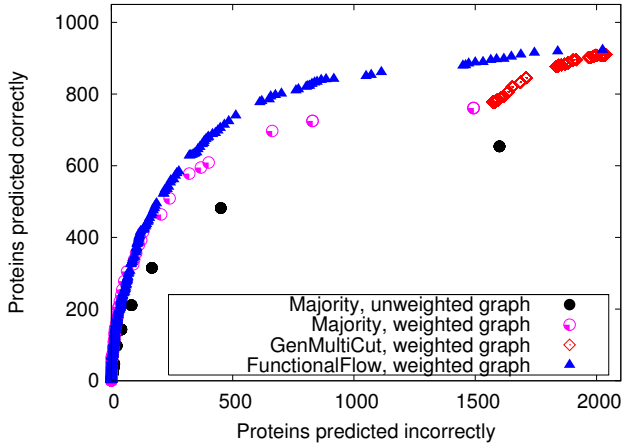

Figure 2.6: Performance of Majority, GenMultiCut and FunctionalFlow on the physical interaction map where experimental reliabilities are incorporated. The performance of Majority on the unweighted graph is also given as a reference.

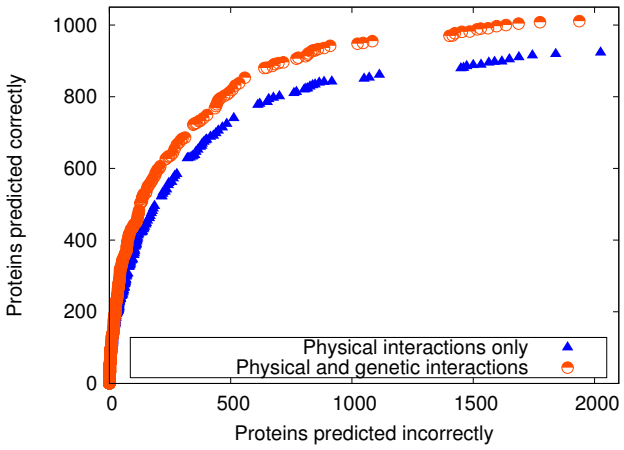

Figure 2.7: Comparison of functional predictions of FunctionalFlow when considering (1) the physical interaction map weighted by experimental source reliability and (2) the integrated physical and genetic interaction map.

we explore the effect of adding genetic linkages to the graph. Reliabilities for genetic interactions are estimated as described earlier, and incorporated into the edge weights. Figure 2.7 shows the performance of FunctionalFlow on the weighted physical interaction network and the weighted physical and genetic interaction network. As is evident, adding genetic interaction data significantly improves prediction quality. Majority and GenMultiCut show similar improvements (data not shown).

2.4 Conclusions

We have shown that our network analysis algorithm FunctionalFlow provides an effective means for predicting protein function from protein interaction maps. Our algorithm utilizes indirect network interactions, network topology, network distances, and edges weighted by reliability estimated from multiple data sources. On the other hand, we have also shown that the simplest methods, such as Majority, perform well if there are enough direct neighbors with known function. In the present work, simple independence assumptions are made for estimating the reliability of interactions.

While these work reasonably well, it may be even more beneficial to use Bayesian techniques to model dependence between data sets, and to perform more complete data integration (using, e.g., a portion of the Bayesian network of [105]). Finally, while we have applied our method to baker's yeast, FunctionalFlow is likely to be especially useful when analyzing largely uncharacterized proteomes where computational methods are used to infer protein interaction maps.

Chapter 3

Analyzing protein interaction networks via pathway schemas

3.1 Introduction

In this chapter, we address the problem of uncovering the organizational principles of cellular interaction networks. Broadly speaking, we aim to answer the following question: "Are there common means by which diverse biological processes are accomplished?" Our approach is to focus on known features of individual proteins—for example, their molecular functions as revealed via sequence motifs—within the larger context of physical interaction networks.

Since the first large-scale interactomes were determined, there has been considerable effort in analyzing the topological properties of protein interaction networks. Early studies of interaction network structure addressed properties like degree distribution, and uncovered their scale-free and small-world properties [46]. Follow-up work generalized this idea by considering the distribution of structures called graphlets [75], and other statistical and theoretical properties of the topological features of these

networks have been studied extensively studied (e.g., the under-reperesentation of interactions between highly-connected nodes [61] and characteristic graph measurements including diameter and clustering coefficient [117]). Such topological analysis can lead to biological insights; for example, it has been shown that highly-connected proteins more likely lead to lethality when deleted than lower-degree ones [46]. There has also been considerable research on identifying modules or dense structures in the interaction networks; we discussed some of this work in the previous chapter in the context of function prediction (see section 2.1).

Recently, topological analysis of networks has been used to uncover patterns of interconnections in networks that occur more frequently than expected by chance [55, 65, 90, 116, 118]. These network motifs, which correspond to over-represented topologies in primarily transcriptional networks, have been proposed to correspond to the building blocks of cellular circuitry [90], and each motif is postulated to play a specific information-processing role in the network. This has been an interesting and influential line of work; however, it does not consider the specific roles of individual proteins. Therefore, while these approaches may give a hint to the general organizational or design principles of biological networks—the syntax, so to speak, of biological networks—they do not capture the "semantics" of the networks, or the tendency of certain types of proteins to act together in order to accomplish diverse biological tasks. Network alignment approaches [24, 49, 52] provide an alternative way to begin to address this issue. Network alignment identify homologous proteins and (nearly) conserved patterns of interactions among them. These techniques, which rely on sequence homology, can identify network components that are conserved across species [24, 52, 76] as well as within species [49].

Complementing the work on analysis of protein interaction networks, several groups have used individual protein-protein interactions, independent of their con-

text within protein interaction networks, and have found protein sequence motifs and domains that co-occur significantly more often in pairs of interacting proteins than in non-interacting pairs [15, 31, 32, 81, 95, 113]. Yet, it may be that larger groups of specific types of proteins work together as the basic functional unit.

Here, we introduce a paradigm for analyzing protein interaction networks that explicitly exploits both local network topology as well as known characterizations of individual proteins. *Pathway schemas* are specified via combinations of topologies and annotations, and are used to describe recurring means with which different biological processes may be carried out. Pathway schemas associated with signaling, for example, can range in complexity from a kinase interacting with another kinase to a path of interacting proteins, where the first protein is a receptor and the last protein is a transcription factor [98]. Pathway schemas need not be linear; any topology of interactions is permitted. The instantiations of a pathway schema in an interactome correspond to all known sets of proteins, as well as the specified interactions between them, that are examples of the schema (Figure 3.1, A,B). Thus, a pathway schema is a diagram of an underlying organizational pattern in an interactome and its instantiations correspond to specific examples in support of that pattern.

In this chapter, we develop computational methods for automatically identifying the pathway schemas that are the *building blocks* of interaction networks. We define building-block pathway schemas as pathway schemas that are both recurrent in the interaction network and are over-represented with respect to their lower-order constituents. A key part of this procedure is identifying higher-order schemas that are not emergent from their constituent lower-order subschemas. We accomplish this computationally by generalizing the stub-rewiring (i.e., degree-preserving) algorithm of [65] to preserve, when appropriate, the distribution of specific labeled subgraphs.

Our work on uncovering pathway schemas differs from previous related work in

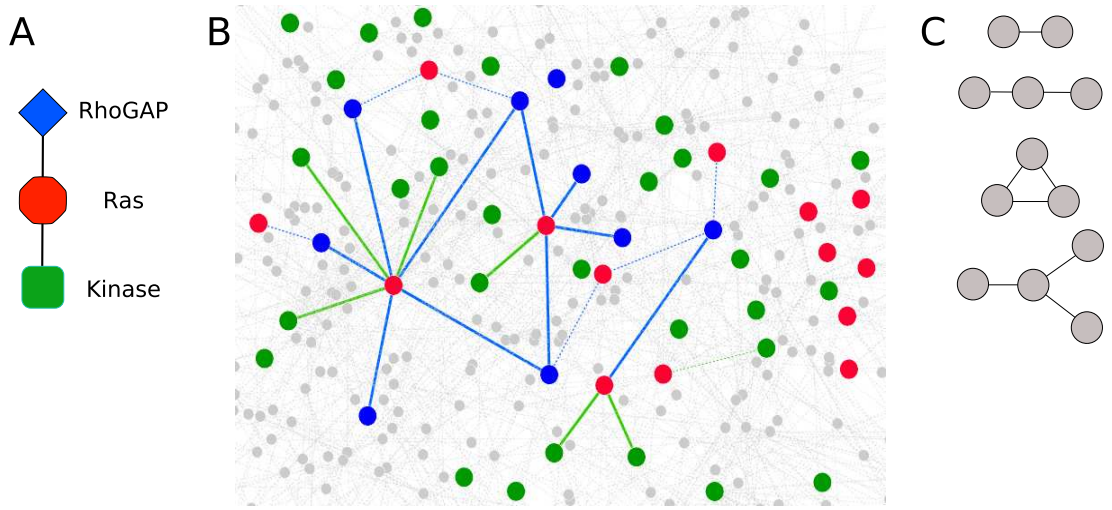

Figure 3.1: A, B: An example of a 3-line schema. A. The GAP-Ras-kinase schema is made up by a small GTPase of the Ras family (in this case, it is, more particularly, in the Rho subfamily), which is regulated by a GTPase Activating Protein (RhoGAP) and in turn regulates its effector kinase. B. Instantiations of the Kinase-Ras-RhoGAP schema in yeast. Here, a portion of the yeast physical protein-protein interaction network is shown. Ras family proteins are shown as red octagons, RhoGAP proteins as blue diamonds, and kinases as green squares. Ras-RhoGAP interactions are marked in blue, Ras-kinase interactions in green. Interactions that are instantiations of the RhoGAP-Ras-kinase 3-protein triplet linear schema are marked by thick solid lines; other RhoGAP-Ras and Ras-kinase interactions are marked by thin dashed lines. See **Methods** for construction of physical interaction network and determination of protein annotations.

C. Schema topologies that are considered in this study.

several key ways. In contrast to the work on uncovering network motifs [65], we endeavor to incorporate known features of individual proteins. In contrast to the work on network alignment [49], we aim to incorporate information beyond sequence homology, and focus on patterns of interactions that are significant even when considering their lower-order constituents. Finally, in contrast to the work on uncovering sequence motifs that co-occur in interacting proteins [95], we move beyond individual interactions and explictly consider local network topology.

We find that building-block pathway schemas are readily identified in the existing

protein interaction networks for yeast and human. Our approach considers four network topologies (Figure 3.1, C), varying from two interacting proteins to higher-order ones containing three edges; these include branched and cyclical topologies. While the concept of schema permits the investigator to focus on any protein property as an annotation, we choose for this purpose Pfam sequence motifs [4]. Pfam sequence motifs may correspond to specific molecular functions (e.g., kinase) or stuctural domains (e.g., PDZ domain), or may be uncharacterized (e.g., all Pfam-B motifs and some Pfam-A motifs). Additionally, it is estimated that nearly three-quarters of proteins sequences have at least one match to Pfam.

We uncover almost 300 network schemas of various complexity in the yeast interactome, and implicate 745 yeast proteins as members of building blocks identified in this manner. We show that instantiations of the automatically uncovered pathway schemas lead to subnetworks whose biological processes are functionally more cohesive, as judged by GO biological process terms, than subnetworks with identical topologies but no constraints on the proteins making them up. Moreover, we find that between 40% and 60% of the uncovered yeast schemas (depending on schema type) have instantiations in the human interactome. Together, these suggest that the uncovered pathway schemas correspond to recurring functional units. We also show how pathway schemas can be used to help assign function to uncharacterized sequence motifs. Overall, our work demonstrates that pathway schemas are a novel means for organizing interaction networks and are likely to play an important role in future attempts to partition, interrogate and annotate protein interaction networks.

3.2 Methods

3.2.1 Preliminaries

Interaction network. Yeast and human protein interaction data was downloaded from the BioGRID [6], release 2.0.20. The network contained in its largest connected component 4656 proteins and 27571 interactions among them for yeast, and 7368 proteins and 20020 interactions among them for human. These networks are filtered to focus on direct physical interactions, as described below.

Protein annotations. Pfam download 18.0 gives 3873 yeast proteins that contain at least one Pfam sequence motif. 1894 are annotated with a curated Pfam-A motif, and 2301 are annotated with an uncharacterized Pfam-B motif.

Network filtering. The physical protein interaction network was considered, with several filters imposed on the network. Attention was restricted to physical interactions that are highly likely to be direct. They were identified as those determined using the following experimental systems: Biochemical activity, Co-crystal structure, Far western, FRET, Protein-peptide, Reconstituted complex and Two-hybrid [37]. Additionally, pairwise interactions determined using Affinity capture-Western and Affinity capture-MS were used (these are interactions when a bait protein identifies at most one prey in an experiment).

Futhermore, in order to limit the effect of experimental error, the per-experiment degree of each protein (i.e., the number of interaction partners of the protein that were found by a single experiment) was considered, and if any protein was found to have per-experiment degree over 30, the interactions for that protein and that experiment were removed. Then, proteins which had no annotations according to the annotation system(s) under consideration were removed. Finally, proteins that had overall degree greater than 50 after this procedure were excluded from further analysis.

