Network-based analysis of protein function

Jimin Song

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
Adviser: Mona Singh

January 2012

Abstract

Large-scale protein-protein interaction networks have been determined for organisms across the evolutionary spectrum. The resulting interactomes are a great resource for furthering our understanding of cellular functioning, pathways and organization. In this thesis, we focus on uncovering the relationship between the topological characteristics of these networks and their underlying functioning.

In the first part of this thesis, we study the problem of network modularity. Cellular networks are known to have modular organization, with groups of proteins working together to perform some larger biological process. Numerous clustering approaches have been applied in order to uncover, from large-scale protein physical interaction data, protein complexes and functional modules. We develop a comprehensive framework to assess how well network clustering approaches perform in uncovering protein complexes and biological processes, and in predicting protein functions. By applying this framework, we find that topological characteristics of networks are a significant factor in the accuracy trade-offs between local and global (i.e. clustering) approaches for uncovering cellular functioning.

In the second half of this thesis, we focus on relating one important aspect of protein functioning, its essentiality, to network topology. A protein is essential if it is vital for a cell's survival and its removal kills the cell. Previously, researchers had observed that essential proteins tend to have many physical interactions. We find that the relationship between essentiality and interaction degree is true at different scales of organization. In particular, we find that the number of intra-complex or intra-process interactions that a protein has is a better indicator of its essentiality than its overall number of interactions. Moreover, we find that within an essential complex, its essential proteins tend to have more interactions, especially intra-complex interactions, than its non-essential proteins. Finally, we build a module-level interaction network, and find that essential complexes and processes tend to have higher interac-

tion degrees in this network than non-essential complexes and processes; that is, they tend to exhibit a larger amount of functional cross-talk than non-essential complexes and processes.

Acknowledgements

I am greatly indebted to many people who have helped me complete my graduate studies. I would not be where I am without them. First and foremost, I would like to thank my advisor Mona Singh. She was a great mentor throughout many years. I am grateful for her constant support, encouragement, advice, and ideas.

I would like to thank my committee members, Tom Funkhouser, Olga Troyanskaya, as readers, and Bernard Chazelle, Andrea LaPaugh, as non-readers, for taking the time to serve on my committee.

I would like to thank the Singh group members, past and present, Eric Banks, Tony Capra, Jesse Farnham, Dario Ghersi, Peng Jiang, Zia Khan, Elena Nabieva, Alex Ochoa, Anton Persikov, Yuri Pritykin, Tao Yue, and Elena Zaslavsky for providing feedback on this work.

My sincere thanks also go to my friends and All Nations Mission Church. They have helped me get through tough times and made my life at Princeton enriched and enjoyable. I cannot forget memorable times spent with them.

I would like to thank my family: my parents, my sister Hanna, and my twin brothers Wonho and Junho. I am always grateful for their constant love and support throughout my life.

I was supported by Princeton University, Samsung Foundation of Culture (Samsung Scholarship), and the following grants: National Institute of Health (NIH GM076275) and National Science Foundation (NSF ABI-0850063).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Virtually all biological processes are accomplished via numerous specific interactions amongst various types of molecules (e.g., proteins, DNAs, RNAs, and small molecules). In the past decade, high-throughput experimental techniques have determined large scale interaction data between molecules in the cell. This data holds great promise in helping to unravel cellular organization and functioning, and in gaining a better understanding of protein function. In this thesis, we develop and apply algorithms for analyzing cellular networks.

Broadly speaking, a cellular network can be modeled as a graph where nodes are proteins and edges are either undirected or directed interactions between proteins [3, 2, 95, 84]. There are several types of interactions which have been determined in a high-throughput manner. Here, we briefly review different types of networks. First, a protein-protein physical interaction corresponds to direct physical binding between proteins. Second, a transcription factor binding interaction between two proteins corresponds to the case where one of the proteins, a transcription factor, binds to DNA to regulate gene expression. In this case, there is a directed interaction from the transcription factor protein to the protein product of the corresponding gene. Third, a phosphorylation interaction is a directed protein-protein interaction where

a kinase protein adds phosphate chemical groups to a substrate protein, thereby leading a change in its functional state. Fourth, a genetic interaction occurs between two genes if they are functionally related; that is, the growth rate of the organism with a deletion or mutation of these two genes is much different from what we expect, based on the growth rates of two single mutants. In the extreme case where the double mutation kills the cell while each single mutation does not, the interaction is a "synthetic lethal." While there are many other types of molecular interactions in the cell, these four interaction types represent the bulk of known network data to date for most organisms [81], and much existing network analysis work has focused on them.

1.1 Experimental techniques for determining protein-protein physical interactions

In this thesis, we primarily focus on protein-protein physical interaction networks among various types of cellular networks. We now briefly describe some of the experimental techniques utilized to determine these networks, as the interactions determined by specific techniques have different interpretations.

There are mainly two types of physical interactions that have been determined at the large scale: direct, binary interactions between two proteins or indirect interactions indicating co-membership of proteins in complexes [92]. The yeast two-hybrid (Y2H) was invented in the late 1980s [28] and since then a series of large scale protein-protein interaction data have been determined across several organisms, including yeast, fly and human [83, 43, 42, 92, 35, 80, 29, 82, 70]. Y2H utilizes a transcription factor (TF) protein GAL4 that consists of a binding domain and an activating domain. The binding domain localizes the TF protein to an upstream activating sequence of a reporter gene. The activating domain helps the TF protein to

activate transcription of a reporter gene. To exploit this modular structure, the transcription factor is first divided into two parts, where one includes the binding domain and the other includes the activating domain. Expression of the reporter gene cannot proceed without both the binding and activating domain. In order to detect physical binding between two proteins, one protein is fused to the binding domain and the other protein is fused to the activating domain. If the two proteins interact with each other, the two binding and activating domains are brought together, and the function of the TF protein is restored. Thus, when the two proteins bind, a reporter gene is successfully transcribed to mRNA, and this can be detected.

Tandem affinity purification-mass spectrometry (TAP-MS) is used to detect protein complexes by identifying a target protein and its interacting proteins [68]. First, a TAP (Tandem affinity purification) tag is fused to a target protein. Since the TAP tag includes an IgG binding unit and calmodulin-binding peptide, two consecutive affinity steps can detect the target protein and its interacting proteins by their binding to IgG matrix and calmodulin. The target protein and its interacting proteins are then identified by mass spectrometry. Using the TAP-MS technique, a series of high-throughput interaction data were released [41, 32, 31, 49, 21, 27], primarily for human and yeast.

It has been argued that Y2H has a higher rate of false positive interactions than TAP-MS [85, 9], however, Yu et al [92] produced better quality interaction data than before and suggested that the quality of Y2H data is not necessarily worse than that of TAP-MS, rather that Y2H uncovers interactions complementary to TAP-MS data. That is, the Y2H approach uncovers transient, binary interactions, whereas the TAP-MS approach uncovers both indirect and direct interactions. In this thesis, due to different weaknesses of each experimental technique, we do not rely on just one type of data. Rather, interactions can have more confidence when detected by more than one experiment [91]. Thus, we utilize several types of protein interaction networks

and all of our analysis is repeated for each network. In this way, we show that our findings are independent of the specific experimental techniques utilized.

1.2 Analysis of cellular networks

Here, we review briefly previous computational approaches to analyze cellular networks. Some of this analysis has been done in one particular type of network (e.g., protein physical network or transcription factor regulatory network), while others have been done in integrated networks consisting of more than one interaction type. The methodologies are easily applicable across a range of networks.

Network topology

Early research characterized the overall topological features of networks, and attempted to relate network topology to cellular functioning. Cellular networks have been shown to be scale-free [45, 87, 11]. That is, there are a few high-degree proteins that interact with lots of other proteins, and a large number of proteins that interact with only a few proteins. Much additional work focused on finding interesting properties of high-degree or hub proteins. For example, it has been observed that in the yeast *S. Cerevisiae*, hub proteins tend to be essential for the survival of the cell. In other words, in optimal conditions, yeast cannot grow and multiply without any one of these essential proteins [44, 11]. In Chapter 3, we will discuss essentiality further and show that essential proteins are more likely to have many interactions with functionally related proteins than with any proteins, and that there is a relationship between essentiality and network topology at several scales of organizations.

Modularity of networks and network clustering

It has been proposed that cellular networks are organized into functional modules; this has been confirmed numerous times in several computational analyses [39, 79]. More specifically, modularity means that proteins tend to interact with other proteins in modular units in order to accomplish specific tasks in the cell. These modules can be treated as more or less as discrete entities–the proteins within them are densely interconnected but are more sparsely connected with the rest of the network. Computationally, this has led to much work on identifying modules from static protein interaction networks, and to relating these modules to function and complexes [79, 8, 18].

With the observation that cellular networks tend to be modular, one of the natural follow-up works is to cluster networks. A myriad of network clustering approaches have been developed for many different applications [14, 26, 18, 71, 8, 72, 86, 7, 48, 65, 67, 5, 25, 63, 75, 4, 19, 61, 53, 88, 51, 33, 60, 20, 46]. Broadly speaking, there are mainly two types of approaches–global and local. Global or top-down approaches divide a network into subnetworks in an iterative way. For biological networks, such approaches include all proteins for clustering and almost all proteins belong to one cluster. Typically most methods result in clusters that are not overlapping. This is a disadvantage in uncovering biological functional modules as these highly overlap each other in reality. On the other hand, local or bottom-up approaches typically find dense regions from the networks. For biological networks, they tend to leave many proteins unclustered and these unclustered proteins are usually not considered further. This is a disadvantage in clustering sparse networks since it is possible that only a small portion of these networks are considered. In Chapter 2, we will discuss how well computational clustering of cellular networks reveals cellular modules and organization.

Protein function prediction

Cellular networks have been used to predict unknown function for proteins. As a direct consequence of modularity, a protein's biological process can be effectively predicted from protein interaction data. This is an important problem, as even for a well-studied genome, we do not know what about 30% of the proteins do. In its simplest form, biological process prediction is based on local guilt-by-association approaches, where a protein's function is predicted by looking at the annotations of its neighbor proteins [73, 55]. Another common way of predicting protein function is to first cluster networks and then ask what functions are enriched among proteins within each cluster [76]. In Chapter 2, we will discuss when local vs. clustering approaches should be used for functional annotation.

1.3 Our contributions

In Chapter 2, we develop a framework to evaluate how well network clustering algorithms uncover functional modules and predict protein function. Clustering of protein-protein physical interaction networks is one of the most common approaches for predicting functional modules and protein functions but until our work there was not a rigorous and comprehensive framework for evaluating clusters utilizing known functional data as a gold standard. To evaluate clustering algorithms, previous works focused on internal measures (i.e., the quality of the clusters are judged without desired groups in mind), or focused on how well they recapitulate protein complexes [17]. By applying our framework, we re-examine when and how clustering approaches should be applied to physical interactomes, and parameterize performance based on how well annotated genomes are. We also establish specific guidelines by which novel clustering approaches for cellular networks should be justified and evaluated with respect to functional analysis.

In Chapter 3, we show that protein essentiality is correlated to network topology at different scales of organization. Numerous studies have confirmed the "centrality-lethality" rule that hub proteins in the *S. cerevisiae* physical interaction network are enriched in essential proteins, and the prevailing view has been that hubs tend to be essential due to their participation in essential complexes and processes. By considering proteins within the functional organization of the yeast interactome, our main finding is that the centrality-lethality rule is true not just at the protein level but also at the module level, with complexes and processes that are essential tending to interact with many functional groups.

Chapter 2

How and when should interactome-derived clusters be used to predict functional modules and protein function?

2.1 Introduction

Proteome-scale physical interaction data have become available for a large number of organisms, including human and most model organisms. Global analyses of the resulting protein interaction networks provide new opportunities for uncovering cellular organization and revealing protein functions and pathways. Beyond the basic characterization of these interaction networks with respect to their topological features (e.g., [11]), arguably the most widespread approach for analyzing biological networks is to cluster or partition them into subcomponents. Clustering of biological networks has revealed a modular organization [39], with highly connected groups of proteins taking part in the same biological process or protein complex [69, 79, 8, 64]. Indeed,

dozens of papers for analyzing protein interaction networks have focused on finding clusters within them and novel network clustering methods continue to be developed (e.g., [26, 18, 71, 72, 86, 7, 48, 65, 67, 5, 25, 1, 4, 19, 61, 53, 88, 60]).

Most frequently, computationally-derived clusters within physical interaction networks are used to uncover protein complexes and functional modules, as well as to predict protein function. Typically, a cluster is associated with a known complex or function by determining whether the number of proteins known to be part of the complex or annotated with the function is enriched, as judged by the hypergeometric distribution. Within a cluster, enriched functions, perhaps also required to annotate a suitable fraction of member proteins, can then be transferred to other member proteins. While these types of analysis are commonplace in interactomics, how effective are they for the tasks at hand?

In this Chapter, we focus on the task of utilizing network-derived clusters to uncover functional modules and predict protein functions. Evaluating how well clusters correspond to functional modules is a challenging task. Central to this is that while functional modules are commonly defined as groups of proteins that work together to accomplish a biological process, there is no widely accepted formal definition of a module; many have been proposed, though typically based on topological features of the network (e.g., [67]). We utilize an external measure—the Gene Ontology (GO) [6]—to derive functional modules. That is, for a GO biological process or cellular component functional term, the corresponding module contains all the proteins that are annotated with that term. Since GO relates functions in a hierarchical fashion, the next challenge for evaluating clusters is to deal with this hierarchy. At first glance it may appear that functions can be chosen at a particular resolution in the hierarchy. For example, it is possible to utilize the high-level GO "slim" functional terms, and then clusters can be evaluated in how well they recapitulate these terms, using sensitivity and positive predictive value measures, as introduced in an influ-

ential quantitative assessment of how well clustering approaches can uncover known protein complexes [17]. However, for evaluating functional modules, this approach has the weakness that a clustering that finds many small tightly connected clusters corresponding to very specific biological processes would be unfairly penalized.

Our main technical contribution is a series of measures that can be used to compare and evaluate network clustering algorithms with respect to how well they perform in uncovering known, potentially overlapping functional modules. We demonstrate the quality of our measures by using them on random networks, and on clusters derived from the annotations themselves (i.e., these two extremes represent the noisy vs. ideal scenarios). With this evaluation framework in hand, in order to make general conclusions about the efficacy of network clustering-based approaches, we experiment with six available clustering algorithms on four different high-throughput derived *S. cerevisiae* physical interaction networks. We find that clustering algorithms exhibit a wide range of performances in recapitulating functional modules, derived from either biological process or cellular component GO terms, even when run on the same network, and that the relative performance of clustering algorithms varies depending on the network at hand. In particular, we find that topological features of the network should guide algorithm choice. Given the vast differences we find in how well clustering algorithms recapitulate functional modules, this is an important practical consideration. As a by-product of our analysis, we can also make conclusions about individual clustering algorithms: overall, though there are some clustering approaches which clearly outperform others, there is no single network clustering approach that dominates the rest in all cases.