After this, Pfam annotation terms that describe fewer than two proteins in the resulting graph were removed from the list of terms under consideration; vertices which became unannotated as a result of this step were removed as well. Finally, to prevent redundancies in the resulting schemas, annotation terms that always co-occurred with some other annotation term were removed from the list of terms under consideration (i.e., term a is redundant with term b if the proteins annotated with a are always annotated with b as well).

After all filtering steps, the resulting yeast network has 2073 proteins described by 472 Pfam terms and 3871 interactions between the proteins, and the resulting human network has 4062 proteins described by 669 terms and 7284 interactions between the proteins.

Terminology. For completeness and (hopefully) clarity, we formally describe what we mean by pathway schemas and their instantiations.

A protein interaction netbwork is represented as a labeled graph $G = (V_1, E_1)$, with a vertex $v \in V_1$ for each protein and an edge $(u, v) \in E_1$ between vertices whose corresponding proteins interact. Each vertex v is labeled with a set of sequence features S_v; in this work, the features are limited to be Pfam sequence motifs. In the general case, each edge (u, v) may labeled by a set $T1_{(u,v)}$ containing the types of interactions observed (e.g., genetic or co-expression); however, for the purposes of this work, only physical interactions are considered.

A *pathway schema* is a graph $H = (V_2, E_2)$ where each vertex $v \in V_2$ is specified by a description d_v, which is a set of protein features, and each edge (u, v) is labeled by a set $T2_{(u,v)}$ of allowed edge labels. Here, we only allow one description per protein and consider only physical interactions, so $|d_v| = 1$ and $T2_{(u,v)} = physical$.

An *instantiation* of a pathway schema H in an interaction network G is a graph (V, E) where $V \subset V_1$ and $E \subset E_1$ such that there is mapping $f : V_2 \rightarrow V$ where for

each $v \in V_2$, $S_{f(v)}$ satisfies the description d_v, and for each $(u, v) \in E_2$, $(f(u), f(v)) \in E$ and $T1_{(u,v)} \cap T2_{(f(u),f(v))} \neq \emptyset$. In our case, an instantiation is just a subgraph of the physical interaction network with the same topology and protein properties as given by the schema. Note that two instantiations of schema may share proteins and/or interactions; however, two instantiations must differ in at least one protein.

Two instantiations (V, E) and (V', E') of the same pathway schema are *independent* if $V \cap V' = \emptyset$ (i.e., they are made up of non-overlapping proteins); along with the total number of instantiations, the number of independent instantiations of a schema is used to evaluate how prevalent a particular schema is in the interaction graph.

3.2.2 Uncovering pathway schemas

We consider schemas of four topologies containing up to three edges: *pairs*, consisting of annotations a and b, and denoted as $p(a, b)$; *triplets*, denoted as $t(a, b, c)$; *triangles*, denoted as $\Delta(a, b, c)$; and Y-shaped network schemas, $Y(a, b, c, d)$ (Figure 3.1, C). In the case of triplets and Y's, we allow instantiations to have additional edges (i.e., the endpoints of the triplet or any pair of endpoints of the "spokes" of the Y may be connected with an edge).

The overall procedure for uncovering over-represented pathway schemas is as follows; the steps are described in more detail below. First, each schema topology is considered independently, and for each topology, the number of instantiations for each schema s found in the interactome is tabulated. Second, for each possible schema s, its average number of instantiations in randomized networks is computed. Third, the schema is scored to favor frequently occuring schemas that also occur more often than expected by chance. Fourth, significance of scores is judged via computation of a false discovery rate (FDR). Finally, the schemas are filtered in order to focus on

those that are most interesting.

Randomized networks for computing scores. For each schema s that recurs in an interactome (i.e., has at least 2 instantiations), we compute how often it occurs in randomized networks. That is, we wish to know whether the schema occurs more often than expected. For each pairwise schema, we count how often it occurs in randomized networks that have been generated using the stub-rewiring approach of [61, 90]. For each triplet and triangular schema, we count how often the schema occurs in networks randomized so as to preserve the distributions of the pairs making them up. For each Y schema, we use the same approach, but consider random networks that preserve the distribution of triplets making up the Ys. In this manner, we aim to identify the schemas that are over-represented even when considering the distribution of the lower-order schemas making them up.

More specifically, for triplet schema $t(a, b, c)$, we generate random graphs that attempt to maintain the original number of interactions where one protein is labelled a and the other protein is labelled b, where one protein is labelled b and the other is labelled c, and where one protein is labelled a and the other is labelled c. This is accomplished by considering only proteins labelled with a, b or c. An edge is added between two proteins labelled with (say) a and b proportional to how much closer it brings the distribution to what is desired. As with the stub-rewiring approach, the degrees of these proteins are maintained; thus, edges are added only if doing so would not exceed the original degree of either protein. This process is continued until the three pairwise distributions are satisfied or no further edges can be added.

The same process is used to generate random graphs for triangular schemas $\Delta(a, b, c)$.

For Y schemas $Y(a, b, c, d)$ (with a at the center), edges are added in a way that maintains the distribution of triplets b-a-c, c-a-d, and b-a-d.

For pairwise schemas, 65 randomized stub-rewired networks are generated, and the average number of times that each schema occurs in these networks is computed. For each triplet, triangular and Y schema s, 65 randomized networks are generated as described above, using the appropriate method, and the average number of times that it occurs in these networks is computed.

Scoring schemas. For each schema s, let $count_s$ be the number of times it occurs, and avg_s be the average number of times it occured in randomized networks. The score for schema s is given by

$$(count_s + 1) \log \left(\frac{count_s + 1}{avg_s + 1} \right).$$

The addition of the pseudocount of 1 downweighs the contribution of very rare schemas that could otherwise obtain abnormally high scores due to very small (or 0) average counts in the random graphs.

Significance model. For each putative recurring schema found in the real network, we obtain a score reflecting its frequency and overrepresentation compared to the random graphs. In order to evaluate the significance of the scores thus obtained, for each topology of schema considered, we repeat this procedure with 40 *iteration* graphs created by stub-rewiring algorithm of [61]. Since all associations in these randomized networks occur by chance, we can use them to calculate the FDR for each score, or the fraction of schemas with score $\geq x$ that arise from chance alone. It can be computed as

$$\frac{\frac{1}{40} \sum_{\text{iteration graph i}} \#\text{ putative schemas in graph i with score } \geq \text{s}}{\#\text{putative schemas in the real graph with score } \geq \text{s}}.$$

Scores with FDR < 0.05 are considered significant.

3.2.3 Evaluating functional coherence

In order to test the biological relevance of the schemas discovered, we look at the functional coherence of groups of proteins that form instantiations of significant schemas and compare then to background. By functional coherence we mean the tendency of proteins to participate in the same biological process. We use the Gene Ontology Biological Process annotations [2] as the "gold standard" of fuctional annotation. One complication that arises in this situation is the hierarchical (or, more precisely, DAG) nature of the Gene Ontology. We address this issue by mapping the different terms of the GO to a common scale. In our case, this scale is the probability of a randomly chosen group of proteins of a given size having that annotation, as determined by the hypergeometric distribution based on the number of proteins annotated with this term.

We consider each schema topology separately. For each topology, we compile the set of instantiations of all significant and recurrent ("building-block") schemas of that topology, excluding schemas consisting of the same proteins as some other schema already in the set. For the "background" set, we enumerate all subgraphs of a given topology that occur in the interaction network. In order to avoid any bias that might arise from Pfam annotations, only proteins having any Pfam annotation are considered when building the sets of subgraphs. Furthermore, both for the building-block-schema instantiations and for the background sets, we require all or most proteins in the subgraph to have non-trivial biological process annotation. In the case of subgraphs corresponding to schemas with 3 or fewer proteins, we require all proteins to have biological process annotations; in the case of 4-protein subgraphs, we permit one "central" protein (i.e., a protein whose degree in the subgraph is $>$ 1) to be unannotated (if a "peripheral" node is not annotated, then the annotated portion of the subgraph would just be a 3-protein subgraph).

For each schema topology, for both building-block-schema instantiations and the background set of subgraphs, we use the following method to evaluate the functional coherence of a set of schema instantiations of a given topology. For each subgraph consisting of N proteins, n of which have non-trivial biological process annotation, we determine the least common ancestor of their annotation in the GO graph; if there are multiple LCA's that are not comparable in the partial order imposed by the GO structure, we select the one that annotates the smallest number of proteins in yeast, breaking ties arbitrarily. Note that if the proteins are not known to be functionally related, the LCA of their annotations would be the trivial annotation of biological_process. Then, the "specificity" of this LCA is calculated as the probability of n proteins having that annotation, using the hypergeometric distribution: $p = \frac{\binom{A}{n}\binom{T-A}{N-n}}{\binom{T}{N}}$, where A is the number of proteins annotated with the term or its descendants in the reference genome, and T is the total number of annotated protens in that genome. Then, for a given value of p, for both the significant-schema interactions and the background set of subgraphs, we can measure the functional coherence of each as the fraction of subgraphs whose constituent proteins have annotation LCA with p-value at most p.

3.3 Results

3.3.1 Pathway schemas in the yeast interactome

In the filtered yeast interaction network, there are 2838 pair, 23395 triplet, 999 triangular, and 101840 Y schemas consisting of proteins with Pfam annotations. Of them, 831 pairs, 8491 triplets, 283 triangles, and 52019 Ys occur at least twice. Due to the multiplicity of annotations on many proteins, there is often overlap between schemas of the same topology. Therefore, we remove schemas that are "subsumed" by another

schema of the same topology; i.e., we remove schemas whose instantiations are a subset of the instantiations of some other schema of the same topology. This leaves 657 pairs, 5080 triplets, 161 triangles, 7849 Ys. All numbers reported from now on will be computed after application of this filter. Using our scoring procedure with background averages for schemas computed using the stub-rewiring randomizations of [61] for pairs, the pair-distribution preserving randomizations for 3-protein schemas, and triplet-distribution preserving randomizations for Y schemas, and a false discovery rate ≤ 0.05, we get: 196 pairs, 213 triplets, 89 triangles, 242 Ys. We note that the false discovery rate has a built-in multiple-hypothess correction.

Before we analyze these results further, we apply an additional filter to focus on significant schemas that are truly recurrent. Rather than just looking at schemas that have at least two instantiations, we impose a stricter requirement, namely that they have at least two independent (i.e., non-overlapping) instantiations. We note that there are many interesting schemas whose instantiations would not be independent; for example, the Skp1-cullin-F-box (SCF) complexes associated with ubiquitination only vary in the identity of the F-box protein. After applying these filters, we are left with 156 pair, 85 triplet, 25 triangle and 31 Y building-block schemas. These results are summarized in Table 3.3.1.

These building-block schemas in yeast are displayed in Figures 3.2 -3.5. For the purposes of visualization, higher-order schemas are represented as graphs of lower-order schemas: triplets as connected vertices in a graph of pairs, triangles as colored triangles in a graph of pairs, and Ys as triangles in the graph of triplets. The uncovered building-block schemas are comprised of a wide-variety of Pfam sequence motifs. For example, the network comprised of pairwise building-block schemas (Figure 3.2) consists of more than 140 Pfam motifs representing a wide-variety of functions including signaling, transporter activity, intracellular trafficking, and DNA packaging.

Topology	Total	Recurring	Non-redundant	Significant	Building Block
Pairs	2838	831	657	196	156
Triplets	23395	8491	5080	213	85
Triangles	999	283	161	89	25
Ys	101840	52019	7849	242	31

Table 3.1: Statistics for uncovering schemas in the yeast interactome. **Total** gives the total number of schemas of each topology that is found in the filtered yeast interactome; each schema is described via Pfam motifs. **Recurring** gives the number of these schemas when each is required to have at least two instantiations. **Non-redundant** gives the number of recurring schemas that are additionally filtered so that each schema whose instantiations are a subset of another schema is removed. **Significant** gives the number of non-redundant, recurring schemas that are also found to be significant at an FDR level of 0.05. **Building block** gives the number of significant, non-redundant, recurring schemas that occur at least twice independently in the filtered yeast interactome.

Many of the uncovered schemas recapitulate known biology. For example, many of the triangular schemas correspond to known complexes, such as the spliceosome (reflected by the LSM motifs) or the SNARE vesicle-fusion machinery, and several of triplet schemas contain combinations of domains associated with signaling [72]. (See Section 3.3.5 for a detailed analysis of the uncovered schemas relating to the Ras family of signaling proteins.)

3.3.2 Pathway schemas are functionally coherent

Though a case-by-case analysis of uncovered building-block schemas is illustrative, here we systematically evaluate the biological significance of these schemas by analyzing their functional coherence. That is, given an instantiation of one of these schemas, does it show enrichment for a particular biological process? As described in the methods, for each schema topology, we compile the building-block schema set and the background set, and for each instantiation determine its most descriptive biological process annotation. We visualize the functional coherence of the schema

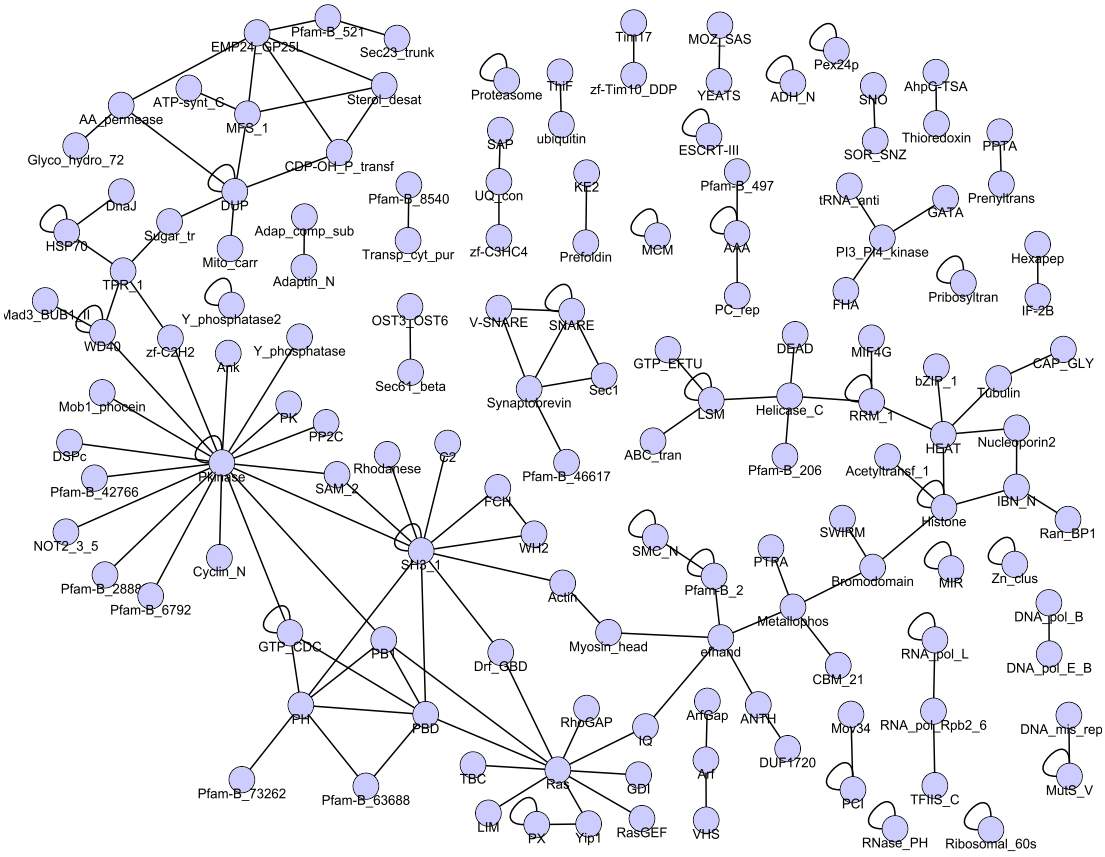

Figure 3.2: Building-block pairwise schemas uncovered in the yeast interactome. Each node is labelled with a Pfam domain and an edge between two Pfam domains corresponds to a building-block pathway schema.

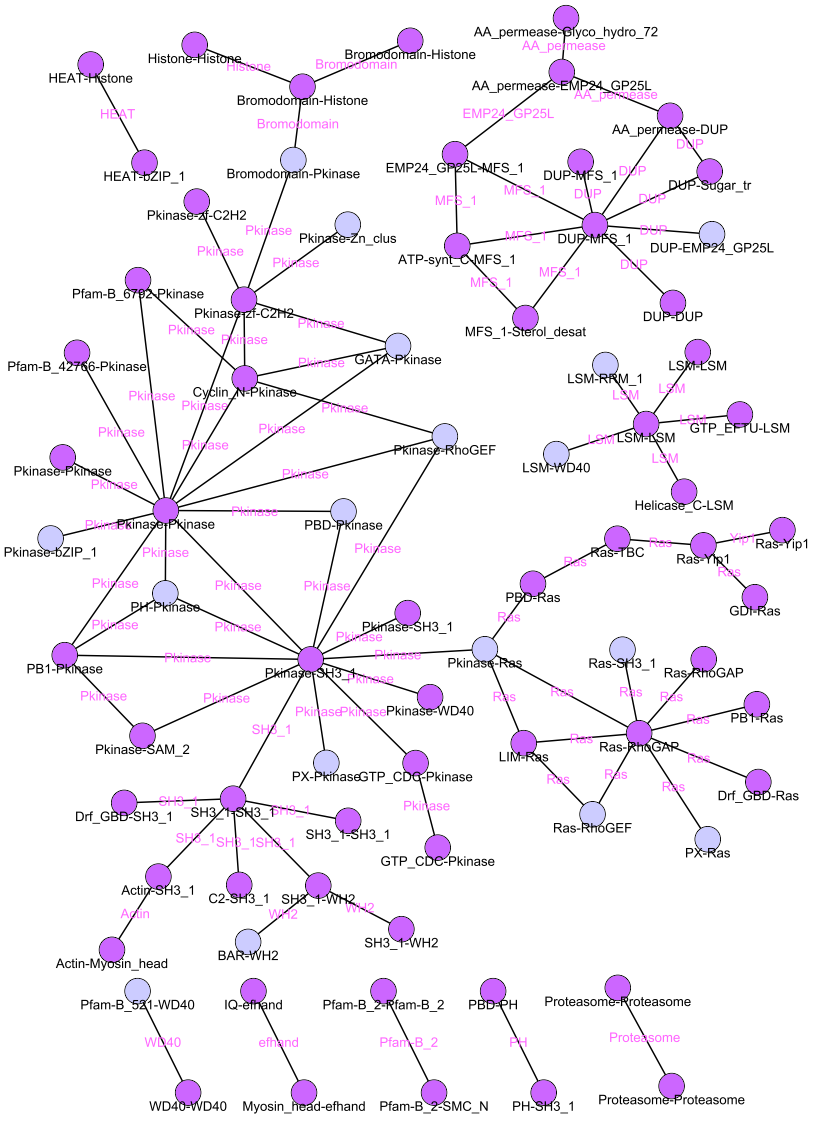

Figure 3.3: Building-block triplet schemas uncovered in the yeast interactome, represented as connected vertices in a *pair graph*. The central domain of the triplet is shown in pink as the label of the corresponding edge. Pairs that are "building blocks" are represented as purple vertices, other pairs as blue vertices.

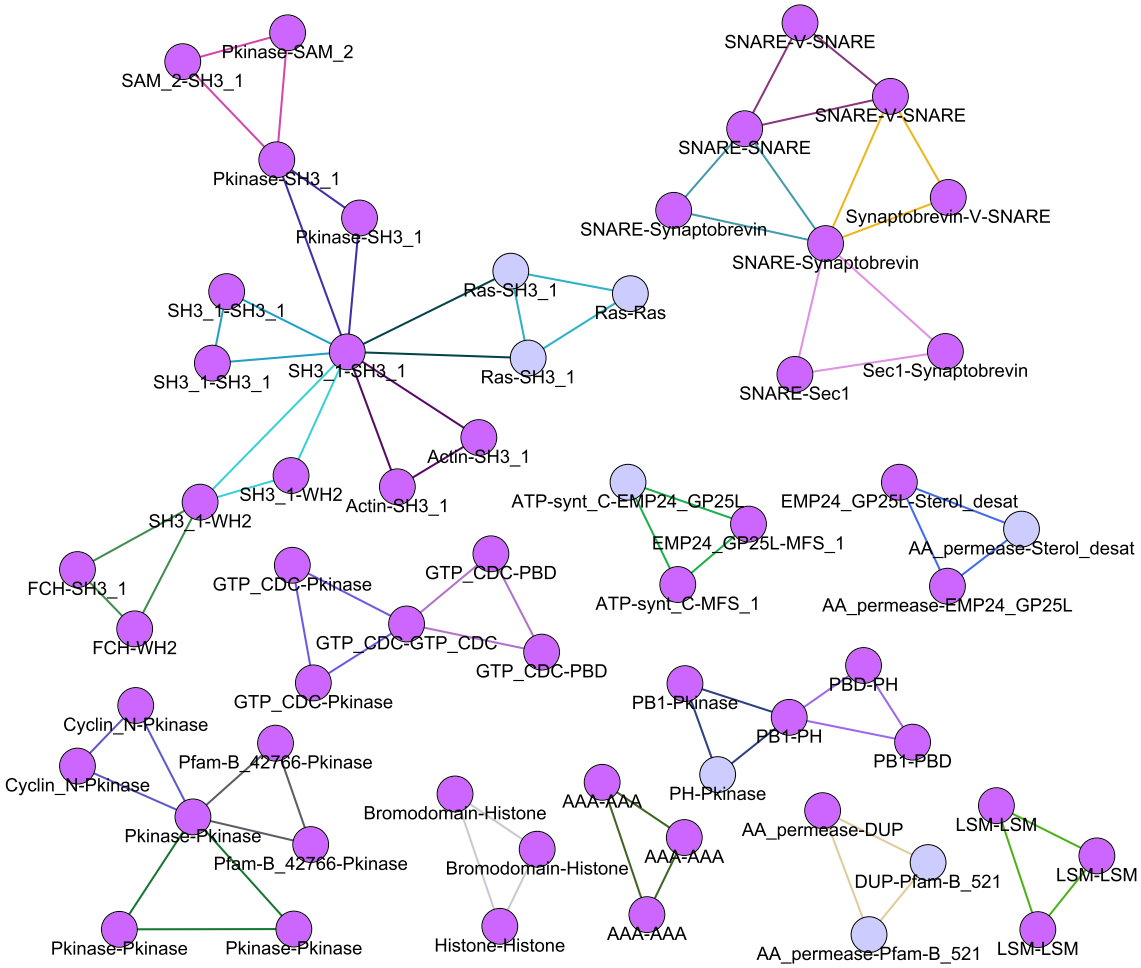

Figure 3.4: Building-block triangular schemas uncovered in the yeast interactome. represented as triangles in a *pair graph*. Edges that belong to the same triangle are colored with the same color. Pairs that are "building blocks" are represented as purple vertices, other pairs as blue vertices.

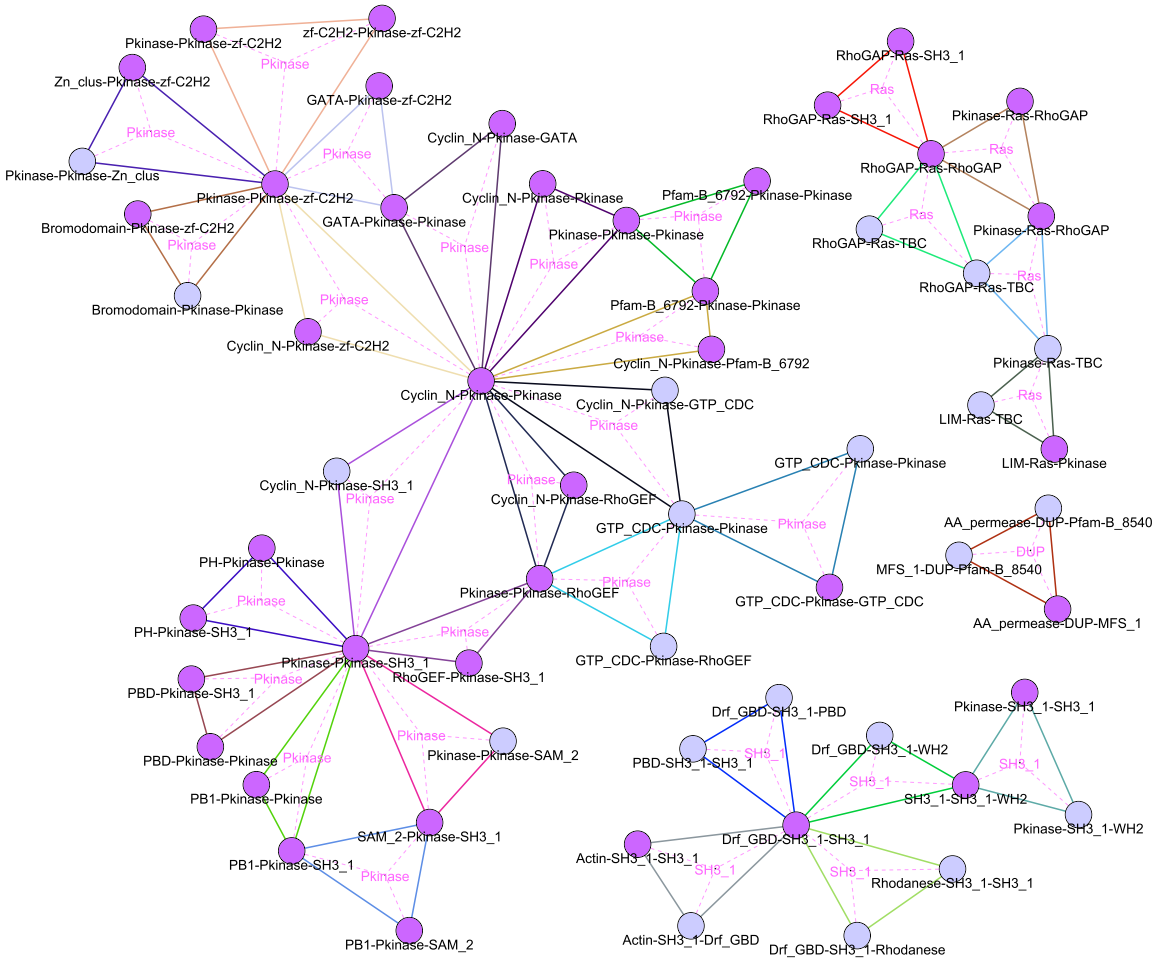

Figure 3.5: Building-block Y schemas uncovered in the yeast interactome, represented as colored *triangles* in a *triplet graph*. For ease of visualization, the central node of each Y is given inside the triangle and connected to the vertices by pink dashed lines. Triplets that are "building blocks" are represented as purple vertices, other triplets as blue vertices.