Since module finding in biological networks is often motivated by the task of function prediction, we also perform a comprehensive evaluation in this scenario. Surprisingly, we find that for *S. cerevisiae*, the common practice of annotating a protein with the over-represented biological process or cellular component terms in its cluster

is less accurate than simple guilt-by-association approaches based on considering just the annotations of direct interaction partners. This is true regardless of which underlying clustering approach is used. Additionally, as annotations are removed from the network, the relative performance of clustering-based function prediction improves in comparison to the simple scheme that just considers the annotations of interacting proteins. This suggests that clustering-based methods are most useful in networks obtained for genomes with fewer protein annotations.

In addition to characterizing the utility of network-derived clusters in uncovering functional modules and predicting protein functions, a major contribution of our work is a framework that can be used in the future for evaluating how well a new clustering approach performs for these tasks. Importantly, our testing suggests that while clustering of networks is often motivated by the goal of predicting protein function, if new clustering approaches are evaluated with respect to function prediction, it is important to demonstrate how much, or in which circumstances, improvement is obtained over guilt-by-association approaches. Overall, we hope that our testing framework as well as our findings about the utility of interactome-based clustering will inform future methodological advances in clustering biological networks.

2.2 Materials and methods

2.2.1 Interaction and functional module datasets

We use *S. cerevisiae* protein interaction data from BioGRID [81], release 2.0.20, and generate four different networks in order to analyze how the underlying characteristics of the networks affect the performance of clustering algorithms. The first network contains all *S. cerevisiae* genetic and physical interactions in BioGRID. The second network contains all physical interactions. The third network consists of high-throughput physical (HTP) interactions from large datasets [83, 43, 42, 41, 32, 31, 49],

in case the small-scale experiments in BioGRID overlap the protein complexes used for evaluation, and the last network consists of physical interactions derived via three large-scale experiments utilizing the yeast two-hybrid (Y2H) technique [83, 43, 42]. For each network, we filter the data to remove proteins which interact with more than 50 other proteins. Furthermore, self-interactions are ignored. The resulting four networks have different topological properties (Table 2.1), as judged by the average number of interactions per protein, and the average node clustering coefficient. While we utilize all of these networks in our analyses, in the main body of this Chapter, we focus on the third and fourth networks, which we will refer to as the HTP network and the Y2H network. The HTP network has 4,160 proteins with 11,928 interactions, and the Y2H network has 2,828 proteins with 3,170 interactions.

Network	#proteins	#interactions	Avg. #neighbors	Avg. NCC
All genetic and physical interactions	4516	17843	7.902	0.15
All physical interactions	4319	13692	6.34	0.152
High-throughput physical interactions	4160	11928	5.735	0.136
Y2H physical interactions	2828	3170	2.242	0.05

Table 2.1: **Topological features of the four different yeast protein interaction networks considered.** For each network we give: the number of proteins, the total number of interactions, the average number of interactions per protein in the network and the average node clustering coefficient (NCC). The NCC for a protein is defined as the number of interactions amongst its interacting proteins, normalized by the total number of possible interactions amongst them.

We derive our gold standard groups from MIPS complexes [56] and GO [6]. We utilize 220 *S. cerevisiae* protein complexes from MIPS; this is the same set as used in the study of [17]. For each network described above, we remove from consideration any complex that has two or fewer proteins in the network. This leaves between 107 and 133 protein complexes for each network. In GO, there are 1963 Biological Processes (BP) and 551 Cellular Components (CC) terms. We remove GO annotations with evidence codes IEA, RCA and IPI. For each network, we remove BP and CC terms that annotate more than 100 proteins or fewer than 3 proteins. This leaves from 954

to 1090 BP terms and from 324 to 357 CC terms for each *S. cerevisiae* network. For the HTP network, 66% of the proteins are annotated with one of these BP terms, and 41% are annotated with one of these CC terms. For the Y2H network, these numbers are 70% and 45% respectively. For each BP and CC term under consideration, we define a functional module consisting of the proteins in the organism annotated with it. This gives us sets of potentially nested functional modules that range in specificity and size.

We also use *Homo sapiens* protein interaction data from BioGRID, release 2.0.55 for further analysis. The human network consists of all physical interactions and we filter the data to remove proteins which interact with more than 50 other proteins. The network has 7148 proteins with 18236 interactions. In GO, there are 5186 BP and 793 CC terms. We remove BP and CC terms that annotate more than 200 proteins or fewer than 3 proteins. This leaves 2777 BP terms and 451 CC terms for this network. Among the 7149 proteins in the network, 3653 proteins are annotated with one of these BP terms and 2181 proteins with one of these CC terms.

2.2.2 Clustering algorithms

We consider six diverse network clustering algorithms: *NetworkBlast* [75], *CFinder* [1], *MCL* [26], *DPClus* [4], *Mcode* [8] and a spectral approach based on modularity [61], which we refer to as *SpectralMod*. Brief descriptions of the clustering algorithms along with parameter settings are given here.

Network Blast [75] is originally designed for comparison of multiple protein networks but can also be applied to cluster a single protein network. This approach scores a cluster based on a log likelihood ratio, where the likelihood of observed interactions given that the cluster is modeled as a clique is compared to the likelihood of observed interactions given that the cluster is modeled as a part of a degree-preserving randomized subgraph on the same set of nodes. Clusters are grown greedily, and are

limited to contain at most 15 proteins. The final output set of clusters are filtered by the program to remove those that are highly overlapping. We run *NetworkBlast* with parameters beta set to 0.9 and true factor set to 0.5.

Clique Finder (CFinder) [1] finds a set of k-clique percolation clusters, each of which consists of a maximal connected component of adjacent cliques of size k where two cliques are adjacent if they share $k-1$ nodes. We use *CFinder* to find k-clique percolation clusters for all $k \geq 3$ and filter clusters whose size is greater than 500.

Markov clustering (MCL) [26] is a global clustering approach based on modified random walks on networks. It converts the adjacency matrix of a network into a stochastic matrix, and then clusters by repeating two steps: expansion and inflation. Expansion squares the matrix, and corresponds to taking another step in a random walk. Inflation is a boosting step, where each entry in the matrix is raised to the r-th power ($r > 1$) and then renormalized; this point-wise exponentiation amplifies higher probability transitions. These two steps are repeated until there is no change on the matrix. Finally, blocks of non-zero elements in the resulting matrix are taken as clusters. We use the inflation factor 1.8, which was found to be the best parameter in a recent study [17].

Density-periphery based clustering (DPClus) [4] is a greedy approach that grows clusters based on adding nodes that are well connected to other nodes in the cluster and that maintain cluster density. Here, a cluster is grown so as to maintain the density of the cluster above a particular threshold, and to ensure that each vertex that is added to the cluster is connected to a large enough number of vertices already in the cluster (its "cluster property"). Once a cluster can no longer be expanded, it is removed from the network and the process is repeated to find other clusters. We use density threshold 0.5 and cluster property threshold 0.5.

Molecular Complex Detection (MCODE) [8] is one of the first approaches for clustering interactomes. It also greedily grows clusters from a seed node. *Mcode* weights

each node according to the density of its "k-core" neighbors (i.e., the density is computed using only proteins of degree $\geq k$). A highest weight node is selected as seed and its neighbor nodes are added based on their weights. *Mcode* has many parameters, and we set them as: "include loops" false; "degree cutoff" 2; "haircut" true; "fluff" false; "node score cutoff" 0.1; "k-core" 2; "max. depth" 3.

Modularity-based spectral clustering (SpectralMod) is a global procedure that iteratively cuts the network so that there are more than the expected number of edges within clusters [61]. It approaches typically aim to divide a graph into a set of clusters in such a way that the number of edges between clusters is minimized, while subject to additional constraints (e.g., a balanced cut). We use a recently introduced spectral clustering approach that assumes that the probability of having an edge between two nodes is proportional to the degrees of the nodes, and tries to find a cut so that there are more than the expected number of edges within clusters [61]. There are no parameters for *SpectralMod*.

For each of these algorithms except *SpectralMod*, we download the software made available by the authors. For *SpectralMod*, we use software obtained from the author. For a baseline comparison, we also include a trivial algorithm, *OneCluster*, which always outputs a single cluster that includes all proteins in the network.

2.2.3 Evaluation measures for clustering

We evaluate clustering algorithms by judging how well the clusters correspond to groups of proteins as specified by MIPS complexes or functional modules as derived from either GO BP or GO CC annotations. Throughout, we refer to the output of the clustering algorithms as "clusters" and the proteins comprising complexes or functional modules as "groups." Though cluster validation approaches are well-developed (e.g., see [37]), much of this work has focused on either internal measures (i.e., the quality of the clusters are judged without desired groups in mind) or exter-

nal measures where groups partition the data (i.e., the groups are non-overlapping). Since our groups are overlapping, these external measures are not directly applicable. For each of the three tasks we are considering (uncovering complexes, BP functional modules, and CC functional modules), we utilize several measures to ascertain 1) how well each cluster maps to a known group and 2) how well each group maps to a cluster. Depending on what we want to test, we utilize either one direction of these mappings (e.g., clusters to groups) or both directions. (See Figure 2.2.3.)

When we consider clustering in order to uncover protein complexes, there should be a one-to-one mapping of clusters and protein complexes. Thus, both directions of mappings are utilized. On the other hand, GO annotations are organized in a hierarchical fashion with respect to each other. So, even for a high-quality clustering where each cluster corresponds to a functional module, there may be functional modules to which no clusters correspond. Also, while proteins interacting with each other tend to have the same GO term and thus highly connected regions or clusters are likely to be enriched with GO terms, it may be less likely that all proteins with the same GO term are together in the same cluster. Therefore, in the case of functional modules, we evaluate a clustering only by mapping clusters to groups. It is important to note that when mapping a cluster to a GO term, each of the overlap measures we introduce below considers the total number of proteins annotated with that term (i.e., if that term annotates many proteins that are not part of the cluster, then the score for mapping that cluster to the term will be lower).

Overlap measures

We utilize three measures for evaluating clusters that are based on overlaps between clusters and known groups of proteins. Each measure gives a value in the range of 0 and 1, where higher numbers correspond to better overlaps. Before describing these measures, we give some preliminaries. Let M be the number of clusters given

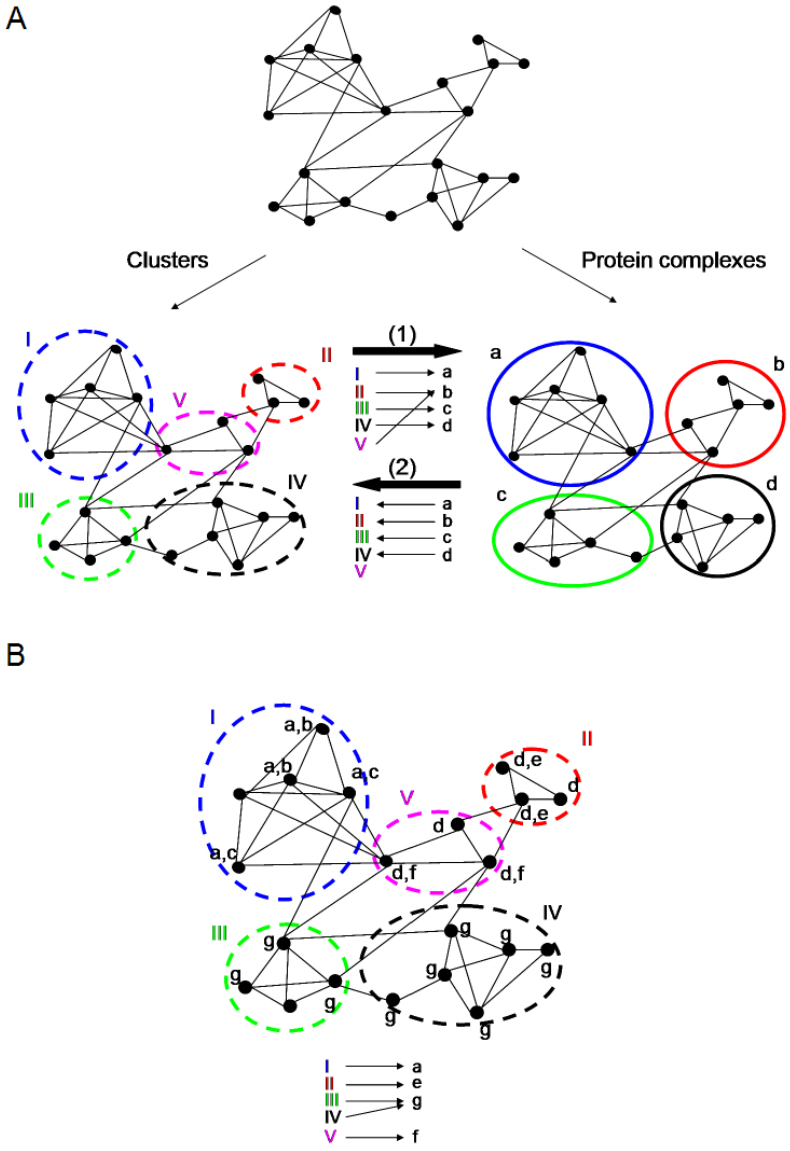


Figure 2.1: **Mapping between computationally-derived clusters and complexes and functional modules.** (A) A schematic network is shown with computationally-derived clusters on the left, and constituent protein complexes in the network on the right. Roman numerals refer to clusters and lowercase letters refer to protein complexes. The mapping from clusters to protein complexes (1) and from protein complexes to clusters (2) are given in the middle. (B) A schematic is shown with proteins annotated by GO terms of interest (in lowercase letters), and clusters outlined (and referred to by Roman numerals). A mapping of clusters to GO terms is also given. Note that both clusters II and V have proteins that are annotated with d, but map better to the more specific modules corresponding to annotations e and f respectively.

by a particular clustering, and N be the number of groups against which we are evaluating. Let C_j be the set of proteins within cluster j and let G_i be the set of proteins associated with the i-th group (e.g., in the i-th complex or annotated with i-th function). Our measures are as follows:

Jaccard measure. Given two sets, the Jaccard similarity coefficient is defined as the size of the intersection over the size of the union. For sets of proteins corresponding to cluster j and group i, let $Jac_{ij} = \frac{|G_i \cap C_j|}{|G_i \cup C_j|}$ denote their Jaccard coefficient.

PR measure. For sets of proteins corresponding to cluster j and group i, let $PR_{ij} = \frac{|G_i \cap C_j|}{|C_j|} \cdot \frac{|G_i \cap C_j|}{|G_i|}$ denote their precision-recall based score. The first part $\frac{|G_i \cap C_j|}{|C_j|}$ measures what fraction of the proteins in the cluster correspond to the grouping at hand (i.e., precision with respect to group i). The second part $\frac{|G_i \cap C_j|}{|G_i|}$ measures how much of group i is recovered by cluster j (recall). We note that our **PR**-based measures are similar to the F-measure (see, for example, [37]).