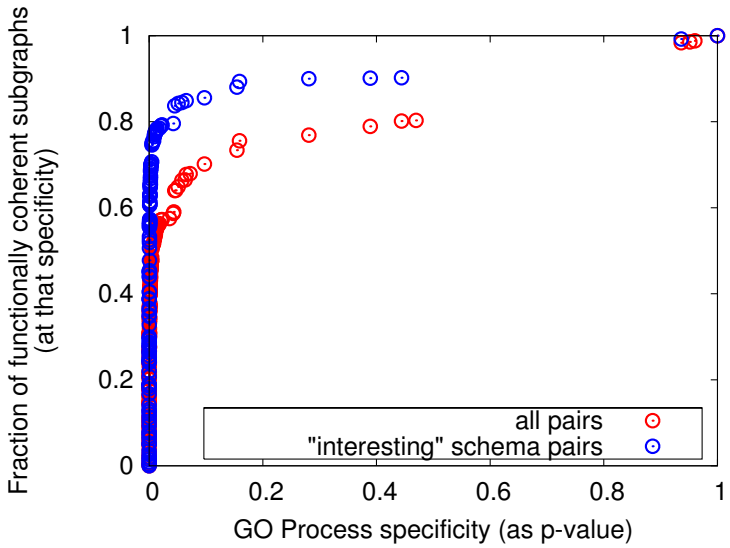

Figure 3.6: Functional coherence of uncovered pairwise yeast schemas. Proteins making up instantiations of building-block pairwise schemas are more functionally coherent than the background set of all interacting annotated proteins. See text for details.

set and the background set by plotting, as a function of the specificity of a biological process term (measured via a p-value, with small p-values corresponding to more specific terms), the fraction of subgraphs in each set whose most descriptive biological process annotation has at most that p-value. Thus, a set of schemas leads to more functionally coherent instantiations than another set if its curve is "above" the others in these plots. The results of the comparison between the schema set and the background are presented in Figures 3.6-3.8. One can see that for all topologies, the instantiations of significant schemas are more functionally coherent throughout the entire range of p-values than background subgraphs of the same topology.

3.3.3 Yeast pathway schemas conserved in human

To study the extent to which building-block yeast schemas are conserved through evolution, we obtained their instantiations in the human interaction network. We

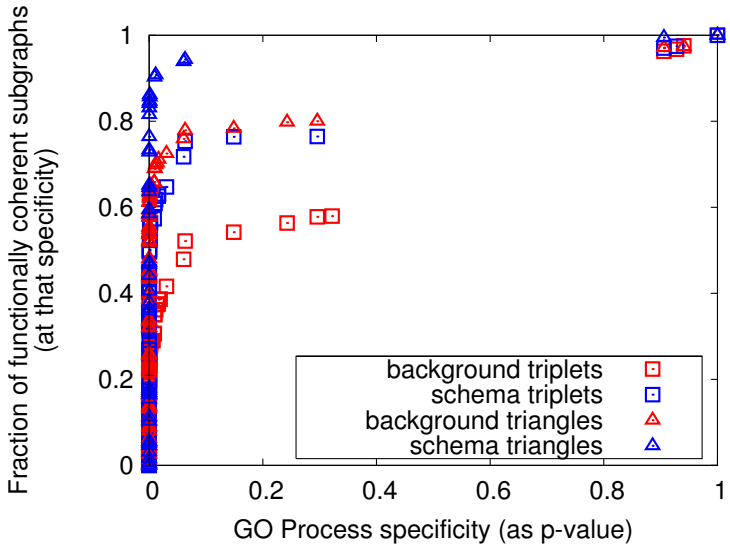

Figure 3.7: Functional coherence of building-block 3-protein triplet and triangular schemas. See text for details.

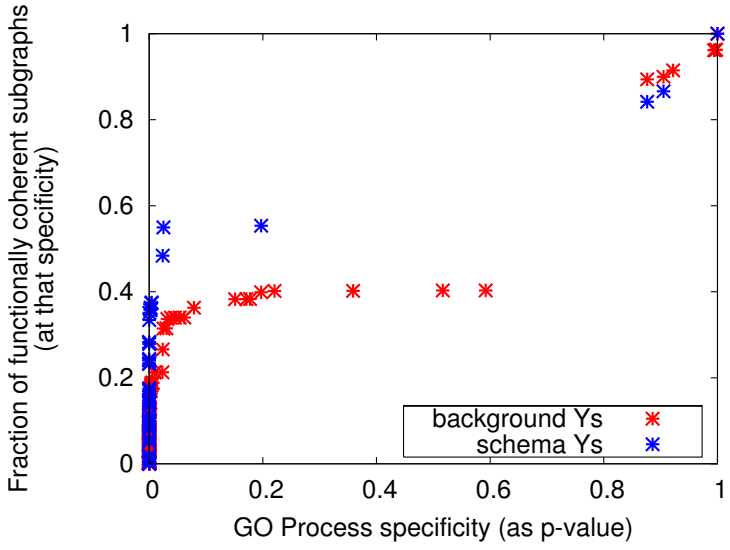

Figure 3.8: Functional coherence of building-block Y schemas found in yeast.

searched the full human BioGRID physical interaction network for instantiations of the yeast network schemas (i.e., no filters were used to build the network).

The results are displayed in Table 3.3.3 and summarized below. Briefly, the fraction of yeast schemas that are conserved in human is similar for pairs and triangles, and decreases slightly for triplets and Ys, with 94/156 (60%) interesting pairs having at least one human instantiation, compared with 45/85 (53%) triplets, 15/25 (60%) triangles, and 13/31 (42%) Ys. For comparison, if we consider 100 randomly selected schemas of each topology, 19% of pairs, 9% of triplets, 5% of triangles, and 4% of Ys have at least one human instantiations.

These numbers show an increase in the "gap" between conservation of "building-block" and random schemas with increasing schema complexity, lending further evidence to the correctness of criteria for finding meaningful schemas.

Topology	**Yeast Building Block Schemas**	**Conserved in Human**	**Background Conservation**
Pairs	156	60%	19%
Triplets	85	53%	9%
Triangles	25	60%	5%
Ys	31	42%	4%

Table 3.2: Conservation of uncovered significant yeast schemas in human. **Yeast Building Block Schemas** gives the number of yeast building block schemas uncovered for each topology. **Conserved in Human** gives the fraction of these schemas which have at least one instantiation in the human interactome. **Background conservation** gives the fraction of background subgraphs of the same topology as the schemas being considered that have an instantiation in the human interactome.

3.3.4 Pathway schemas in the human interactome

In order to compare building-block schemas across genomes, we repeated the pairwise building block schema finding procedure on the human interaction network. We found

29 building-block (i.e., significant and recurrent) pairwise schemas that are found in both yeast and human. As expected, these schemas represent some of the basic processes that happen within the cell: signaling, including the regulation of Ras family GTPases, vesicle fusion, cyclin regulation of kinases, ubiquitination, and so on.

Building-block pairs that are shared between yeast and human as well as those that are building block in one organism but only conserved (not identified as "building blocks") in the other are shown in Figure 3.9. Building-block pairs that are unique to either yeast or human and are not conserved in the other organism, are given in Figures 3.10 and 3.11, respectively.

3.3.5 Schemas recapitulate biology: focus on the Ras family

As an example, we focus on schemas involving proteins of the Ras family (PF00071). Ras is a family of small GTPases, which are active when bound to GTP; they inactivate themselves by slowly converting the GTP to a GDP. The guanyl-nucleotide exchange factor (GEF) facilitates the exchange of the GDP to a GTP (the latter is present at higher concentration) and thus activates the Ras protein, whereas the GTPase activating protein (GAP) increases the GTPase activity of Ras and thus deactivates it.

The pairwise schemas involving the Ras family reflect some of the biology that is known about the Ras family of proteins. As we expect, we see the schemas Ras-RhoGAP and Ras-RasGEF, both of which reflect the regulatory interactions of Ras proteins that were described in the preceeding paragraph (Rho is a subfamily of Ras). In the same category is the pairwise schema consisting of Ras and the TBC domain that belongs to GAPs of the Rab subfamily of Ras proteins. In addition, we have a Ras-GDI pair, which reflects the additional regulation mechanism of the Rab sub-

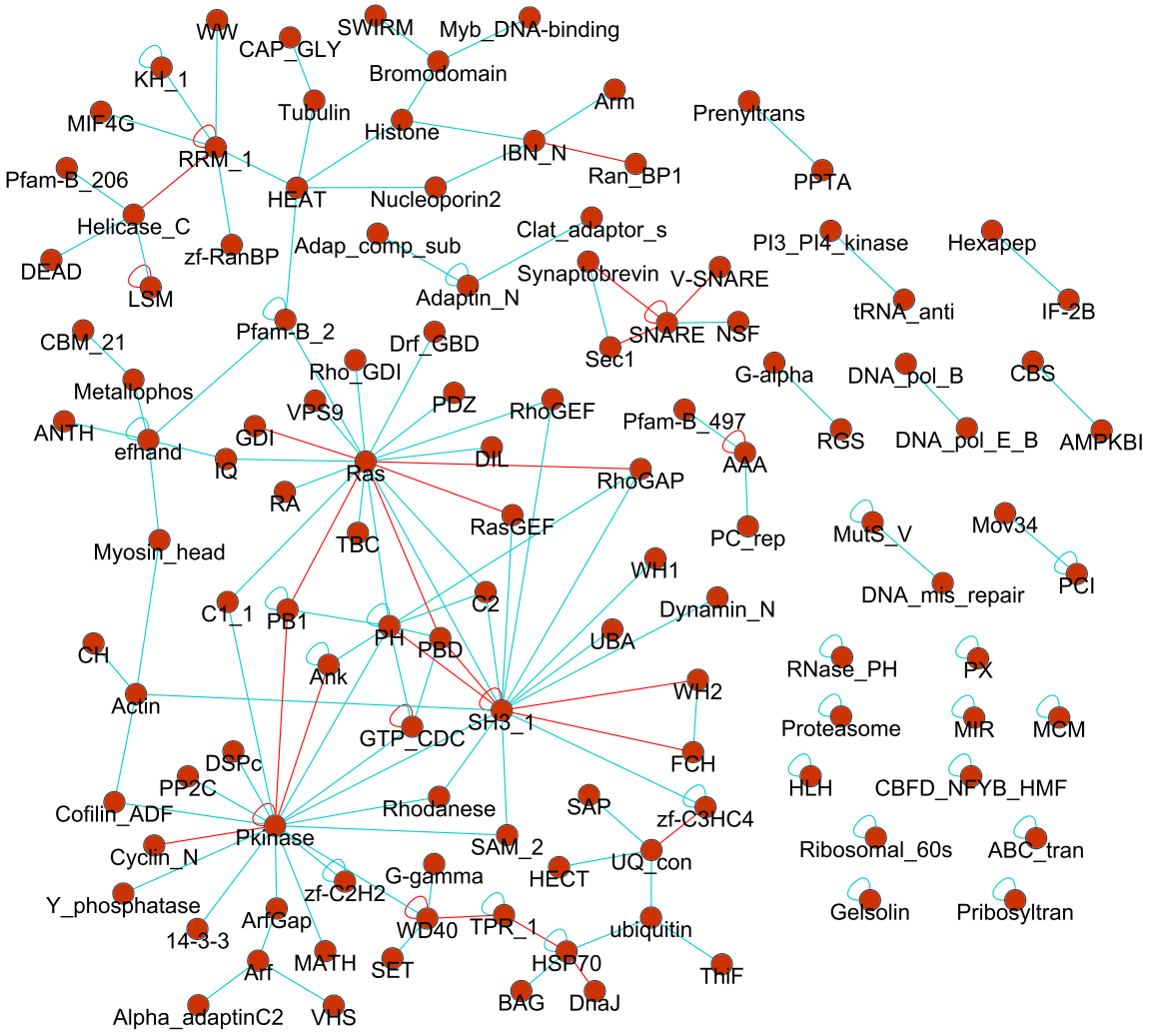

Figure 3.9: Pairwise schemas conserved in human and yeast. Schemas that are building blocks in both organisms are indicated via red edges. Schemas that are building block in one organism and conserved in another are indicated with light blue edges.

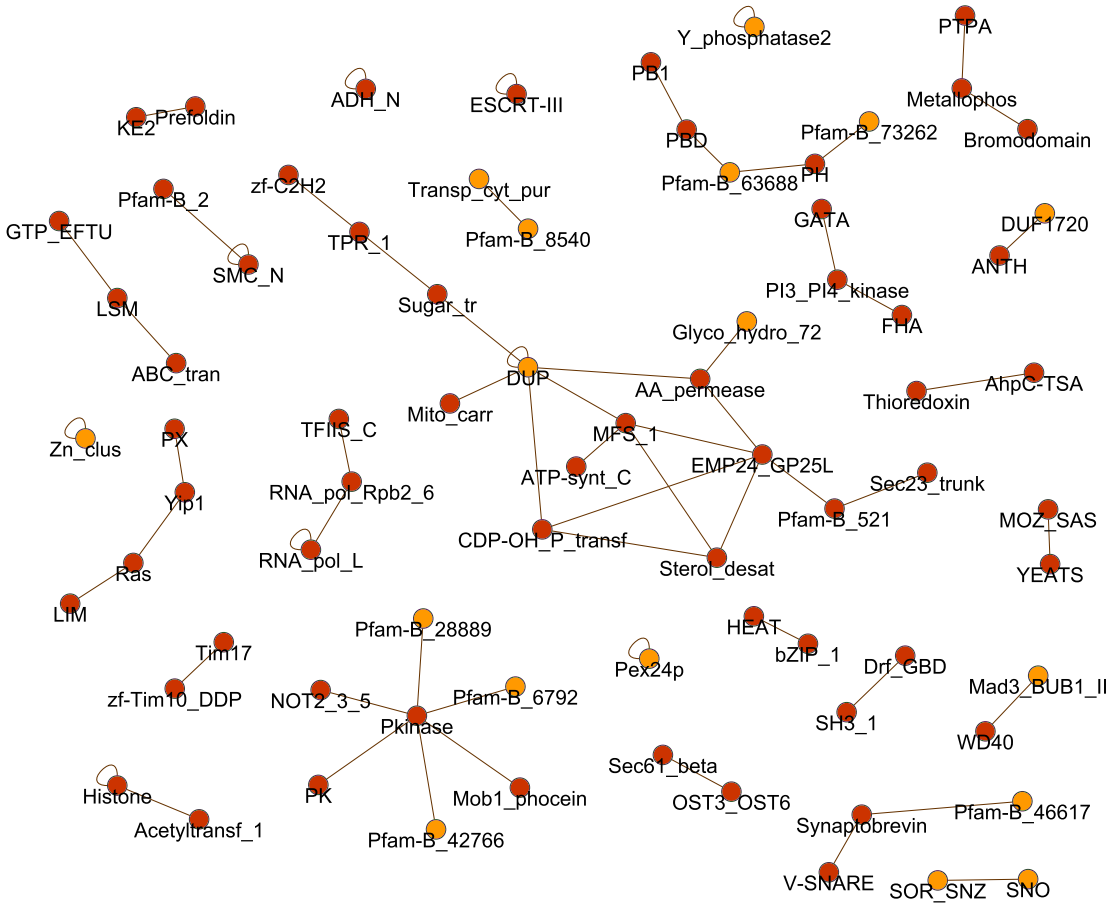

Figure 3.10: Building-block yeast pairwise schemas that are not conserved in the human interactome. Red vertices indicate Pfam motifs that are found in both organisms, orange vertices indicate Pfam motifs that are found in yeast but not human.

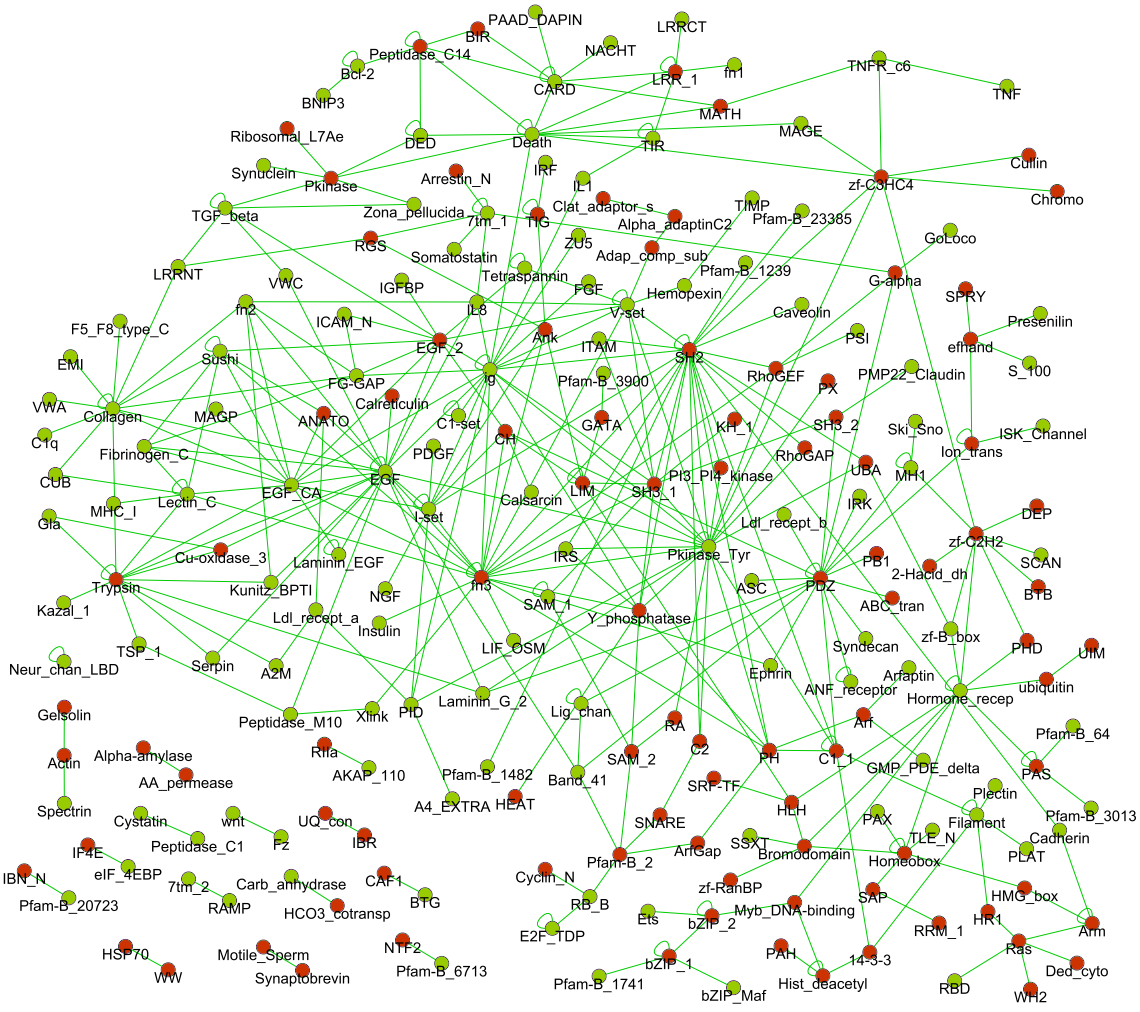

Figure 3.11: Building-block human pairwise schemas that are not conserved in the yeast interactome. Red vertices indicate Pfam motifs that are found in both organisms, green vertices indicate Pfam motifs found in human but not yeast.

familiy of Ras proteins, where the guanyl dissociation inhibitor (GDI) slows the rate of dissociation of GDP from membrane-bound Rab proteins. The Yip1 family of proteins in turn may act as GDI displacement factors [92] for a group of Ras-like proteins associated with Golgi membranes and/or act as membrane recruiters of these proteins [115]. Other Ras pairwise schemas include schemas involving Ras-binding domains, such as the Diaphanous GTPas-binding Domain (Drf_GBD, PF06371) contained by Rho effectors and the P21-Rho-binding domain (PBD, PF00786), or involving motifs that reflect the biological role of Ras families, such as the IQ calmodulin-binding motif (PF00612) and the PB1 family of proteins associated with signaling.

The Ras triplets include those that are built from some of these pairs (because of our methodology, they are significant even conditioned on the fact that they are built of significant pairs), as well as those that contain pairs that were not themselves significant. For example, in the Pkinase-Ras-RhoGAP triplet, the Ras-RhoGAP pair is significant as a pair, but the Pkinase-Ras pair is not. Similarly, Ras Ys are made up of a combinatin of triplets that are "building block" and those that are not (if we consider the constituent pairs, most of them are "building blocks").

The two "building block" triangles involving Ras in yeast are Ras-SH3_1-SH3_1 and Ras-Ras-SH3_1. Interestingly enough, in one of the instantiations of the former triangle, Ras1-CDC25-SDC25, the two SH3_1 proteins are in fact GEFs for Ras1 that are capable of forming a dimer with each other, whereas in the second group of instantiations, in which the Ras protein is CDC42, this is not the case. The Ras-RasGEF-RasGEF triangle, however, occurs only once in the network, and therefore is not recurring. Thus, here the schema-finding algorithm may in fact obscure the "semantics" underlying the schema.

Now we briefly address Ras pairs in human. They are dominated by interactions with Ras-associated protein domains, which, in addition to the familiar regulators

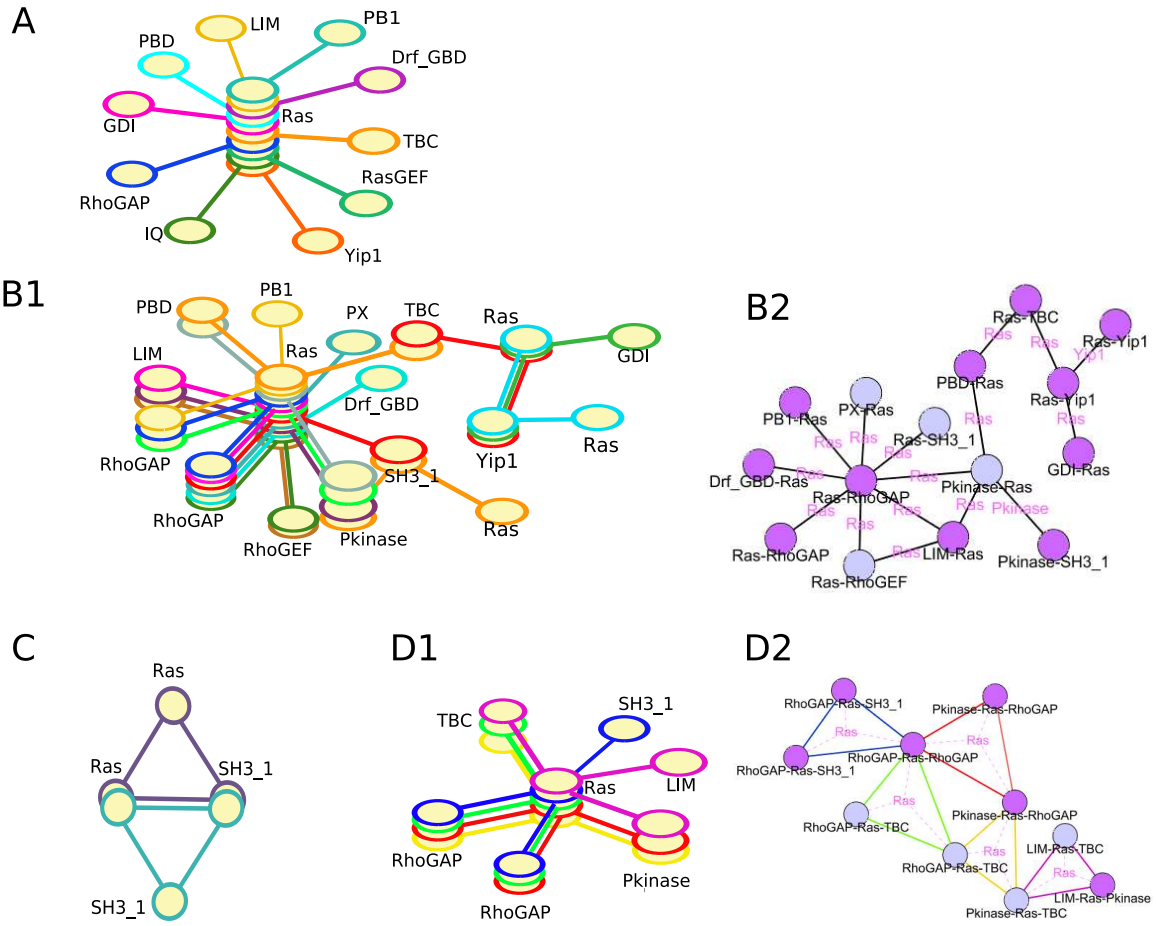

Figure 3.12: Schemas involving Ras proteins: (A) pairs, (B) triplets, represented as graphs (B1) and as a pair graph (B2), (C) triangles, (D) Ys, represented as graphs (D1) and as triangles in a triplet graph (D2).

such as GAP and GEF proteins includes families like RBD (PF02195), the Raf-like Ras binding domain, or the RA (PF00788) family of RasGTP effectors, as well as more general signalling protein motifs such as the PBZ (PF00595).