Semantic density measure. The density of a set of vertices in a network is typically defined as the number of edges among them divided by the maximum number of possible edges. We generalize this notion for protein interaction networks as follows to better recapitulate characteristics of the groups being compared to. For a set of proteins S, each protein $p \in S$ may be associated with labels $A(p) \subseteq \mathcal{A}$. For example, S can be a MIPS complex, $\mathcal{A}$ can be the set of clusters obtained after computational analysis of the entire interactome, and $A(p)$ gives which clusters p belongs to. Alternatively, S can be a cluster of proteins, and $\mathcal{A}$ can be the set of groups of proteins with a shared functional annotation; in this case $A(p)$ gives the groups p is part of. Then,

$$density(S, \mathcal{A}) = \frac{\sum_{\forall (p_1,p_2) \in S} W_{\mathcal{A}}(p_1, p_2)}{\sum_{\forall (p_1,p_2) \in S} (1)}$$

where $W_{\mathcal{A}}(p_1, p_2)$, defined next, is the weight given to a pair of proteins p_1, p_2 and is

in the range of 0 and 1. When considering clusters as $\mathcal{A}$, $W_{\mathcal{A}}(p_1, p_2) = 1$ if $A(p_1) \cap A(p_2) \neq \emptyset$, and 0 otherwise. This weight function is also used when considering MIPS complexes as $\mathcal{A}$. When GO derived functional groups are used as $\mathcal{A}$, the weight function is defined using a standard semantic similarity measure [52]. In particular, let $f(a)$ for functional group a be defined as the fraction of the total number of proteins in the considered network that have annotation a, and let $s(a) = -log(f(a))$ be a measure of how specific the annotation is. Then,

$$W_{\mathcal{A}}(p_1, p_2) = \frac{2 \cdot \max_{a \in A(p_1) \cap A(p_2)} s(a)}{\max_{a \in A(p_1)} s(a) + \max_{a \in A(p_2)} s(a)}$$

Mapping scores

Before describing our mapping scores, we briefly highlight some of our choices in computing these. First, some clustering approaches attempt to cluster all proteins (e.g., MCL and spectral clustering) whereas others leave many proteins unclustered. We chose to consider the unclustered proteins as singleton clusters, instead of ignoring them in the evaluation. Second, we remove from consideration all proteins in the complex and functional module groups that are not included in the network at hand. Third, when mapping a cluster to a group, we did not consider proteins that do not have any annotations from the grouping at hand. This means that clusters are filtered so as to remove any unannotated proteins, thereby potentially changing the size of the cluster. In practice, of course, clusters consisting of mostly unannotated proteins could be considered as putative protein complexes or functional modules in further analysis. Fourth, when mapping a group to a cluster (performed only in the MIPS analysis), all proteins in the clusters, including unannotated ones, are considered.

For each overlap measure described above, we utilize three "scores." First, we define scores for a clustering that measure how well clusters map to known groupings of proteins. For each cluster C_j, we find the group G_i that maximizes the over-

lap between it and cluster C_j. That is, we define **Jaccard**$C_j = \max_i Jac_{ij}$ for the **Jaccard** measure, **PR**$C_j = \max_i PR_{ij}$ for the **PR** measure, and **sDensity**$C_j = density(C_j, \mathcal{G})$ for the **sDensity** measure where $\mathcal{G}$ is the set of groupings we are considering. If cluster C_j is a singleton cluster, then we define **Jaccard**C_j = **PR**C_j = **sDensity**$C_j = 0$. For each measure, we take an average over the clusters, weighted by cluster size, to obtain **Jaccard**C, **PR**C, and **sDensity**C. That is, **Jaccard**$C = \frac{\sum_{j=1}^{M} |C_j| \cdot \mathbf{Jac}C_j}{\sum_{j=1}^{M} |C_j|}$, and the other two measures are defined analogously. Note that **sDensity**C is similar to the biological homogeneity measure utilized previously to evaluate gene expression clusters [22].

Next, we define scores for a grouping that measure how well the known groups of proteins correspond to clusterings. Here, for each group G_i, we try to find cluster C_j such that maximizes the overlap between it and the group G_i. That is, we define **Jaccard**$G_i = \max_j Jac_{ij}$ for the **Jaccard** measure, **PR**$G_i = \max_j PR_{ij}$ for the **PR** measure, and **sDensity**$G_i = density(G_i, \mathcal{C})$ for the **sDensity** measure where $\mathcal{C}$ is the set of clusters we are considering. For each measure, we take an average over the groups, weighted by group size, to obtain **Jaccard**G, **PR**G, and **sDensity**G. That is, **Jaccard**$G = \frac{\sum_{i=1}^{N} |G_i| \cdot \mathbf{Jac}G_i}{\sum_{i=1}^{N} |G_i|}$, and the other two measures are defined analogously.

Finally, we define **Jaccard** as the harmonic mean of **Jaccard**C and **Jaccard**G, **PR** as the harmonic mean of **PR**C and **PR**G, and **sDensity** as the harmonic mean of **sDensity**C and **sDensity**G.

2.2.4 Quality of performance metrics

We demonstrate the utility of our performance metrics by using them to evaluate clusters found in random networks vs. real networks. We generated 10 random networks for each of the four networks under consideration using a degree-preserving stub-rewiring algorithm [62]. We ran the six clustering algorithms on the four original networks as well as their randomized versions using the parameters given above. We

find that the performance, as judged by our introduced measures, of each clustering approach is better in real networks than the corresponding randomized networks (see Table 2.2). For example, when considering *CFinder* on the HTP network and using either MIPS, BP or CC as the desired set of groupings, each of the three measures is > 8.2 times larger on the real network than its average over 10 random networks; these ratios are > 2.3 for *SpectralMod*, > 1.6 for *DPClus*, > 19.2 for *Mcode*, > 1.5 for *MCL*, and > 1.5 for *NetworkBlast* (We note that this analysis also provides some information about the quality of the underlying clustering algorithms; for example, the measures always stay the same for the trivial *OneCluster* algorithm).

When MIPS complexes are gold standard groups:

Measures \ Clustering algorithms	*SpectralMod*	*DPClus*	*Mcode*	*MCL*	*CFinder*	*NetworkBlast*	*OneCluster*
Jaccard	6.5468	4.3031	28.1018	4.4964	9.3768	4.0750	1
PR	20.5910	6.3058	57.6330	6.0946	14.2724	6.1971	1
Density	8.0006	156.0403	595.2888	251.2888	315.2934	63.0358	1

When functional modules relating to BP terms are gold standard groups:

Measures \ Clustering algorithms	*SpectralMod*	*DPClus*	*Mcode*	*MCL*	*CFinder*	*NetworkBlast*	*OneCluster*
Jaccard	2.5416	1.6094	19.2842	1.5173	8.2832	1.5668	1
PR	6.5237	2.3261	38.0035	2.0537	13.5513	2.8768	1
Density	2.3686	10.7267	111.4625	8.4925	47.2376	11.7357	1

When functional modules relating to CC terms are gold standard groups:

Measures \ Clustering algorithms	*SpectralMod*	*DPClus*	*Mcode*	*MCL*	*CFinder*	*NetworkBlast*	*OneCluster*
Jaccard	4.5041	3.6700	44.7057	3.9566	16.1021	4.0501	1
PR	14.8470	6.3049	97.4034	6.8857	28.1849	7.9848	1
Density	4.4346	20.6418	223.9784	20.3107	67.1806	17.8027	1

Table 2.2: **The ratios of performances of the clustering algorithms on the actual network vs. their average performances over the randomized networks.** The ratios of performances of the clustering algorithms on the high-throughput physical interaction network as compared to averages over the randomized networks. For *OneCluster*, the clustering is the same regardless of the network structure, so this ratio is always 1.

The **sDensity** measure seems to be the best with respect to its ratio in real vs. random networks. For example, when considering *CFinder* on the HTP network, and using either MIPS, BP or CC as the desired set of groupings, **sDensity** is > 47 times larger on the real network than its average over the randomized networks (Ta-

ble 2.2). Overall, our performance evaluations metrics are typically much higher when evaluating clusters derived from real networks as compared to those derived from random networks, demonstrating the strength of our measures. Of the 72 evaluations we performed (4 networks, 6 algorithms, 3 groupings), the only exception to this is *NetworkBlast*'s performance in recapitulating BP and CC modules from the Y2H network; in this case, **Jaccard** is on average better in the randomized networks than the actual network (data not shown). We note that the previously introduced separation, positive predictive value and accuracy measures for evaluating interactome-derived clusters [17] were often similar in value on clusters derived from networks corresponding to single high-throughput data sets as they were on the corresponding randomized networks.

As a sanity check, we also constructed three ideal clusterings, each of which corresponds exactly to the groupings we are trying to recover (i.e., protein complexes or functional modules), and evaluated each of those clusterings with respect to all three groupings to compute the maximum performance based on the evaluation framework. We see that performances of ideal clusterings are excellent when compared to the appropriate grouping, as expected. (See Figure 2.2.4.) While ideal clusterings obtain **Jaccard** and **PR** values of 1.0, we note that the performances of ideal clusterings for GO terms evaluated by **sDensity** are lower; this is because of the characteristics of the weight function used.

2.2.5 Subsampling approaches

We subsampled networks from the HTP network in order to obtain networks whose topological features are similar to those of the Y2H network. We tried three subsampling schemes (in succession). Brief descriptions of each are given below.

Scheme 1 samples each interaction from the HTP network with probability proportional to the ratio of the number of edges in the HTP vs. Y2H networks. This

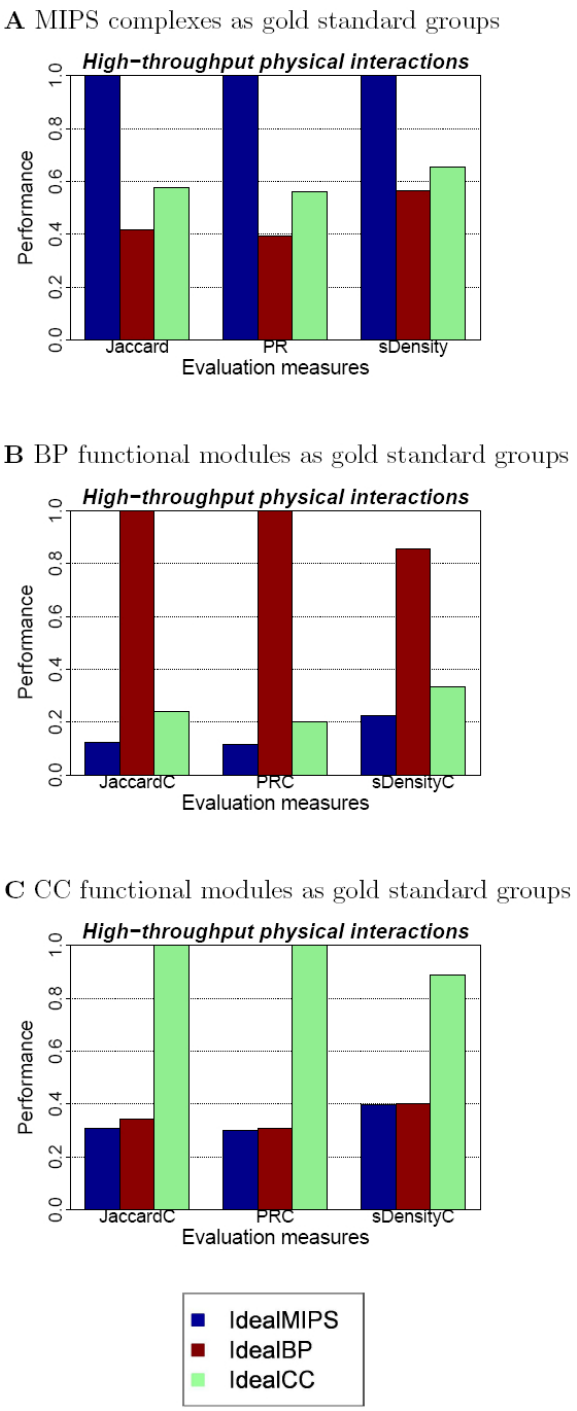


Figure 2.2: **Performances of ideal clusterings.** Three ideal clusterings, "*IdealMIPS*", "*IdealBP*", and "*IdealCC*", are evaluated with respect to how well they recapitulate MIPS complexes, BP modules and CC modules. In *IdealMIPS* all complexes are exactly recapitulated, and in *IdealBP* (respectively *IdealCC*), all BP (respectively CC) modules considered are exactly recapitulated.

results in networks with a noticeably smaller fraction of proteins with degree 1 than in the Y2H network, and noticeably larger fraction of proteins with degree 4 or higher.

Scheme 2 samples proteins and interactions from the HTP network, with the goal of obtaining the same degree distribution in the subsampled network as in the Y2H network. We maintain a degree distribution table which keeps track of how many proteins should be added to the sampled network of each degree in order to match the Y2H network's degree distribution. Initially, the degree distribution table is identical to the degree distribution of the Y2H network. We repeat the following and stop when all elements in the table become 0. For the highest degree d which is not zero in the table, we pick a random protein p from the HTP network whose degree is $\geq d$, where p is not yet in the sampled network. We add p and randomly pick d of its neighbors and add these proteins as well as the interactions between p and them. The desired degree distribution table is updated accordingly (i.e., based on the changes to the node degrees of the sampled network). If adding p makes any element in the table negative, we cancel the addition and pick another protein p at random. This scheme results in sampled graphs that match the desired degree distribution, but the node correlation coefficients are all 0.0.

Scheme 3 is based on our second scheme for subsampling, but attempts to better maintain the node correlation coefficient of the Y2H network. Let $Count(p)$ for protein p be the number of edges amongst the neighbors of p. Let $CountSet(d)$ for degree d be a set of $Count(p)$ for all proteins p whose degree is d. We sample as in Scheme 2 except that at each time we add p and d of its neighbors and the interactions from p to these neighbors, we remove one element from $CountSet(d)$, say *count*. We then pick at random *count* interactions amongst the neighbors of p; if there are fewer than *count* interactions available amongst the neighbors of p, we use all of them. If $CountSet(d)$ is empty, we skip this process. This approach gives excellent agreement with the Y2H network with respect to the number of nodes in the network, the degree

distribution, and the node correlation coefficient.

2.2.6 Protein function prediction

Protein function prediction based on clustering

Given a set of clusters, each protein i is scored with respect to each function f in the following way. For protein i in a cluster, we compute the p-value of all other member proteins in the same cluster having function f based on the hypergeometric distribution (i.e., with parameters as the number of proteins in the entire network, the number of proteins in the cluster, the number of proteins annotated with f in the network, and the number of proteins annotated with f in the cluster). If protein i belongs to multiple clusters, the score for function f is taken to be the minimum p-value computed for this function over all clusters to which it belongs.