3.3.6 Using schemas to annotate domains and proteins

We use schemas to predict the function of domains and proteins of unknown function. Towards this goal, we focus on families of unknown function which are involved in

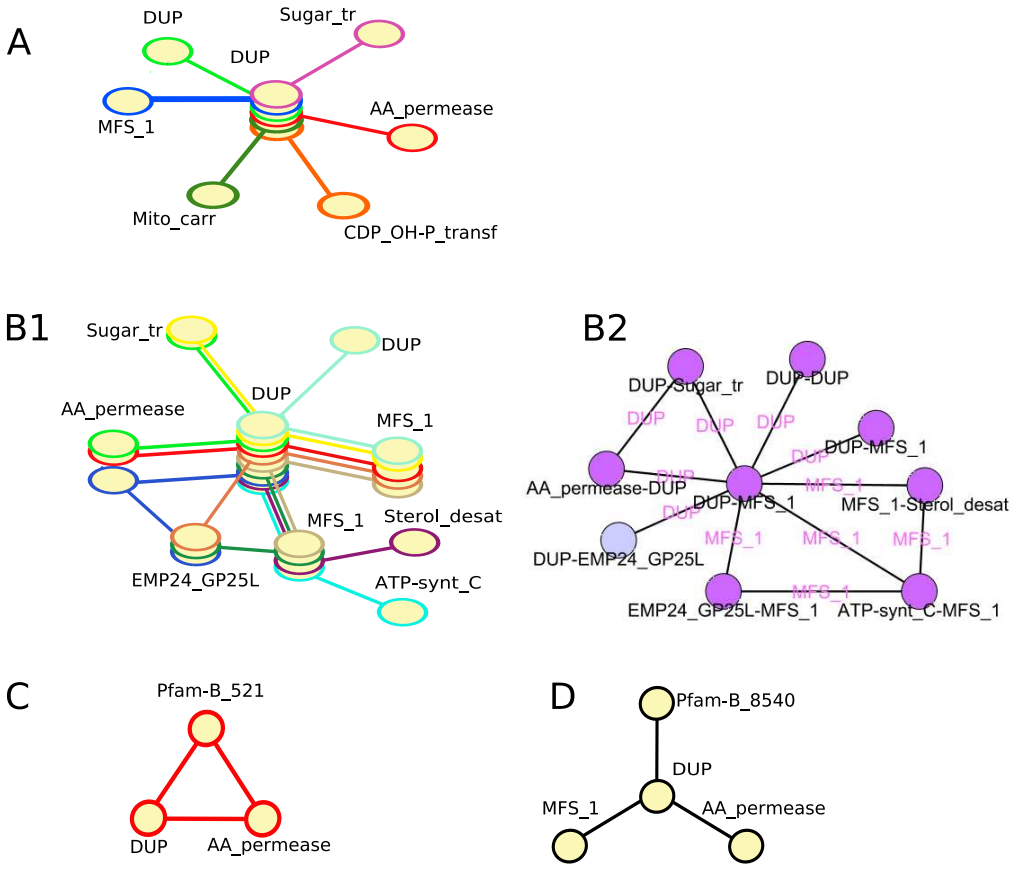

Figure 3.13: Schemas involving DUP proteins: (A) pairs, (B) triplets, represented as graphs (B1) and as a pair graph (B2), (C) triangle, (C) Y.

more than one schema in which they associtate with different domains, as well as on associated proteins. Two families which lend themselves to this task are: DUP and MAGE. The first is a family of yeast membrane proteins of uknown functions, and the second is a "family of genes of unknown cellular function that are expressed in a wide variety of tumours but not in normal cells, with the exception of the male germ cells, placenta, and, possibly, cells of the developing embryo." [4].

DUP

"One of the most curious gene families in yeast" [84], the DUP family consiss of 23 members [19], 21 of which have physical interactions, as listed in the BioGRID, although some of these members may not be true genes. These proteins are specific to the hemaiscomycete phylum [19]. All but one of DUP proteins are localized to different membranes in the cell and, with a few exceptions, have unknown functions. The exceptions are the pairs Mst27/Mst28 and Prm8/Prm9 which form dimers, as well as the Cos3 protein. The first pair of protein are involved in vescicle formation and bind COPI and COPII vesicles [84], and the second pair are pheromone-induced proteins. The Cos3 protein is involved in sodium resistance that interacts with the Na+/H+ antiporter Nha1p [19], and Cos10 which is implicated in endocytosis [19].

We find that the DUP proteins are involved in multiple schemas of every topology (Figure 3.13). These schemas are dominated by interactions with members of transporter families: MFS_1, Sugar_tr, AA_permease. Interactions with proteins of these families form the "core" of the DUP schemas in that these sub-schemas occur as part of every schema involving DUP; the majority of DUP proteins' schema interaction partners belong to these families. We believe that this provides strong evidence that the DUP family consists of proteins that are associated with membrane transporters. This is in line with the suggestion, based on the information about Cos3, that Dup proteins may activate or stabilize membrane proteins [19].

MAGE

The MAGE homology domain consists of 200 amino acids. Originally, MAGE proteins were found to be expressed in tumors, although later members of the family that are expressed in normal tissue were identified. There are 55 MAGE genes and putative genes in human [10], 32 of which are listed as such in Pfam, of which 9 have physical

interactions as described in the BioGRID. The MAGE family is poorly characterized functionally; of the 9 genes that have interactions, one, NRAGE (MAGED1) was implicated in nerve growth factor receptor (NGFR) signalling, and another, necdin was implicated in neuronal differentiation. Some facts that might shed light on the connection between MAGE proteins and tumors are that NRAGE contributes to cell cycle arrest and NGF-dependent apoptosis within sympathetic neuron precursors cells [83]; and that it has been found that NRAGE positively [112] and necdin [102] and members of the MAGE-A subfamily [66] negatively regulate p53, a key tumor-suppressor transcription factor, suggesting an explanation of the link between MAGE proteins and tumors.

We found the MAGE family to participate in pairwise schemas with two protein families: the Death domain and the RING family (zf-C3HC4) (see Figure 3.14). The Death domain is associated with apoptosis, and the RING finger is associated with E3 ubiquitin ligases, which perform the final step in protein ubiquitination. In many cases, ubiquitination targets proteins for destruction by the proteasome, although in other situations, ubiquitination plays other biological roles, for example, acts in DNA repair; in those cases the topology of the ubiquitin chain(s) may be different from that of proteolytic-marker ubiquitination [100]. In addition, RING E3 ligases may attach to proteins other small molecule markers that are similar to but distinct from ubiquitin, such as SUMO (small ubiquitin-like modifier) proteins. The cellular role of these modifications is not always understood [62].

Of 9 MAGE proteins in the BioGRID, 7 are interacting with Death and/or RING proteins. The MAGE-RING schema is more abundant, wtih 6/7 of the MAGE proteins interacting with a RING protein, whereas only 3 of the 7 MAGE proteins interact with Death domain proteins. These 3 proteins form an "interaction cluster", in which two of the three proteins are also interacting with RING finger proteins.

The RING proteins which interact with MAGE proteins are quite diverse. Among them are the inhibitor-of-apoptosis protein BIRC2, ligand of numb-protein X (LNX) that is implicated in tumorogenesis, and two members of the TRIM/RBCC family of proteins, one of which, TRIM27/RFP is implicated in apoptosis, and the other, TRIM31 may be involved in it because it shares the apoptosis-associated RBCC moiety [20]. Other RING members of the schema instantiations, such as ZSWIM2, are less characterized.

We also note that in addition to these proteins, there are other proteins that should be included in instantiations of the schema, but were not because their RING annotation scores from Pfam were not high enough (they were marked as "context" annotations by Pfam). These include another inhibitor of apoptosis BIRC4 (interacts with NRAGE) and another TRIM family member TRIM37 (interacts with MAGEB18).

These data suggest a connection between MAGE proteins and apoptotis; if real, this connection would likely shed light on the association between some of the founding members of the MAGE family members and cancer. It is possible that ubiquitination plays a role in this connection. The link between ubiquitination and apoptosis is a subject of investigation; for example, inhibition of proteasome has been shown to act as both promoter and inhibitor of apoptosis, likely by affecting the degradation of inhibitors or promoters of apoptosis [1, 69]. MAGE proteins may provide a link between these two processes. However, it is also possible that several mechanisms are in fact at work, and that, for example, the MAGE-RING pairwise schema may in reality consist of semantically different patterns of interaction between proteins of these two families. In particular, it is possible that the nature of the various MAGE-RING interactions may be different for different pairs. All these hypotheses require further biological investigation.

Additionally, due to the topology of the instantiations of the MAGE schemas,

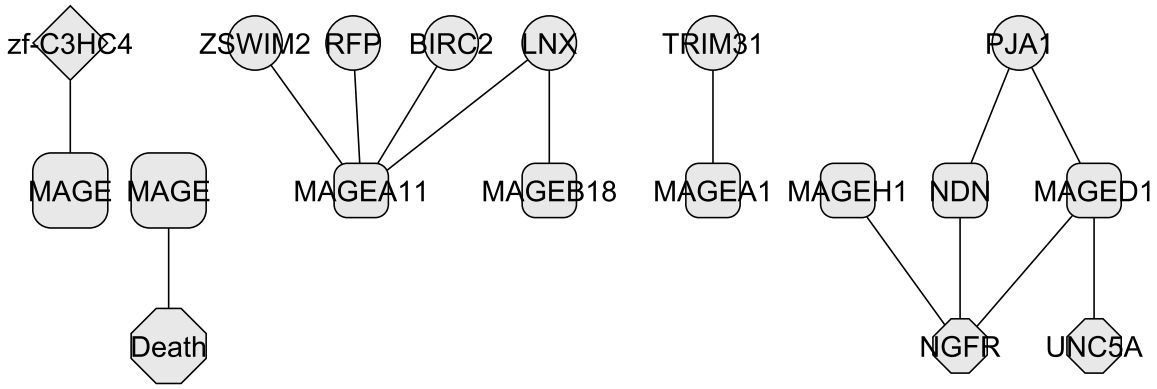

Figure 3.14: MAGE pairs

namely, the existence of a cluster of 3 MAGE proteins that interact with Death proteins (and, in the case of 2 of them, with RING proteins), we may hypothesize the existence of a yet-to-be discvered interaction of the third protein, MAGEH1, with a RING protein, possibly PJA1 (which interacts with the two other members of the cluster). This prediction too needs experimental confirmation or refutation.

3.4 Discussion

3.4.1 Schemas as network building blocks

In this work, we introduce the notion of network schemas and suggest a procedure for discovering schemas that are "building blocks" of the network. At first, one may focus simply on recurrence of schemas (i.e., take as building blocks schemas that have a least two instantiations in the network). As is seen from Table 3.3.1, this gives us a large number of schemas of each topology. While we might try to make our focus more specific by increasing the minimum number of schema occurrences from two

to a higher number, any method based on counting is naive and does not take into account the relative abundance of different types of labels. The dominant practice in this kind of research is to look for statistically significant subgraphs; this is achieved by comparing the actual interaction network to a collection of randomized networks, which are usually built using the stub-rewiring algorithm of [61]. However, there is certain asymmetry to the way the stub-rewiring method treats schemas of different complexities. In the case of pairwise schemas, it evaluates their significance given the lower-order properties of the network, which in the case of pairs include features of vertices such as label and degree. For schemas that are logically built up from pairs, such as triplets or triangles, the lower-order properties would include information about pairs as well, and for Ys, information about triplets. Therefore, we developed a randomization procedure for scoring these schemas that preserves these lower-order properties in random graphs. More importantly, this procedure allows us to look for true building blocks of the network, since a building block should not be reducible to the lower-order building blocks, but should contain new information.

Many schemas we discover recapitulate known biology and include, for example, well-known components of signaling pathways, such as many "star"-shaped triplet or Y schemas that are built up of protein kinase and signaling domains, reflecting the importance of phosphorylation; or components of known complexes, such as the spliceosome (reflected by the LSM triangle), or the SNARE vesicle-fusion machinery. Furthermore, a more systematic analysis of the schemas we uncovered has shown that they are enriched for biological process.

3.4.2 Pairwise schemas across genomes

The existence of several genomes that have been covered to a large extent by high-throughput physical interaction experiments allowed us to begin to take a cross-

genomic approach to the study of network schemas. We focused on baker's yeast and human as the two orgnaisms for which a large amount of physical interaction data exists, and which in a way represent the two ends of the eukaryotic complexity spectrum: from a unicellular yeast to the complex multicellular human.

The nearly 30 pairwise schemas that are found to be significant and recurring in both yeast and human may represent the "core" pairwise schemas that provide the building blocks of core processes in the cell. Indeed, these schemas represent some of the basic processes that happen within the cell: signaling, including the regulation of Ras family GTPases, vesicle fusion, cyclin regulation of kinases, ubiquitination, and so on.

The second immediately obvious feature of the human pairs is that they reflect the much richer landscape of biological functions that the human has and the novel molecular "tools" that human has compared to yeast. Thus, a large number of human pairs involve elements of phosphotyrosine pathways, such as the tyrosine kinase motif (PF07714), tyrosine phosphatase motif (PF00102) and the SH2 (PF00017) domain that binds phosphotyrosines.

3.4.3 Schemas for domain and protein annotation

Using two poorly understood gene families, one from human, one from yeast, we show how schema analysis can be used to annotate protein families and their individual members. One may wonder what schema analysis gives us that we would not be able to get from simple interaction analysis such as guilt-by-association. For example, if we consider the 41 interaction partners of the 9 of the MAGE proteins that have interactions, and use the GO Term Finder [5], we see that the only significant common biological process among them is apoptosis, whereas 21 of the 41 proteins have no biological process annotation. Using this information, we might hypothesize that

MAGE proteins are involved in apoptosis, although the signal for this implication may be considered fairly weak. Schema analysis, on the other hand, acts as a lens that focuses the investigator's attention on patterns of interaction that are statistically significant. By "pulling out" interactions that are likely to be meaningful from the total wealth of interaciton data, it allows him or her to concentrate on promising directions of exploration.

3.4.4 Interpretation of schemas

Here, we discuss the schemas from the point of view of their role in network organization. Although each schema is in a way unique, due to the fact that schema definition includes protein annotations, we can begin to group them into broader categories that correspond to different network organizational principles. The most basic level of schema organization is represented by the pairs. Since our approach does not limit itself to search for domain-domain interactions, the pairs we find include both putative domain-domain interactions and more abstract patterns of organization. The pairwise schemas include both homotypic and heterotypic interactions. In addition, the meaning of a pair can be of several types. The pair may either represent two proteins working together (such as, but not necessarily, forming a dimer or part of a complex), or one protein acting on the other, e.g., activating or deactivating it. The Ras-RhoGAP pair is an example of the latter kind (RhoGAP regulates Ras).

Among the higher order topologies that we considered, we have triangles and "star schemas"; the latter group includes triplets ("2-stars") and Ys ("3-stars"). The triangles are most intuitively associated with protein complexes, whereas the "star topologies" can be more easily associated with "flow of information". In this case, the spokes of the star may represent the "flow of information" in the pathway; the "flow of information" along each spoke may have any direction: either one or more

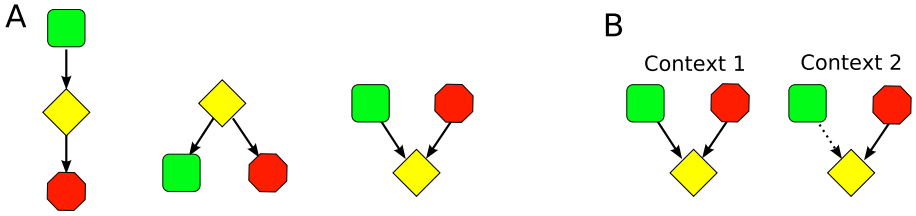

Figure 3.15: Possible flow of information in "star" schemas. Any arrow may also be replaced by an undirected edge (representing a protien dimer w/o associated flow of information).

of the spoke acting on the central node, which in turn acts upon the remaining spoke(s), or all of the spokes acting upon the central protein, or the central protein acting upon all of the spokes (Figure 3.15, A). In addition, the "star" schemas may represent "switch-like" patterns, in which some of the spokes of the "star" are active in different contexts (Figure 3.15, B). This is especially likely to be the case if the spokes represent the same motif.

We can see instances of these types of schemas turning to our running example of Ras proteins. The Pkinase-Ras-RhoGAP triplet is an example of a schema that represents linear flow of information. Here, the Ras family protein (either CDC42 or Rho1) is regulated by a RhoGAP, and in turn regulates its effector kinase. Thus, this schema represents linear information flow from the RhoGAP protein to the kinase. The Ras-Yip1-Ras triplet, on the other hand, may represent a "star" schema in which the central protein "acts upon" the spokes, since the Yip1 protein is believed to recruit the Ras proteins to the membrane.

This triplet probably represents "switch-like" behavior. Since the two Yip1 proteins have reported interactions with 8-9 Ras proteins, it is likely that these interactions take place at different times, depending on the context or simply on chance. It is worthwhile to remember, though, that since the significance of this schemas was established using pairwise-count-preserving randomizations, the "switch-like" nature of the interaction between the Yip1 proteins and their Ras-like partners is itself significant.

Our decision to stop the study at 3-edge schemas was pragmatic; schemas of higher order may very well exist. Moreover, by looking at the schemas we find, one can see that many of them overlap "horizontally", that is, triplets or triangles share common pairwise subschemas, and Ys share common triplet subschemas (see Figures 3.3, 3.5). This may represent larger, "mezoscale" graph substructures of which schemas are building blocks. Moreover, even a single schema may in fact be a part of a larger structure. However, the significance of these structures can only be determined by doing the significance analysis using randomizations conditioned on lower-order graphs, and there may be higher-order significant schemas whose subgraphs are not by themselves significant schemas (in the same way that significant triplet schemas do not have to be made up of significant pairwise schemas).

3.4.5 Semantics and context of interactions

A noticeable feature of this analysis is that the underlying data treats all interactions as being the same. In reality, the interactions have both meaning and contextual information. As we point out in the analysis of "star schemas", the same schema topology may in fact represent several underlying "biological reality" schemas, some of which have directed edges (e.g., representing (de)activation of one of the interactors by the other), with corresponding temporal information, and some of which present

a combination of multiple subschemas which are active at different times or in different cellular contexts. This information becomes especially important when studying multicellular organisms, in which different interactions may take place in different cell types altogether. In this respect, the schema finding approach has two alternative or complementary potential directions. As more specific interaction data is gathered, such as that which comes from high-throughput phosphorylation experiments [43], it will become possible to assign direction to more and more of the potentially directed edges in the interaction networks. Then, when enough such data exist, we could look for schemas that include edge direction as part of their topology. Similarly, if contextual information for a large number of interactions becomes known and systematized, we can look for schemas either within each context separately, or include contextual information as part of the schema definition. Alternatively, we could attempt to extract contextual information ourselves, focusing on the individual undirected schemas that we presently find, and devising computational means for predicting such information based, for example, on expression information or literature search.

Finally, the utility of schema analysis is intimatey connected to the data that is used as source of interactions and labels. In order to get a more complete biological picture, it is advisable to use several complementary systems of protein labels; the schemas obtained using different classes of protein features as labels would display the multidimemsnional problem of protein function from different angles or perspectives.

Chapter 4

Conclusion

Protein interaction networks hold promise in shedding light on one of the most important problems in biology—understanding what proteins do and how they work together to accomplish their tasks. At the same time, the size and complexity of protein interaction networks, especially those that are based on high-throughput experiments, challenge the computational biologist to understand the organizational principles governing the networks. In this thesis, we addressed both of these problems. We discussed some aspects of the relationship between interaction topology and function in physical protein-protein interaction networks. In Chapter 2, we focused on the problem of predicting biological process for unannotated proteins using physical interaction networks. Network-based function prediction has been dominated by the assumption of guilt by association, and we built upon this principle. We began by discussing what features of the interaction network need to be taken into account by a function prediction algorithm, using as illustration several existing methods that address the same problem. We then introduced a novel network-based function prediction algorithm which outperforms those methods thanks to its use of network connectivity and distance.

Whereas Chapter 2 used interaction networks as a tool for function prediction, in Chapter 3, we made the object of study the interplay between the interaction network and individual protein features. We proposed a novel framework for the study of organizational principles of interaction networks, by introducing the concept of pathway schemas to describe patterns of interaction between different types of proteins. We introduced a statistical framework for finding overrepresented schemas that act as network building blocks; this framework relies on comparing the abundance of a schema in the real interaction network as compared to a collection of random networks. By using a randomization procedure that preserves lower-order schema statistics, our methodology builds a collection of schemas such that the more complex ones contain novel information relative to the lower-order ones. Coming back to the problem of predicting biological processes, we demonstrate the use of schema analysis for predicting the function of uncharacterized proteins and protein families.

In this work, we were guided by the assumption that different interpretations of protein function—the more abstract and high-level biological processes versus the more concrete biochemical properties or molecular functions—require different approaches when studied in the context of protein interactions. When the biological process view of protein function is taken, the basic assumption is guilt by association, which predicts that interacting proteins tend to belong to the same process, whereas when dealing with the molecular features of proteins, it is meaningful to ask about patterns of integration between proteins of different types. Of course, like any simplifying assumptions, the guilt by association is not without exceptions. If the guilt by association assumption were unequivocally accurate, either the interactome would break up into separate disconnected components—which is certainly not the case–or the vast majority of proteins, which form the largest connected component of the interaction network, would participate in the same biological process. The latter is in

fact true in a certain way, since all proteins do participate in one process that is life of the cell, but this view is obviously too trivial to be meaningful. Therefore, it may prove useful to look beyond the "guilt by association" assumption and examine the interplay between proteins of different biological processes. Recent work in network-based prediction of protein function has begun to utilize inter-protein correlations between functional terms [51]. Our schema-finding methodology is directly applicable to uncovering over-represented pairs (and higher order topologies) of biological processes that are found to annotate interacting proteins. In future work, these uncovered inter-process schemas may prove useful in moving beyond guilt-by-association approaches for network-based protein function annotation.

Another area in which we see further work is addressing the dynamics of protein interactions. So far, we have taken a static view of the interactome; this is in part due to the nature of the interaction data that is currently available. We see two possible directions to pursue with respect to the dynamic view of the network. One is to perform further analysis on the current results. For example, we can focus on the schemas that we have find, and analyze them from the perspective of possible network dynamics. This is likely to be particularly useful for understanding the meaning of "star" (linear triplet and Y-shaped) schemas, and preliminary analysis has revealed pairs of interactions that are unlikely to be present at the same time and place in the cell (i.e., putative "switches" in the interactome). An alternative approach would be first to obtain a dynamic view of the interaction network and then extend our methods to to take this view into account by developing a dynamic-network version of function-prediction and schema-finding algorithms. Of course, obtaining such a dynamic view is a problem in itself; since the current state of experimental data does not readily permit dynamic interpretation of the majority of interactions, this problem will need to be addressed computationally.

In this thesis, we have taken two somewhat orthogonal approaches for considering protein function in the context of protein interaction networks. However, since biological processes are accomplished via the interplay of proteins of different molecular functions, the problem of predicting the biological processes of proteins is intimately linked to understanding how proteins with different biochemical roles interact with each other. This thesis has taken a first step towards providing computational methodology that helps unify and exploit these two differing but related views of protein function.

Bibliography

[1] Almond, J., , and Cohen, G. M. The proteasome: a novel target for cancer chemotherapy. *Leukemia 16* (2002), 433–443.

[2] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet. 25*, 1 (2000), 25–29.

[3] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. Reverse engineering of regulatory networks in human b cells. *Nature Genetics 37* (2005), 382–390.

[4] Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., and Griffiths-Jones, S. The Pfam protein familes database. *Nucleic Acids Res. 32* (2004), D138–D141.

[5] Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics 20*, 18 (2004), 3710–3715.

[6] Breitkreutz, B., Stark, C., and Tyers, M. Osprey: a network visualization system. *Genome Biol. 4* (2003), R22.

[7] Breitkreutz, B. J., Stark, C., and Tyers, M. The GRID: The general repository for interaction datasets. *Genome Biol. 4* (2003), R23.

[8] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Gunoche, A., and Jacq, B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol. 5* (2004), R6.

[9] Chien, C., Bartel, P., Sternglanz, R., and Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA 88* (1991), 9578–9582.

[10] Chomez, P., De Backer, O., Bertrand, M., De Plaen, E., Boon, T., and Lucas, S. An Overview of the MAGE Gene Family with the Identification of All Human Members of the Family. *Cancer Res 61*, 14 (2001), 5544–5551.

[11] Cormen, T. H., Leiserson, C. E., and Rivest, R. L. *Introduction to Algorithms*. MIT Press/McGraw-Hill, 1990.

[12] Dalhaus, E., Johnson, D. S., Papadimitriou, C., Seymour, P., and Yannakakis, M. The complexity of the multiway cuts. In *Proc. 24th Annual STOC* (1992), ACM, pp. 241–251.

[13] Dandekar, T., Snel, B., Huynen, M., and Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci. 23*, 9 (1998), 324–328.

[14] de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. Dynamic complex formation during the yeast cell cycle. *Science 307* (2005), 724–727.

[15] Deng, M., Mehta, S., Sun, F., and Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res. 12* (2002), 1540–1548.

[16] Deng, M., Sun, F., and Chen, T. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Pac. Symp. Biocomput.* (2003), pp. 140–151.

[17] Deng, M., Tu, Z., Sun, F., and Chen, T. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics 20*, 6 (2004), 895–902.

[18] Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. Prediction of protein function using protein-protein interaction data. *csb 00* (2002), 197.