Protein function prediction via the neighborhood algorithm

The *Neighborhood* algorithm scores each protein i with respect to function f using the hypergeometric distribution to compute the p-value of protein i's direct interactions having function f.

Evaluation of algorithms for protein function prediction

Since there are parent-child relationships between terms in the BP and CC GO ontologies, for each protein, we update the predictions to deal with such a hierarchy. In particular, for each protein, we update p-values for the functions so that the p-value of a parent functional term is set to be less than or equal to the p-value of any of its children. Thus, given a threshold, if a term is predicted for protein i, then its parent terms are always predicted for protein i (the rationale being that a protein cannot have the more specific functional annotation without having the more general terms as well). We utilize a precision-recall (PR) curve, as suggested by [23], where we vary

the p-value threshold from 0 and 1. For protein i, let m_i be its functional annotations, n_i be a set of predicted functions for i based on the p-value threshold, and k_i be the overlap between m_i and n_i. Then, recall and precision are defined as:

$$\textbf{recall} = \frac{\sum_i |k_i|}{\sum_i |m_i|}, \ \textbf{precision} = \frac{\sum_i |k_i|}{\sum_i |n_i|}.$$

We note that, as outlined earlier, we do not consider overly general or specific functional terms within the ontology. Moreover, proteins within the network that are not annotated by any of these terms are ignored in computing the precision and recall.

2.3 Results

2.3.1 Recapitulating protein complexes and functional modules

We give our performance metrics measuring how well the uncovered clusters correspond to protein complexes, BP functional modules and CC functional modules (Figure 2.3) using the six studied algorithms applied to the HTP and Y2H networks. The analogous results on all four networks are given in Figures 2.4,2.5,2.6. The runtimes of the clustering algorithms on the HTP network are given in Table 2.3. The clustering algorithms vary in the number of clusters they find in each network, as well as the number of singleton proteins left after clustering (see Table 2.4). On HTP network, the algorithms find between 40 and 913 clusters of size > 1 covering between 631 and 4160 proteins, and on the Y2H network, the algorithms find between 34 and 815 clusters of size > 1 covering between 133 and 2828 proteins.

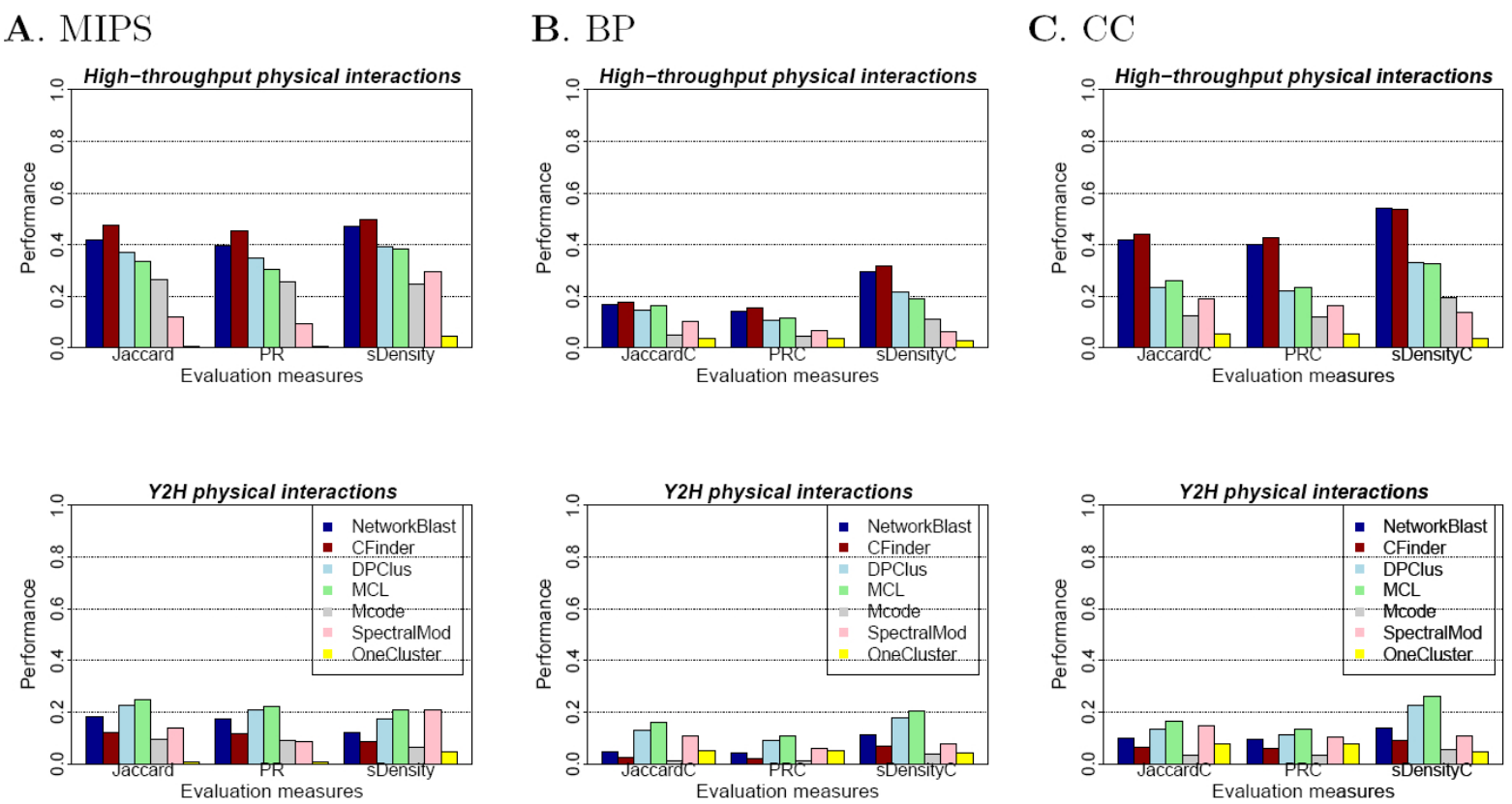


Figure 2.3: **The performance of the clustering algorithms in recapitulating functional modules in the HTP and Y2H networks.** Performance as judged via three measures (**Jaccard**, **PR**, and **sDensity**) of six clustering algorithms and *OneCluster* in how well they recapitulate (A) MIPS complexes, (B) BP modules and (C) CC modules from the HTP (top) and Y2H (bottom) *S. cerevisiae* networks.

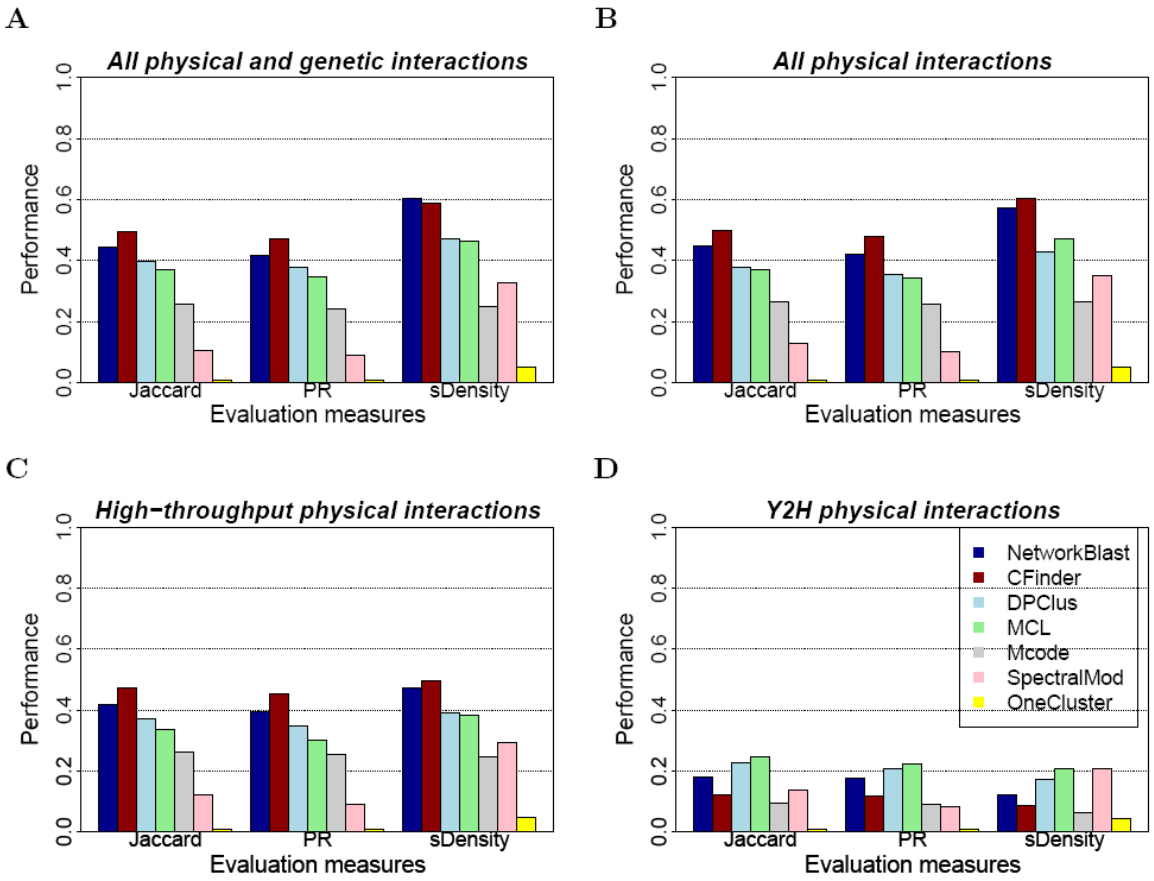


Figure 2.4: **The performance of the clustering algorithms in recapitulating MIPS protein complexes.** Six clustering algorithms and *OneCluster* are evaluated using three measures (**Jaccard**, **PR**, and **sDensity**) on four different networks in how well they recapitulate MIPS protein complexes.

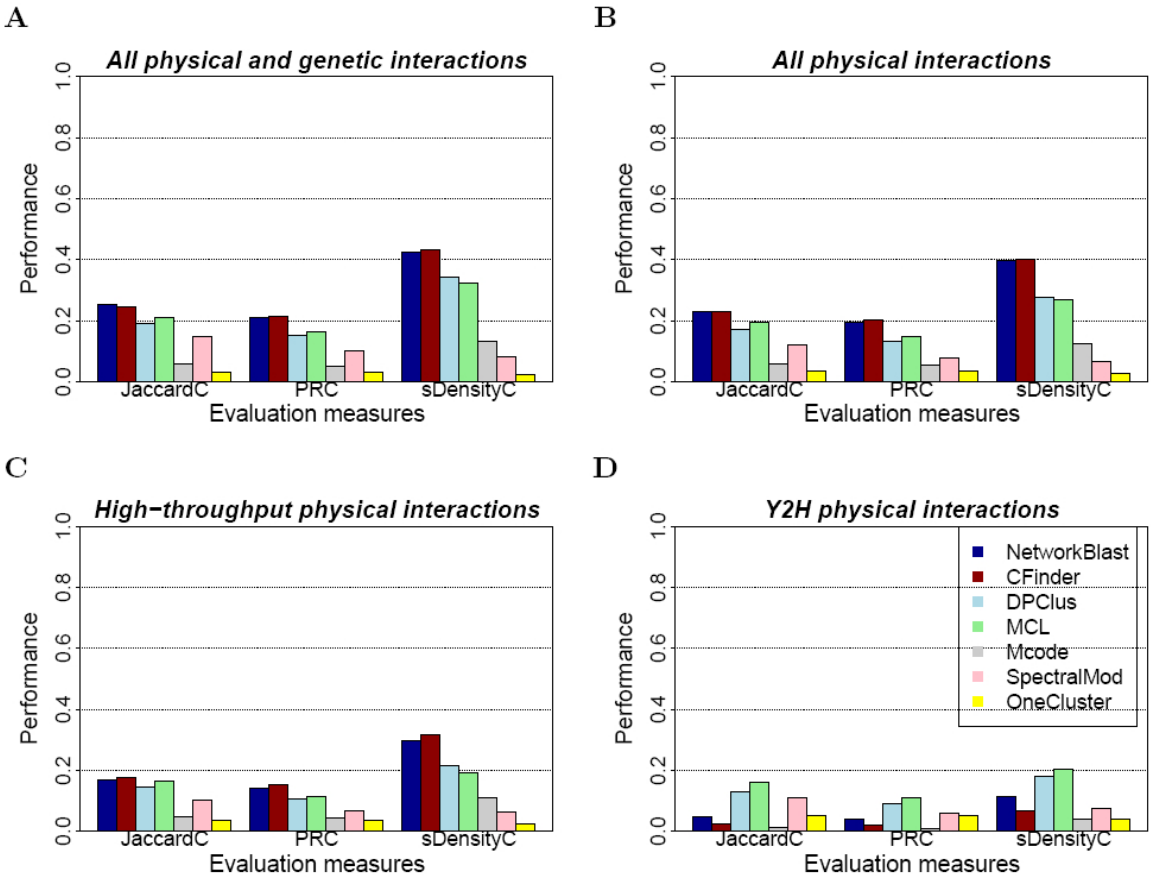


Figure 2.5: **The performance of the clustering algorithms in recapitulating biological process (BP) modules.** Six clustering algorithms and *OneCluster* are evaluated using three measures (**Jaccard**, **PR**, and **sDensity**) on four different networks in how well they recapitulate BP modules.

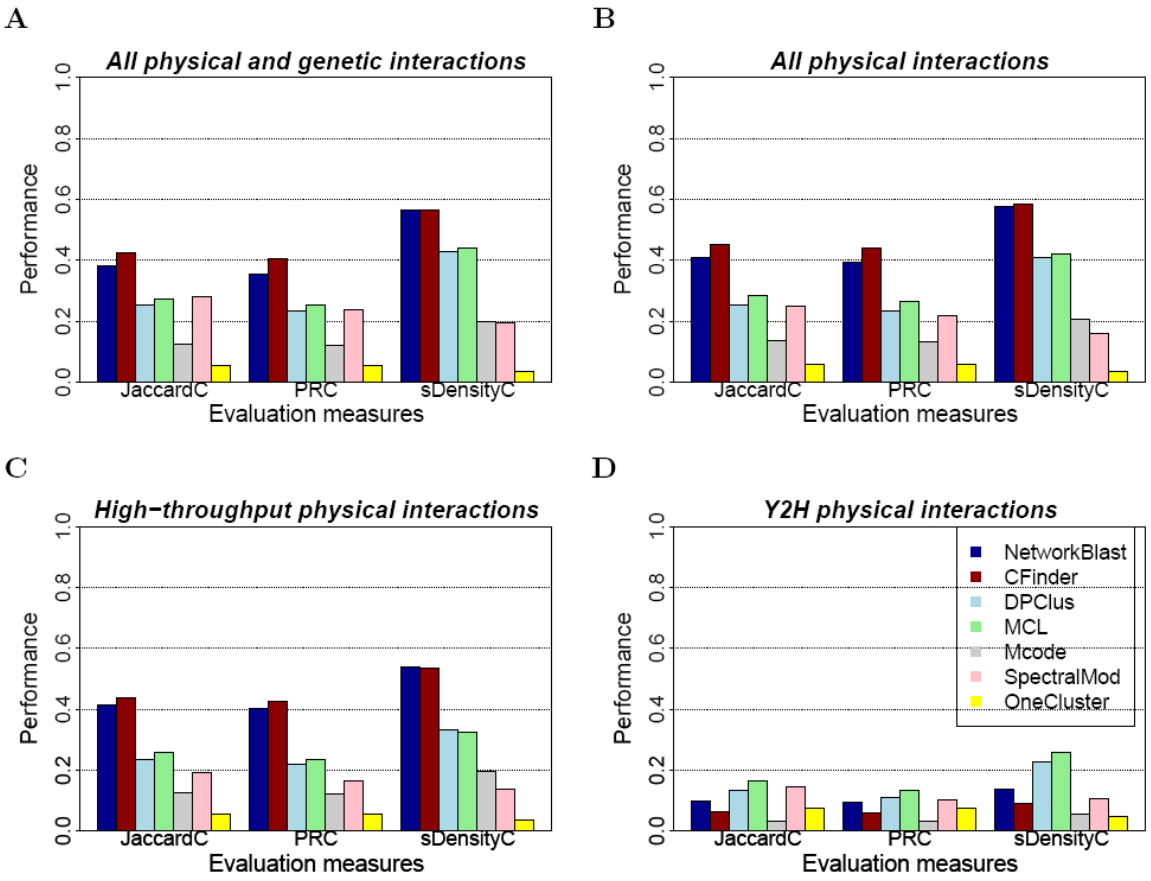


Figure 2.6: **The performance of the clustering algorithms in recapitulating cellular component (CC) modules.** Six clustering algorithms and *OneCluster* are evaluated using three measures (**Jaccard**, **PR**, and **sDensity**) on four different networks in how well they recapitulate CC modules.

Clustering algorithms	SpectralMod	DPClus	Mcode	MCL	Cfinder	NetworkBlast
Run-time (s)	282	713	5	21	4	36

Table 2.3: **Run-times of six clustering algorithms on the *S. cerevisiae* HTP network.**

Network	Spectral-Mod	DPClus	Mcode	MCL	CFinder	Network-Blast
All genetic and physical interactions	23 (4516)	813 (3459)	148 (821)	930 (4287)	728 (2068)	909 (2659)
All physical interactions	33 (4319)	828 (3250)	122 (697)	922 (4129)	576 (1701)	561 (1833)
High-throughput physical interactions	40 (4160)	822 (3117)	122 (631)	913 (3994)	469 (1335)	396 (1371)
Y2H physical interactions	118 (2828)	755 (2229)	34 (133)	815 (2789)	59 (197)	75 (222)

Table 2.4: **Cluster statistics of algorithms on the four *S. cerevisiae* networks.** For each network and each clustering algorithm, we give the number of clusters of size more than 1, as well as the total number of proteins in these clusters (in parentheses).

Stark performance differences in clustering algorithms. We find that there are significant differences in how well the clustering algorithms perform in recapitulating functional modules and protein complexes. For instance, on the HTP network, the best performing approaches for recapitulating complexes as well as biological process and cellular component functional modules are *CFinder* and *NetworkBlast*. Their performance measures for these three tasks on this network are 1.6 to 5.0 times larger than that of *SpectralMod* (Figures 2.3,2.4,2.5,2.6). These two approaches also significantly outperform *Mcode*. Part of the performance difference is due to the number of unclustered proteins: *Mcode* only clusters 631 of the proteins in this interaction network, whereas *NetworkBlast* and *CFinder* cluster 1371 and 1335 respectively. The significant differences in the performances of these algorithms on the various networks confirm that algorithm choice plays an important role in interactome analysis.

No one clustering approach performs best on all networks. Different algorithms perform better on the Y2H network in recovering protein complexes and functional modules than those that perform best on the other networks studied. In particular, on the Y2H network, *MCL* outperforms the other approaches, with

DPClus and *SpectralMod* also demonstrating good performance, whereas *NetworkBlast* and *CFinder* outperform the other algorithms on the other networks. The Y2H network is significantly different from the other three; for example, its average node degree and average node clustering coefficient are significantly lower (see Table 2.1). Relative changes in the performances of the clustering methods are also evident in networks obtained from subsampling from the original HTP network (see section 2.2 **Materials ans methods**), and these changes vary depending on the network at hand. In Figure 2.7), we show the performance of the approaches on these three types of networks subsampled from the HTP network as well as on the Y2H network. Changes in performances of the approaches are evident from the HTP network (Figure 2.3, top). In the subsampled networks whose degree distribution and node clustering coefficients most closely match the Y2H network (Table 2.5), the clustering methods' relative performances are similar to those seen in the Y2H network (Figure 2.7).

The relatively good performance of *MCL* on the Y2H network, which is primarily comprised of a subset of the other networks, is consistent with earlier findings that *MCL* is robust to edge deletions in the network [17]. Overall, the relative change in performance of the clustering algorithms on different networks suggests that there is no clearly superior algorithm in all cases, but that instead algorithm choice should depend on network characteristics. In particular, for the well-studied *S. cerevisiae* interactome, *NetworkBlast* and *CFinder* give superior performance in uncovering complexes and modules from the full network as compared to the other methods, but for less studied organisms with sparser experimentally determined interaction networks, *MCL* may be a better choice.

Specific algorithmic properties of the approaches give hints to the situations to which they are well suited, and can be used to guide algorithm choice. For instance, *SpectralMod* tends to output large clusters as compared to other clustering algorithms

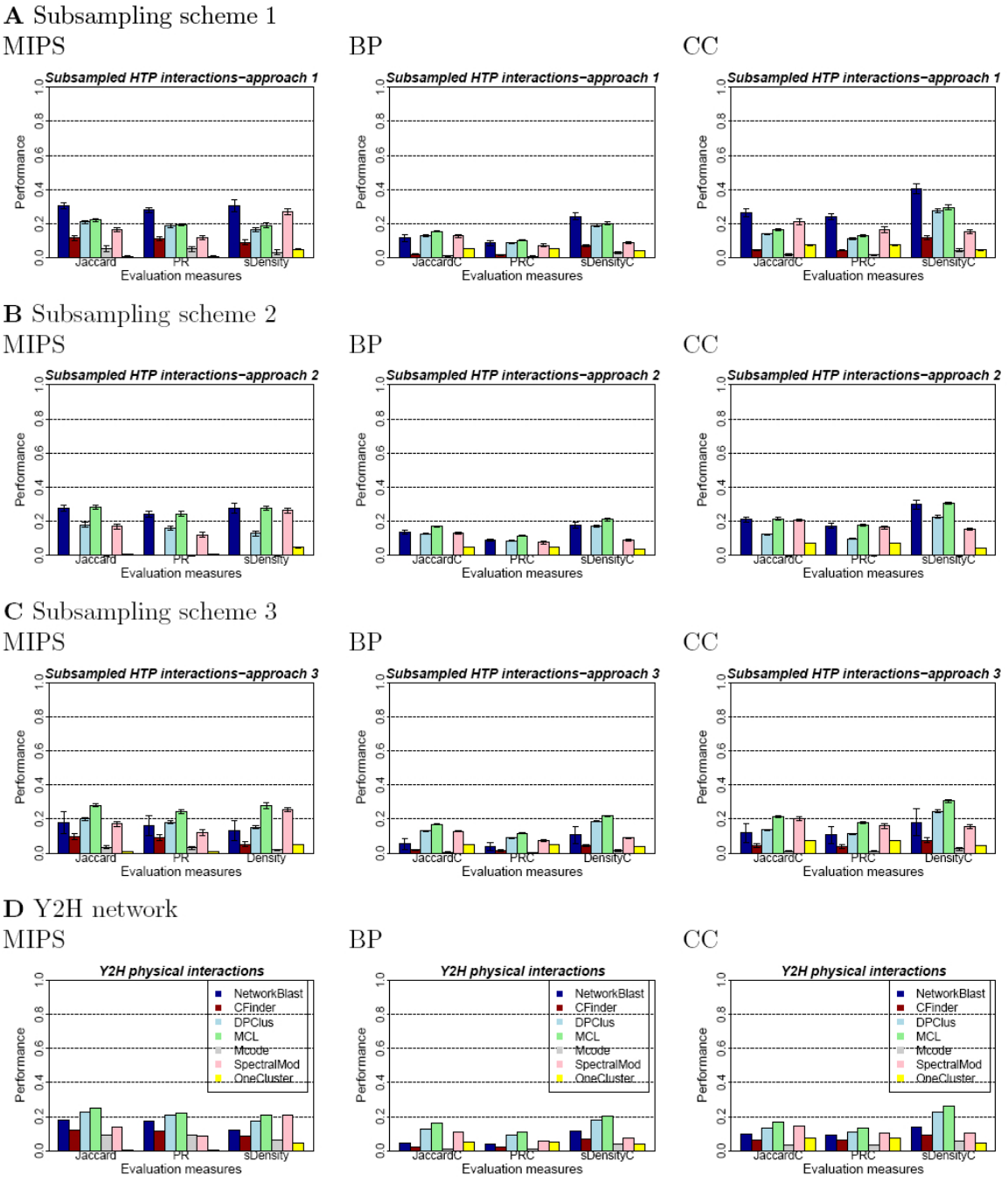


Figure 2.7: **The performance of the clustering algorithms on four types of networks.** **A**, **B** and **C** correspond to networks subsampled from the HTP network using schemes 1, 2 and 3, respectively. For each, 10 networks are subsampled, and average performance measures are shown with error bars indicating one standard deviation above and below this. **D** shows, for reference, the performance of the clustering algorithms on the Y2H network.

Network	#proteins	#interactions	Avg. #neighbors	Avg. NCC
HTP network	4160	11928	5.735	0.136
Y2H network	2828	3170	2.242	0.050
Avg. of subsampled networks with scheme 1	2583	3165.6	2.451	0.048
Avg. of subsampled networks with scheme 2	2829.9	3172.5	2.242	0.000
Avg. of subsampled networks with scheme 3	2830.7	3174	2.243	0.051

Network \ degree distribution	1	2	3	4	5	>=6
HTP network	0.26	0.16	0.11	0.08	0.06	0.33
Y2H network	0.52	0.22	0.11	0.05	0.03	0.06
Avg. of subsampled networks with scheme 1	0.46	0.21	0.12	0.07	0.05	0.09
Avg. of subsampled networks with scheme 2	0.52	0.22	0.11	0.05	0.03	0.06
Avg. of subsampled networks with scheme 3	0.52	0.22	0.11	0.05	0.03	0.06

Table 2.5: **Topological features of the three types of subsampled networks, the HTP network and the Y2H network.** For each network we give: the number of proteins, the total number of interactions, the average number of interactions per protein in the network, the average node clustering coefficient (NCC) and the degree distribution. The NCC for a protein is defined as the number of interactions amongst its interacting proteins, normalized by the total number of possible interactions amongst them. For the networks subsampled from the HTP network, these numbers are averages over 10 networks.

when the network is dense. Thus, it appears to be more suitable for finding large functional modules which correspond to general GO terms as opposed to uncovering more specific functional modules. If a network is very sparse, *SpectralMod* will divide it into many more clusters and will have relatively better performance in our framework. Indeed, *SpectralMod* works comparatively better in the Y2H network than in the other networks. On the other hand, *CFinder* is based on finding "dense" regions in the network; it detects a fewer number of clusters in the sparse Y2H network as compared to the other networks and leaves about 90% of the proteins as singletons. This is a major contributing factor as to why its performances deteriorates in this network.

Advantages of using more complete networks in uncovering complexes and modules. In general, clusters obtained using the full network consisting of all physical and genetic interactions (Figures 2.4,2.5,2.6 (A)) better recapitulate functional modules and protein complexes than those obtained using the other networks. An interesting exception is that cellular component functional modules are somewhat better recapitulated (Figure 2.6 (A), (B)) when using physical interactions only. Genetic interactions are found between (related) pathways [47], though are also found within essential complexes [15]. Depending on the task at hand, it may be advantages to treat these physical and genetic interactions separately [16]. Importantly, clusters obtained using just the Y2H network are significantly worse using all measures in recapitulating functional modules and protein complexes. Additionally, clusters obtained from the HTP network better recapitulate modules and complexes than those obtained from networks subsampled from the HTP network (Figure 2.7).

2.3.2 Predicting protein function

Protein physical interaction data is often utilized to predict protein function. The simplest approach is based on guilt-by-association [73], where a protein is assigned a function based on those that are found frequently amongst its interacting proteins. Alternatively, to better utilize global information, a physical interactome can be clustered first, and then a protein is assigned the functions that are found to be over-represented in its clusters. This more sophisticated cluster-based approach is widely used to obtain hints about protein function. We utilize leave-one-out cross-validation to compare clustering-based methods to a variant of a neighbor majority algorithm, *Neighborhood*, which makes a prediction for a protein based on the over-represented functions found amongst its interacting proteins.

Local approaches outperform clustering in predicting protein function. Table 2.6 (A) shows the area under the PR curve (AUC) for each algorithm in the HTP interaction network. Surprisingly, the simple *Neighborhood* approach has a higher AUC than all clustering algorithms in predicting either BP or CC terms. These results are consistent with earlier work showing that a "neighborhood majority" approach based on total counts performs as well or better than several sophisticated global network approaches [59], as well as earlier work suggesting that *Neighborhood* performs better than *Mcode* for the task of function prediction [76].

In order to assess whether the lower accuracies of clustering algorithms come from their inability to predict function for proteins in singleton clusters, we also considered function prediction when factoring out singleton clusters. We note that is necessary to have the same test set when comparing different clustering approaches with PR AUCs, as baseline performance varies with different test sets. Since each of the clustering approaches leaves a different number of proteins in singleton clusters, for all approaches, we use *Neighborhood* to predict the functions of these proteins, so that we are only considering the performance of transferring function within larger clusters

(Table 2.6 (B)). The *Neighborhood* approach still outperforms the clustering based approaches, though clustering algorithms such as *Mcode* which leave many proteins unclustered see a clear boost in performance.

In order to assess whether the lower accuracies of the clustering algorithms come from transferring functions within large clusters, we next additionally exclude large-sized clusters. For clusters with size greater than 50, as well as those with size 1, we again use *Neighborhood* for proteins within those clusters. For the remaining clusters, we transfer functions according to hypergeometric distribution (Table 2.6 (C)). We still observe that *Neighborhood* has a higher AUC than the clustering-based approaches.