[19] Despons, L., Wirth, B., Louis, V., Potier, S., and Souciet, J.-L. An evolutionary scenario for one of the largest yeast gene families. *Trends in Genetics 22* (2006), 10–15.

[20] Dho, S. H., and Kwon, K.-S. The Ret Finger Protein Induces Apoptosis via Its RING Finger-B Box-Coiled-coil Motif. *J. Biol. Chem. 278*, 34 (2003), 31902–31908.

[21] Edgar, R., Domrachev, M., and Lash, A. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucl. Acids. Res. 30* (2002), 207–210.

[22] Eisen, M., Spellman, P., Brown, P., and D, D. B. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 25 (1998), 14863–14868.

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

[24] Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H., and Batzoglou, S. Graemlin: General and robust alignment of multiple large interaction networks. *Genome Res.* (2006), gr.5235706.

[25] Fong, J., Keating, A. E., and Singh, M. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biology 5*, 2 (2004), R11.

[26] Fourer, R., Gay, D. M., and Kernighan, B. W. *AMPL: A Modeling Language for Mathematical Programming.* Brooks/Cole Publishing Company, Pacific Grove, CA, 2002.

[27] Fromont-Racine, M., Rain, J., and Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genetics 16* (1997), 277–282.

[28] Gaasterland, T., and Ragan, M. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics 3* (1998), 177–192.

[29] Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415*, 6868 (2002), 141–147.

[30] Giot, L., Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., et al. A protein interaction map of Drosophila melanogaster. *Science 302* (2003), 1727–1736.

[31] Gomez, S., Lo, S.-H., and Rzhetsky, A. Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics 159* (2001), 1291–1298.

[32] Guimaraes, K., Jothi, R., Zotenko, E., and Przytycka, T. Predicting domain-domain interactions using a parsimony approach. *Genome Biology 7* (2006), R104.

[33] Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D., Berriz, G., Zhang, L., et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature 430* (2004), 88–93.

[34] Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., et al. Transcriptional regulatory code of a eukaryotic genome. *Nature 431* (2004), 99–104.

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

[36] Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S., et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature 415*, 6868 (2002), 180–183.

[37] Hong, E., Balakrishnan, R., Christie, K., Costanzo, M., Engel, S. D. S., et al. Saccharomyces genome database. http://www.yeastgenome.org.

[38] Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics 22* (2006), 2890–2897.

[39] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics 18 Suppl 1* (2002), S233–S240.

[40] ILOG CPLEX 7.1, 2000. http://www.ilog.com/products/cplex/.

[41] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA 19* (2001), 4569–4574.

[42] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA 97* (2000), 1143–1147.

[43] J., P., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., et al. Global analysis of protein phosphorylation in yeast. *Nature 438* (2005), 679–684.

[44] Jansen, R. H., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A., Snyder, M., Greenblatt, J., and M.Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science 302* (2003), 449–453.

[45] Jenssen, T.-K., Lgreid, A., Komorowski, J., and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics 28* (2001), 21–28.

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

[47] Johnsson, N., and Varshavsky, A. Split ubiquitin as a sensor of protein interactions in vivo. *Proc. Natl. Acad. Sci. USA 91* (1994), 10340–10344.

[48] Karaoz, U., Murali, T. M., Levotsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA 101* (2004), 2888–2893.

[49] Kelley, B., Sharan, R., Karp, R., Sittler, T., Root, D., Stockwell, B., and Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA 100* (2003), 11394–11399.

[50] King, A., Przulj, N., and Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics 20* (2004), 3013–3020.

[51] Kirac, M., Ozsoyoglu, G., and Yang, J. Annotating proteins by mining protein interaction networks. *Bioinformatics 22*, 14 (2006), e260–e270.

[52] Koyuturk, M., Kim, Y., Subramaniam, S., Szpankowski, W., and Grama, A. Detecting conserved interaction patterns in biological networks. *Journal of Computational Biology 13*, 7 (2006), 1299–1322.

[53] Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics 20*, 16 (2004), 2626–2635.

[54] Lee, I., Date, S., Adai, A., and Marcotte, E. A probabilistic functional network of yeast genes. *Science 306*, 5701 (2004), 1555–1558.

[55] Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., et al. Transcriptional regulatory networks in Saccaromyces cerevisiae. *Science 298* (2002), 799–804.

[56] Letovsky, S., and Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics 19 Suppl 1* (2003), I197–I204.

[57] Li, S., Armstrong, C., Bertin, N., Ge, H., Milstein, S., Boxem, M., et al. A map of the interactome network of the metazoan C. elegans. *Science 303*, 5657 (2004), 540–543.

[58] Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., and Gerstein, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature 431* (2004), 308–312.

[59] Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., and Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science 285* (1999), 751–753.

[60] Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature 402* (1999), 83–86.

[61] Maslov, S., and Sneppen, K. Specificity and Stability in Topology of Protein Networks. *Science 296*, 5569 (2002), 910–913.

[62] Meroni, G., and Diez-Roux, G. Trim/rbcc, a novel class of 'single protein ring finger' e3 ubiquitin ligases. *Bioessays 27* (2005), 1147–1157.

[63] Mewes, H., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res. 20* (2002), 31–34.

[64] Miller, J., Lo, R. S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W. S., and Fields, S. Large-scale identification of yeast integral membrane protein interactions. *Proc. Natl. Acad. Sci. USA 102* (2005), 12123–12128.

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

[66] Monte, M., Simonatto, M., Peche, L. Y., Bublik, D. R., Gobessi, S., Pierotti, M. A., Rodolfo, M., and Schneider, C. MAGE-A tumor antigens target p53 transactivation function through histone deacetylase recruitment and confer resistance to chemotherapeutic agents. *PNAS 103*, 30 (2006), 11160–11165.

[67] Myers, C., D., Robson, Wible, A., M., Hibbs, C., Chiriac, Theesfeld, C., Dolinski, K., and Troyanskaya, O. Discovery of biological networks from diverse functional genomic data. *Genome Biology 6*, 13 (2005), R114.

[68] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps, 2005.

[69] Orlowski, R. Z. The role of the ubiquitin-proteasome pathway in apoptosis. *Cell Death and Differentiation 6* (1999), 303–313.

[70] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., and Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA 96*, 6 (1999), 2896–2901.

[71] Palla, G., Dernyi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435* (2005), 814–818.

[72] Pawson, T., and Nash, P. Assembly of cell regulatory systems through protein interactions. *Science 300* (2003), 445–452.

[73] Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA 96*, 8 (1999), 4285–4288.

[74] Pereira-Leal, J., Enright, A., and Ouzounis, C. Detection of functional modules from protein interaction networks. *Proteins: Structure, Function and Bioinformatics 54* (2004), 49–57.

[75] Przulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics 23*, 2 (2007), e177–183.

[76] R., S., Suthram, S., Kelley, R., Kuhn, T., McCuine, S., Uetz, P., et al. Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA 102* (2005), 1974–1979.

[77] Rachlin, J., Cohen, D. D., Cantor, C., and Kasif, S. Biological context networks: a mosaic view of the interactome. *Mol. Syst. Biol. 2* (2005), 66.

[78] Rain, J., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., et al. The protein-protein interaction map of Helicobacter pylori. *Nature 409*, 6817 (2001), 211–215.

[79] Ren, B., Robert, F., Wyrick, J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. Genome-wide location and function of dna binding proteins. *Science 290* (2000), 2306–2309.

[80] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Sraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology 17* (1999), 1030–1032.

[81] Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol. 6* (2005), R89.

[82] Rives, A., and Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA 100*, 3 (2003), 1128–1133.

[83] Salehi, A., Roux, P. P., Kubu, C. J., Zeindler, C., Bhakar, A., Tannis, L.-L., Verdi, J. M., and Barker, P. A. Nrage, a novel mage protein,

interacts with the p75 neurotrophin receptor and facilitates nerve growth factordependent apoptosis. *Neuron 27* (2000), 279–288.

[84] Sandmann, T., Herrmann, J. M., Dengjel, J., Schwarz, H., and Spang, A. Suppression of Coatomer Mutants by a New Protein Family with COPI and COPII Binding Motifs in Saccharomyces cerevisiae. *Mol. Biol. Cell 14*, 8 (2003), 3097–3113.

[85] Schlitt, T., Palin, K., Rung, J., Dietmann, S., Lappe, M., Ukkonen, E., and Brazma, A. From gene networks to gene function. *Genome Res. 13* (2003), 2568–2576.

[86] Scholtens, D., Vidal, M., and Gentleman, R. Local modeling of global interactome networks. *Bioinformatics 21* (2005), 3548–3557.

[87] Schwikowski, B., Uetz, P., and Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol. 18* (2000), 1257–1261.

[88] Segal, E., Wang, H., and Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics 19 Suppl 1* (2003), 264–270.

[89] Sharan, R., Ulitsky, I., and Shamir, R. Network-based prediction of protein function. *Mol Syst Biol. 3* (2007), 88.

[90] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. Network motifs in the transcriptional regulation network of E. coli. *Nat. Genet. 31* (2002), 64–68.

[91] Shoemaker, B., and Panchenko, A. R. Deciphering proteinprotein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLOS Computational Biology 3* (2007), e43.

[92] Sivars, U., Aivazian1, D., and Pfeffer1, S. R. Yip3 catalyses the dissociation of endosomal rabgdi complexes. *Nature 425* (2003), 856–859.

[93] Smith, G. R., Givan, S. A., Cullen, P., and Sprague, George F., J. GTPase-Activating Proteins for Cdc42. *Eukaryotic Cell 1*, 3 (2002), 469–480.

[94] Spirin, V., and Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA. 100* (2003), 12123–12128.

[95] Sprinzak, E., and Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol. 311* (2001), 681–692.

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

[97] Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res. 34* (2006), D535–D539.

[98] Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. Automated modeling of signal transduction networks. *BMC Bioinformatics 3* (2002), 34.

[99] Strong, M., Graeber, T., Beeby, M., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchial clustering of genome-wide functional linkage maps. *Nucleic Acids Res 31* (2003), 7099–7109.

[100] Sun, L., and Chen, Z. J. The novel functions of ubiquitination in signaling. *Current Opinion in Cell Biology 16* (2004), 119–126.

[101] Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA 101* (2004), 2981–2986.

[102] Taniura, H., Matsumoto, K., and Yoshikawa, K. Physical and Functional Interactions of Neuronal Growth Suppressor Necdin with p53. *J. Biol. Chem. 274*, 23 (1999), 16242–16248.

[103] Tong, A., Evangelista, M., Parsons, A., Xu, H., Bader, G., Page, N., et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science 294* (2001), 2364–2368.

[104] Tong, A., Lesage, G., Bader, G., Ding, H., Xu, H., Xin, X., et al. Global mapping of the yeast genetic interaction network. *Science 303* (2004), 808–813.

[105] Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., and Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). *Proc. Natl. Acad. Sci. USA 100* (2003), 8348–8353.

[106] Tsuda, K., and Noble, W. Learning kernels from biological networks by maximizing entropy. *Bioinformatics 20 Suppl. 1* (2004), I326–I333.

[107] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., et al. A comprehensive analysis of protein-protein interactions in S. cerevisiae. *Nature 403* (2000), 623–627.

[108] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol. 21* (2003), 697–700.

[109] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res. 31* (2003), 258–261.

[110] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature 417* (2002), 399–403.

[111] von Mering, C., Zdobnov, E., Tsoka, S., Ciccarellli, F., Pereira-Leal, J., Ouzounis, C., and Bork, P. Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. USA 100*, 26 (2003), 15428–15433.

[112] Wen, C.-J., Xue, B., Qin, W.-X., Yu, M., Zhang, M.-Y., Zhao, D.-H., Gao, X., Gu, J.-R., and Li, C.-J. hnrage, a human neurotrophin receptor interacting mage homologue, regulates p53 transcriptional activity and inhibits cell proliferation. *FEBS Letters 564* (2004), 171–176.

[113] Wojcik, J., and Schacter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics 17 Suppl 1* (2001), S296–S305.

[114] Yamanishi, Y., Vert, J. P., and Kanehisa, M. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics 20* (2004), I363–I370.

[115] Yang, X., Matern, H. T., and Gallwitz, D. Specific binding to a novel and essential golgi membrane protein (yip1p) functionally links the transport gtpases ypt1p and ypt31p. *The EMBO Journal 17* (1998), 4954–4963.

[116] Yeger-Lotem, E., Sattath, S., Kashtan, N., Izkovitz, S., Milo, R., Alon, U., and Margalit, H. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc. Natl. Acad. Sci. USA 101*, 16 (2004), 5934–5939.

[117] Yu, H., Zhu, X., Greenbaum, D., Karro, J., and Gerstein, M. TopNet: a tool for comparing biological subnetworks, correlating protein properties with topological statistics. *Nucleic Acids Research 32* (2004), 328–337.

[118] Zhang, L., King, O., Wong, S., Goldberg, D., Tong, A., Lesage, G., Andrews, B., et al. Motifs, themes and thematic maps of an integrated saccharomyces cerevisiae interaction network. *J. Biol 4* (2005), 6.