In order to assess whether the lower accuracies of the clustering algorithms come from transferring functions that are infrequent, we next exclude poorly annotated clusters. That is, we filter out clusters where there are no functions annotating more than 50% of the member proteins, and use *Neighborhood* for proteins within those clusters as well as within singleton clusters. For the remaining non-singleton and well-annotated clusters, we transfer functions according to hypergeometric distribution but require that they annotate at least half the proteins in the cluster (Table 2.6 (D)). We still observe that *Neighborhood* has a higher AUC than the clustering-based approaches.

Combining local and clustering approaches for function prediction. One hypothesis for why cluster-based approaches do not work as well as local approaches for function prediction is that a cluster may be composed of several functional modules, and while functions may be statistically enriched within them, they should not be transferred to all the members of the cluster. Accordingly, we next combine clustering information with neighbor annotation information. That is, for each protein within a cluster, we use the *Neighborhood* approach but only consider its interacting proteins within the same cluster while ignoring other proteins within its cluster as

(A) Function prediction.

	Spectral Mod	*DPClus*	*Mcode*	*MCL*	*CFinder*	*Network Blast*	*Neighborhood*
BP	0.0411	0.1557	0.0975	0.1213	0.1276	0.1082	0.1784
CC	0.1337	0.3309	0.1890	0.3003	0.2789	0.2506	0.3743

(B) Function prediction when factoring out singleton clusters.

	Spectral Mod	*DPClus*	*Mcode*	*MCL*	*CFinder*	*Network Blast*	*Neighborhood*
BP	0.0411	0.1593	0.1625	0.1251	0.1498	0.1432	0.1784
CC	0.1337	0.3467	0.3512	0.3084	0.3140	0.2985	0.3743

(C) Function prediction when factoring out singleton clusters and large clusters.

	Spectral Mod	*DPClus*	*Mcode*	*MCL*	*CFinder*	*Network Blast*	*Neighborhood*
BP	0.1320	0.1593	0.1625	0.1251	0.1491	0.1432	0.1784
CC	0.3189	0.3467	0.3512	0.3084	0.3253	0.2985	0.3743

(D) Function prediction when factoring out singleton clusters and poorly annotated clusters.

	Spectral Mod	*DPClus*	*Mcode*	*MCL*	*CFinder*	*Network Blast*	*Neighborhood*
BP	0.1715	0.1755	0.1677	0.1654	0.1787	0.1444	0.1784
CC	0.3331	0.3604	0.3577	0.3420	0.3577	0.3054	0.3743

(E) Function prediction when local topology is considered for clustering algorithms.

	Spectral Mod	*DPClus*	*Mcode*	*MCL*	*CFinder*	*Network Blast*	*Neighborhood*
BP	0.1716	0.1679	0.1710	0.1546	0.1827	0.1815	0.1784
CC	0.3535	0.3521	0.3646	0.3496	0.3772	0.3676	0.3743

Table 2.6: **PR AUC for BP and CC predictions of the clustering algorithms and *Neighborhood* in the HTP *S. cerevisiae* network.**

well as interacting proteins that are in different clusters. For proteins that are not clustered, the *Neighborhood* approach is used while considering all its interactions. In this case, clustering approaches such as *NetworkBlast* or *CFinder* have slightly higher AUCs than *Neighborhood* (Table 2.6 (E)). The drastic improvement of *SpectralMod*'s PR-AUC also supports the idea that *SpectralMod*'s large clusters consist of several smaller functional modules.

Characterizing cluster-based function prediction based on number of annotations. While the *Neighborhood* approach has better performance than cluster-based methods in predicting protein functions for *S. cerevisiae*, clustering approaches have other advantages. In particular, they can uncover structure in networks with no additional information and can make predictions for proteins that interact only with proteins of unknown function. Thus, we expect that for proteomes with larger numbers of unannotated proteins, the performance of the *Neighborhood* approach should decrease faster than that of clustering-based approaches. In order to systematically test this, we analyzed how the algorithms perform as we remove annotations from the proteins in the network. That is, we selected 10%, 30%, 50%, 70%, and 90% of the proteins in the HTP network at random and removed all of their annotations to make an artificial network with fewer annotations. Figure 2.8 shows the PR AUC as a function of the fraction of proteins whose annotations are removed (the values at 0% annotations removed correspond to those in Table 2.6 (A)).

As a large fraction of the yeast proteins' annotations are removed, the clustering-based approaches begin to outperform *Neighborhood*. For example, when nearly 70% of the annotations are removed, *CFinder* outperforms *Neighborhood* for BP prediction and *DPClus* outperforms *Neighborhood* for CC prediction. Most clustering algorithms predict better than *Neighborhood* once 90% of the annotations are deleted. We find the same trends when running this procedure on the other yeast networks. See Figure 2.9 for our results on the Y2H network, where *Neighborhood* still outperforms

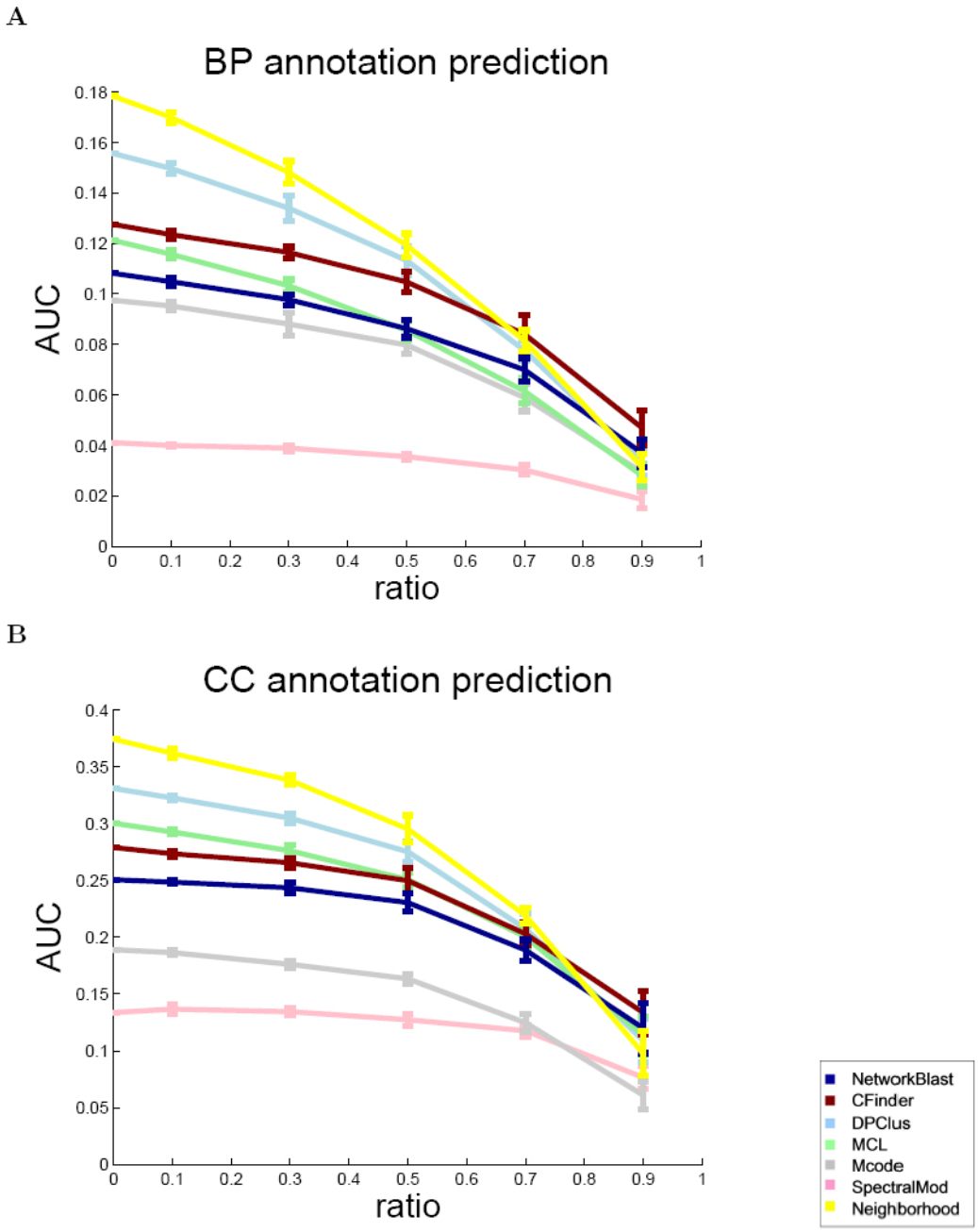


Figure 2.8: **Function prediction performance as protein annotations are removed from the *S. cerevisiae* HTP network.** As BP or CC annotations are removed for 10%, 30%, 50%, 70%, and 90% of the proteins in the high-throughput physical interaction network, the PR-AUC of *Neighborhood* deteriorates more rapidly than that of the six clustering algorithms. The average PR-AUC over 10 networks is plotted, with each error bar showing plus and minus one standard deviation from the average.

the clustering approaches in function prediction on the original network and on networks with a considerable fraction of annotations removed. On the human physical interaction network from BioGRID, where a somewhat smaller fraction of proteins are annotated (51% with BP terms and 31% with CC terms), *Neighborhood* still outperforms the clustering approaches, though *MCL* is competitive with it (see Figure 2.10) and performs better than it when approximately 10% of the annotations are removed. If predictions on unannotated proteins are pessimistically counted as false positives (instead of ignored), then the *Neighborhood* method outperforms the other approaches until 50% of the annotations in the human network are removed (Figure 2.11). In all networks studied, the relative performances of clustering methods in function prediction as compared to the *Neighborhood* method improve as annotations are removed.

2.4 Conclusions

While clustering has become a standard first-line tool in the analysis of physical interactomes, no previous study has systematically assessed how well such an approach performs in predicting protein function and functional modules. In this Chapter, our research establishes guidelines on how and when clustering should be utilized for analyzing physical interaction networks.

Perhaps most importantly, we find that the common practice of looking for enriched functions within clusters is not the best approach for predicting protein function, at least for the yeast proteome. Instead we find that, overall, it is better to use a simple local method such as *Neighborhood* or to use clustering algorithms combined with *Neighborhood* in networks with sufficient annotations. From a computational perspective, this also suggests that clustering algorithms should not be judged solely based on the number of functionally enriched clusters they find, as this may not be

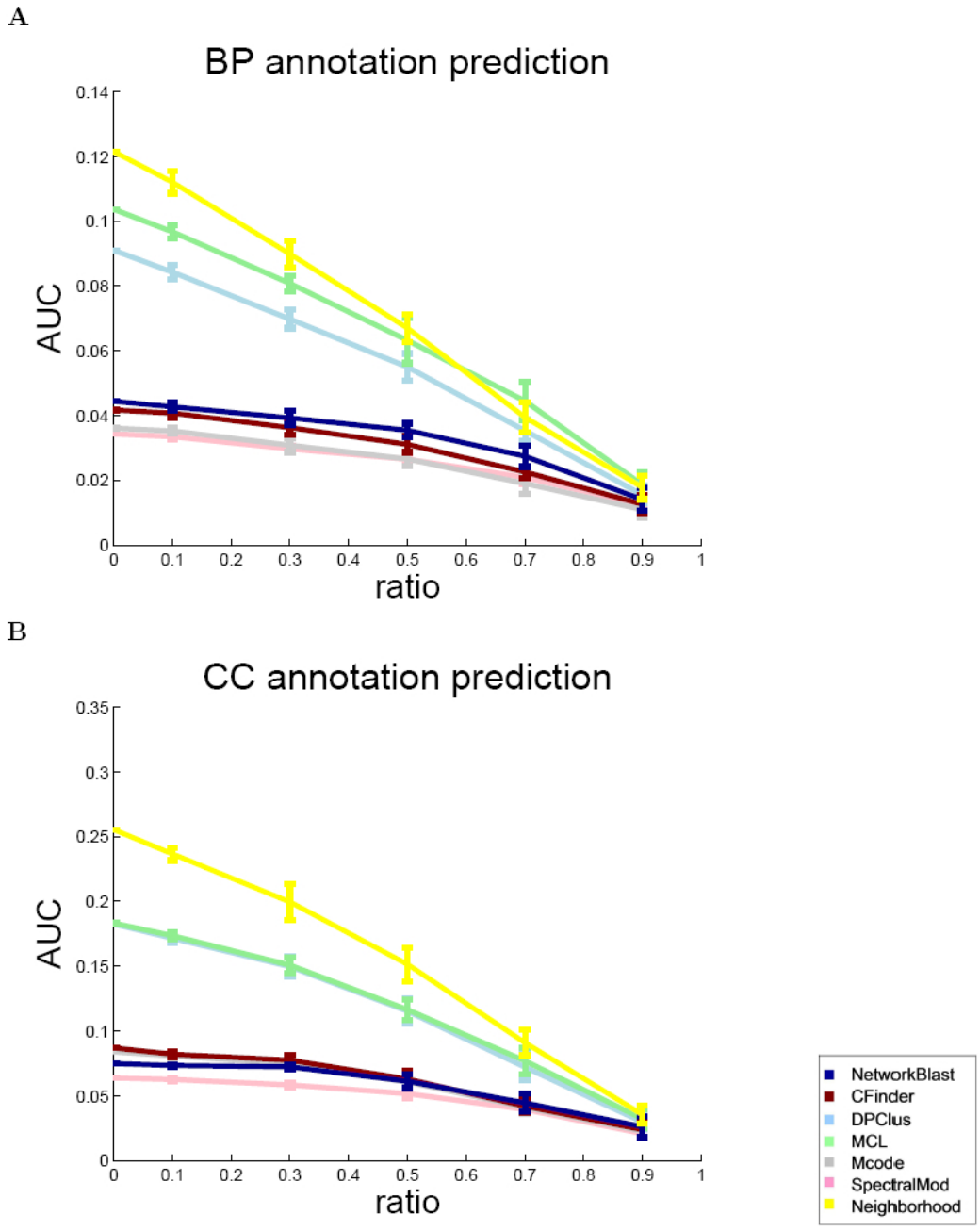


Figure 2.9: **Function prediction performance as protein annotations are removed from the *S. cerevisiae* Y2H network.** As biological process (**A**) or cellular component (**B**) annotations are removed for 10%, 30%, 50%, 70%, and 90% of the proteins in the Y2H network, the PR-AUC of *Neighborhood* deteriorates more rapidly than that of the six clustering algorithms. For annotation removal, we repeated the process ten times and plotted average AUC values, with error bars indicating plus and minus one standard deviation.

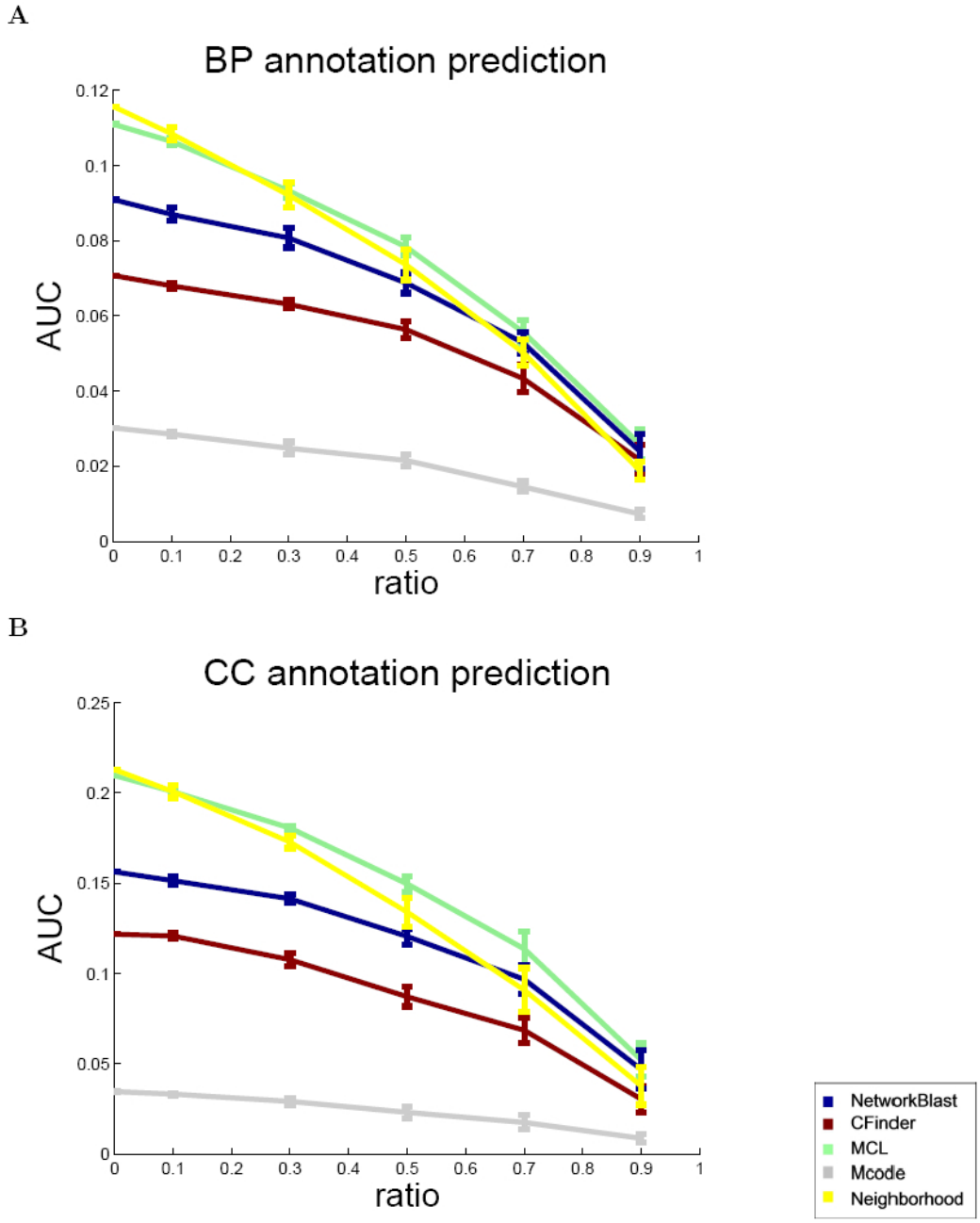


Figure 2.10: **Function prediction performance as protein annotations are removed from the human network.** Biological process or cellular component annotations are removed for 10%, 30%, 50%, 70%, and 90% of the proteins in the human physical interaction network. The removals are repeated ten times. Average PR-AUC values are plotted, with error bars indicating plus and minus one standard deviation. (*SpectralMod* and *DPClus* did not successfully cluster these large networks.)

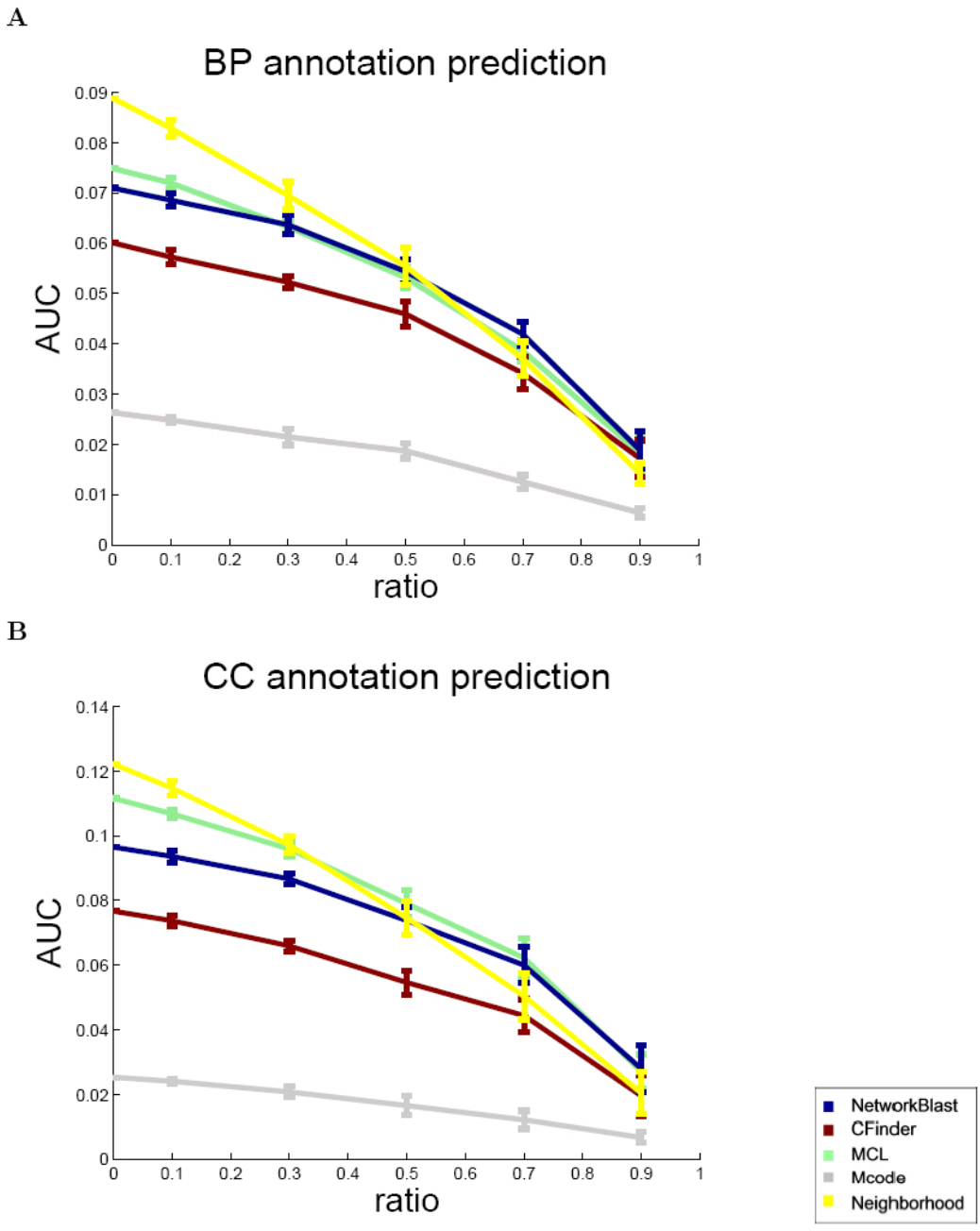


Figure 2.11: **Function prediction performance as protein annotations are removed from the human network, while keeping in the evaluation proteins not annotated with any function.** Biological process or cellular component annotations are removed for 10%, 30%, 50%, 70%, and 90% of the proteins in the human physical interaction network. The removals are repeated ten times. Average PR-AUC values are plotted, with error bars indicating plus and minus one standard deviation. (*SpectralMod* and *DPClus* did not successfully cluster these large networks.)

the best way to do interactome-derived function prediction for the proteome at hand. The strength of clustering is that it uncovers structure within biological networks, even when nothing is known about individual proteins. Thus, for less annotated proteomes, or even biological processes that have not been well-studied, the advantages of clustering over local methods are more likely to be apparent. Indeed, our simulations show that the relative performance of clustering approaches as compared to a simple neighborhood functional annotation scheme improves with fewer annotations. In the future, it would be desirable to characterize which method should be used for function prediction at the per-protein level; this could depend on, for example, the number of annotated interacting proteins, local measures of network topology, the density and size of the clusters it is found within, and the particular functions being predicted.

We also find that the topological features of networks can vastly affect the performance of clustering algorithms in recapitulating functional modules; in particular, some of the best performing algorithms on the more dense HTP network are among the poorest performing in the Y2H network as well as in networks subsampled from the HTP network to resemble the Y2H network with respect to network topological features (Figure 2.7). This suggests that network characteristics should guide algorithm choice, and there is no one algorithm that always outperforms others in predicting functional modules. It is possible that for some clustering approaches, more fine-tuned parameter choices may lead to better results; however, for approaches such as *CFinder* and *SpectralMod*, which have one and zero parameters respectively and whose relative performances swap between the HTP and Y2H networks, this is not the case. Moreover, we note that while MCL has become the algorithm of choice, and the one to which new approaches are most commonly compared to, as a result of its excellent performance in recovering complexes [17], we find that other approaches have better performance depending on the application and network at hand.

Looking forward, we hope that our evaluation framework will be helpful in gauging how well future methodological improvements in clustering translate to improved detection of functional modules and protein complexes from interactomes.

Chapter 3

From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interaction network at different scales of organization

3.1 Introduction

High-throughput experimental approaches for determining protein interactions have resulted in large-scale cellular networks for numerous organisms. Graph-theoretic analyses of these networks have been a great aid in advancing our understanding of cellular functioning and organization (review, [2]). One of the most fundamental discoveries is that there is a strong relationship between the topological characteristics of cellular networks and their underlying functioning. For example, cellular networks consist of tightly clustered groups of interacting proteins, and these proteins work

together as protein complexes or biological processes to achieve specific biological functions [39, 69, 79, 8, 64, 77]. An orthogonal decomposition reveals that there are recurring and over-represented topological and functional patterns within larger cellular networks, and these network motifs [57, 54] and network schemas [10] can be associated with dynamic regulatory properties and shared mechanisms of functioning. Here, we revisit perhaps the most basic structure-to-function relationship in cellular networks—that between the number of interactions which a protein has and its overall functional importance.

The importance of a gene to a cell or an organism can be quantitatively measured by considering the phenotypic effects of gene deletion or disruption. Experimental studies in the baker's yeast *S. cerevisiae* have demonstrated that approximately 19% of its proteins are essential; that is, the deletion of these proteins results in cell death, even in optimal growth conditions [90, 34]. Early computational analysis of the yeast *S. cerevisiae* protein-protein physical interaction network revealed a scale-free topology where a few "hub" proteins have many interactions, and also showed that hub proteins are more likely to be essential than other proteins [44]. Numerous subsequent studies have also confirmed this centrality-lethality relationship, not only in yeast [93, 30, 13, 94, 96] but also in other organisms [36]. While the positive correlation between protein interaction degree and essentiality is widely accepted, the proposed reasons underlying this relationship have been more controversial and intensely studied.

Initial work suggested that high-degree proteins may be essential due to their role in global network connectivity [44]; however, this is unlikely to be the case as it was subsequently shown that non-essential hubs are just as important as essential hubs for maintaining connectivity, and that essentiality is better correlated with local, rather than global, measures of connectivity in protein-protein interaction networks [94, 96]. It was alternatively proposed that essentiality is a property of interactions; that is,

there are essential protein interactions, without which an organism cannot survive, and these are randomly distributed across the network and hubs tend to be essential as they are more likely to participate in essential interactions [40]. However, this model implies that the probabilities that two non-interacting proteins are essential are independent of each other, and this is not the case [96]. Instead, Zotenko et al. [96] argued that the correlation between degree and essentiality is due to the participation of essential proteins in essential functional modules consisting of groups of densely clustered and functionally related proteins. They further showed that the essentiality of hubs that are not in these computationally extracted modules are only weakly correlated with degree [96]. Indeed, it had previously been found that essential proteins tended to be densely connected to each other [93] and concentrated in complexes [24, 38], suggesting that essentiality is a modular property rather than a property of individual proteins. Building upon this, it has been argued that essential complexes tend to be large, and thus proteins within them have a larger number of interactions, and that this explains why hubs tend to be essential [89].

While essentiality appears to be a modular property in protein-protein interaction networks, it is clear that complexes and processes do not consist entirely of essential or non-essential proteins. Do essential proteins within a complex essential differ from the non-essential ones? Further, not all complexes and processes contain essential proteins. Do such essential complexes have distinctive roles in cellular networks? In this Chapter, we aimed to discover whether, within complexes, their essential and non-essential proteins differ in their interaction properties, and at a more global scale, whether essential and non-essential complexes differ in their network-level properties. To accomplish this, we developed a computational framework that incorporates known functional information about proteins into network analysis techniques. Because of quality concerns about protein interaction networks and protein functional annotations, we performed our analysis on three different yeast interac-

tion networks and utilized functional information arising from GO biological process annotations [6] at different levels of resolution as well as information about protein complex membership.

We began by re-examining the relationship between protein essentiality and network modularity. We hypothesized that if essentiality is a modular property, then a protein's intramodular physical interaction degree should be better predictor of a protein's essentiality than its intermodular physical interaction degree. To test this, we utilized biological process functional annotations of proteins and classified physical interactions into intraprocess interactions within processes and interprocess interactions between processes. We found that essential proteins tend to have many interactions with proteins within the same functional modules and that the intraprocess interaction degree is more correlated with essentiality than overall degree. Further, we found that the relationship between overall degree and essentiality is significantly weakened when controlling for intramodular degree, but is not affected when controlling for intermodular degree. These findings confirm in a more direct manner previous work [96] arguing that proteins are essential due to their interactions within essential modules consisting of functionally similar proteins.

To further ascertain whether the modularity of essential proteins is due to their participation in essential protein complexes or more generally within essential biological processes, we repeated this analysis while first exclusively focusing on proteins within protein complexes and next focusing only on proteins that are not within known protein complexes. We found that most essential proteins with many intraprocess interactions in fact participate in essential protein complexes or in essential biological processes that include one or more protein complexes; that is, essentiality appears to be a property of protein complexes and not a more general property of biological processes.

Next, we examined complexes and processes that contain essential proteins, and

found that their essential proteins tend have to more interactions, particularly intra-complex interactions, than their non-essential proteins. That is, while essentiality is a modular property, the degree of a protein is associated with essentiality within essential complexes; this suggests that these essential proteins may play a more important role in maintaining the functioning of complexes.

Finally, we analyzed modules containing essential proteins within the context of other functional modules. We inferred significant "cross-talks" between protein complexes and biological processes and used them to build module-level networks, in which two processes are linked if they have a statistically enriched number of physical interactions between them. Using these module-level networks, we uncovered that functional modules with essential proteins tend to have high degree; that is, degree in the module-level network is positively correlated with module essentiality.

Overall, by considering proteins within the functional context of the yeast interactome, we show that there is a relationship between essentiality and network topology at different levels of cellular organization; that is, the centrality-lethality rule is true not just at the protein level but also at the module level, with complexes and processes that are essential tending to interact with many functional groups.

3.2 Materials and methods

3.2.1 Physical interaction datasets

We utilized three physical interaction datasets. First, physical interactions were gathered from BioGRID [81], release 3.1.78, using all evidence codes indicative of physical interactions except "Affinity Capture-RNA" and "Protein-RNA," and including only core data for [42]. If a protein has more than 30 interactions from a single experimental data source, we removed these interactions. Second, we extracted a network from BioGRID that is focused on direct physical interactions by utilizing interactions

determined from one of the following experimental systems: Biochemical activity, Co-crystal structure, Far western, FRET, Protein-peptide, Reconstituted complex, and Two-hybrid. Third, we used a network consisting of interactions determined via Affinity capture-Western and Affinity capture-MS. We refer to the three networks as *Full*, *Direct* and *Pull-down*, respectively, and their sizes are given in Table 3.1.

Network	**# Proteins**	**# Interactions**	**Fraction of essential proteins**
Direct	4031	15073	0.22
Pull-down	4449	36455	0.22
Full	5167	50170	0.20

Table 3.1: **The number of proteins, the number of interactions and the fraction of essential proteins for each of the three physical interaction networks considered.**

3.2.2 Protein complexes and biological processes

We used the set of 430 protein complexes compiled in [12], which includes the SGD Macromolecular Complex GO standard [74], the CYC2008 protein complex catalog [66] and a set of manually curated complexes. From this initial set, we removed highly overlapping complexes as follows. First, if the proteins comprising one complex are a subset of the proteins comprising another complex, the smaller complex is removed. Next, for any two complexes, if the Jaccard index of the proteins making them up (i.e., the number of overlapping proteins divided by the size of the union of the protein sets) is ≥ 0.5, we removed the smaller complex. This resulted in 394 complexes. There are four complexes consisting of subunits of the ribosome. These complexes were removed from consideration, leaving 390 complexes comprised of 1593 proteins.

For our functional analysis, we worked with a subset of specific Gene Ontology (GO) Biological Process (BP) terms [6] that were derived from the entire GO (version 1.1.2130) as follows. First, we extracted 1418 BP terms, each of which annotates at

least 5 yeast proteins and at most 50. Next, to hone in on the contribution of a specific biological process (as opposed to the effects arising from proteins that are annotated with that process but are also within protein complexes), we pruned the set of proteins that are associated with these functional terms. More specifically, if the size of the intersection between a biological process and one of our original set of 430 protein complexes is ≥ 2, the proteins in the intersection were no longer associated with the process. If this left fewer than 2 proteins associated with the process, or with less than half the number of proteins that it is known to annotate, then this term was removed from consideration. Finally, highly overlapping processes were removed in the same manner as described above for complexes. This procedure resulted in 391 processes, with 2567 proteins associated with at least one of these processes.

3.2.3 Detecting cross-talk between complexes and processes

We determined within a given network whether certain pairs of functional modules are enriched in the number of interactions found between them in the following way. Functional modules consist of either proteins within the same complex, or that have a shared process annotation from the 391 filtered processes considered. We consider modules arising from complexes or processes in turn. Briefly, we limited the network to include those proteins that are associated with one of the modules that we are considering, and all intermodular interactions amongst these proteins. Next, for any two modules c_1 and c_2, we counted the number of "cross-talk" interactions between the proteins comprising each of these modules. Note that interactions where either of the proteins is annotated with both c_1 and c_2. were not included in the network as these are intramodular interactions. To determine whether the number of observed cross-talk interactions is more than would be expected by chance, we randomized the network 100 times using stub-rewiring (as in [57]), thereby preserving degree distribution and module annotation. For each randomized network, the number of

cross-talk interactions are counted for all pair of modules. Lastly, if $count_{\{c_1,c_2\}}$ is the number of cross-talk interactions between c_1 and c_2 in the real network, and $avg_{\{c_1,c_2\}}$ is the average number of corresponding cross-talk interactions in randomized networks, the likelihood of the module pair is as follows:

$$\log \frac{(\ count_{\{c_1,c_2\}} + 1)}{(\ avg_{\{c_1,c_2\}} + 1)}$$

The addition of the pseudocount of 1 downweighs the contribution of very rare cross-talks that could otherwise obtain high scores simply due to very small (or zero) average counts in the randomized graphs. We required, in order for a module pair to be considered a cross-talk, that there should be at least two independent (i.e., non-overlapping) cross-talk interactions, and that its likelihood should be at least 2.

3.2.4 Semantic similarity

The semantic similarity between two GO terms within the same ontology is an estimate of the functional similarity between the terms. We use the semantic similarity measure introduced by [50]. In particular, let $f(a)$ be the fraction of proteins in yeast annotated with term a among the total number of proteins. Then $s(a) = -log(f(a))$ is a measure of how specific a term a is. We compute the term semantic similarity of a and b, $tSS(a,b)$ as $tSS(a,b) = \frac{2 \cdot s(LCA(a,b))}{s(a)+s(b)}$, where $LCA(a,b)$ is a least common ancestor of a and b in the GO ontology.

Note that if the LCA of two terms is a root term (e.g., GO:0008150 'biological process'), then $tSS(a,b) = 0$. Moreover, if two terms are the same, then $tSS(a,b) = 1$.

This measure is naturally extended to functional relationship between proteins that have multiple annotations. For a protein p, let $A(p)$ be the set of terms with which p is annotated. If a term annotates p, then all its parent terms are naturally included in $A(p)$. Then, between proteins p and q, the protein semantic similarity

(pSS) is defined as follows [77]:

$$pSS(p,q) = \frac{2 \cdot \max_{a \in A(p) \cap A(q)} s(a)}{\max_{a \in A(p)} s(a) + \max_{a \in A(q)} s(a)}$$

3.3 Results

We analyzed 5640 proteins that were tested for essentiality [34] in the context of three large-scale *S. cerevisiae* protein physical interaction datasets; each of these networks captures different features of biological interactions. The first network is a *Direct* interaction network, where an interaction between two proteins corresponds to a direct physical contact; this network includes interactions determined by yeast two-hybrid. Next, we considered a *Pull-down* network, where an interaction between two proteins corresponds to their being members in the same multiprotein complex. Finally, we considered the *All physical* network consisting of all physical interactions in BioGRID [81]; in this case, the interactions can represent either direct or indirect interactions. In this Chapter, we focus on our results on the *Direct* interaction network, which contains 4031 proteins (898 of which are essential) and 15,073 interactions. We also report the full analysis on the other two networks.

3.3.1 Categorizing interactions as intramodular or intermodular

For a given interaction network, we labeled protein interactions as either "intramodular," "intermodular" or neither using two sources of functional data in turn. In particular, we utilized yeast protein complex data compiled in [12] and Gene Ontology (GO) Biological Process (BP) annotations [6]. Thus, intramodular interactions can arise from either intracomplex or intraprocess interactions, and intermodular interactions arise as either intercomplex or interprocess interactions; we will separately consider

both types of intramodular and intermodular interactions. For protein complex data, "intracomplex" interactions are between all pairs of proteins that participate in a shared complex and "intercomplex" interactions are between pairs of proteins that are found in at least one complex but are never found in the same complex (See Figure 3.1).

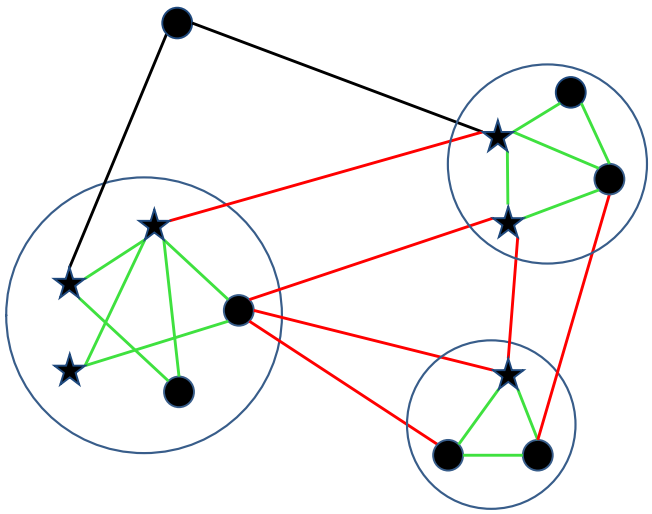


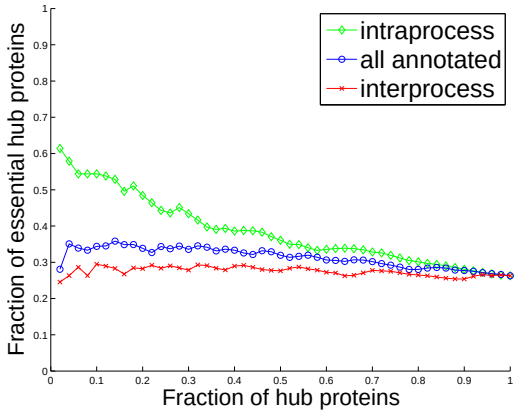
Figure 3.1: **Schematic showing how interactions are categorized into intramodular and intermodular given a set of functional modules.** Each circular node represents a non-essential protein and each star node represents an essential protein. Blue circles represent functional modules, either derived from a protein complex or a biological process, and proteins within a circle are associated with the corresponding module. Green, red, and black interactions represent intramodular, intermodular, and unannotated interactions, respectively.

It is more complicated to characterize interactions as intramodular or intermodular using GO BP terms, as the terms are hierarchically related and annotate different numbers of proteins, with some very general terms. To get only informative and specific terms, we considered GO BP terms that annotate at most 50 proteins in the yeast proteome. An interaction is unannotated unless both proteins are annotated with one of these specific GO BP terms. An interaction is "intraprocess" if it is between two proteins sharing one of these specific BP terms. If two proteins with an interaction are annotated with specific GO BP terms but do not share any specific

BP terms, the interaction is "interprocess." We note that while physical interactions are largely thought of as "within process," especially as compared to other types of interactions [73], a significant fraction of physical interactions are interprocess (Table 3.3); this is true even as the threshold for choosing specific terms is varied from 50.

3.3.2 Intraprocess interactions are a main factor in the relationship between protein essentiality and interaction degree.

As a first step towards relating protein essentiality to network modularity, for each protein, we computed its number of intramodular interactions, intermodular interactions, and total annotated interactions. We then considered each of the intramodular, intermodular and total annotated interaction degrees in turn, and ordered all proteins from high to low degrees with respect to it. As we varied the threshold for the number of proteins considered, we computed the fraction of essential proteins in the "high degree" or "hub" set. Over the range of thresholds, the high degree proteins, as ranked by intramodular degree, have a larger fraction of essential proteins than the high degree proteins as ranked by either total annotated degree or intermodular degree (Figure 3.2 (a), 3.3 (a), and 3.4 (a)). Further, in general, for all three networks, the fraction of essential proteins decreases as the threshold for intramodular, intermodular or total degree is lowered. In the *Direct* network (Figure 3.2 (a)), this trend is only true for intramodular interaction degree and is not true for total degree; this is consistent with previous work showing that the relationship between essentiality and overall interaction degree is weak in networks consisting of interactions determined by the yeast two-hybrid method [13, 96].

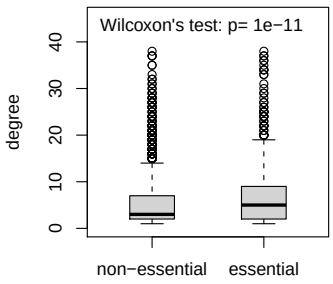


(a) correlation between degree and essentiality for proteins in BPs

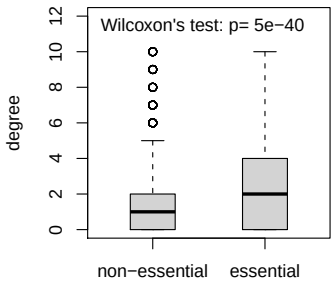
Interaction type	SRCC (p-value)
Intraprocess	0.2472 (0)
All annotated	0.1253 (2e-11)
Interprocess	0.0000 (1*)

(b) Spearman's rho rank correlation coefficient (SRCC)

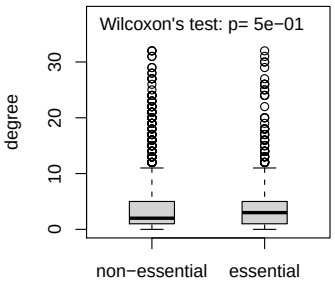
Interaction type	Partial SRCC (p-value)
Intraprocess	-0.0255 (0.2*)
Interprocess	0.2466 (0)



(c) all annotated degree

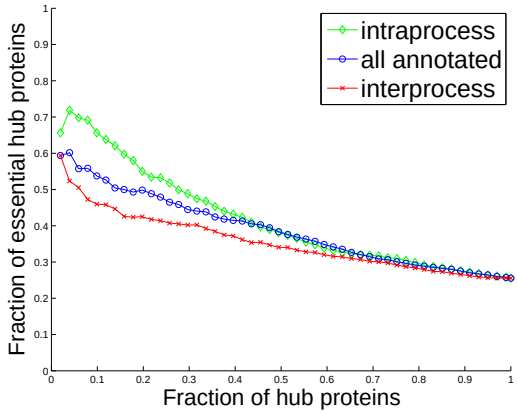


(d) intraprocess degree



(e) interprocess degree

Figure 3.2: **The intraprocess interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Direct* network, when interactions are categorized with specific Gene Ontology (GO) Biological Process (BP) terms, each of which annotates at most 50 proteins. (a)** The fraction of essential proteins among hub proteins decreases as more proteins are considered hub proteins; this is done by adding proteins in a non-increasing order of the interaction degree. The correlation with the intramodular degree (green) is highest, followed by the all annotated degree (blue) and then the intermodular degree (red). **(b)** Spearman's rho rank correlation coefficient (SRCC) is given to measure the correlation between degree and protein essentiality. The partial correlation is also computed between all annotated degree and essentiality when controlling for either intramodular or intermodular degree. * indicates a non-significant p-value > 0.05. **(c)-(e)** The degree distribution of non-essential proteins is compared to that of essential proteins for all annotated (c), intramodular (d), and intermodular (e) degree, respectively. Essential proteins tend to have a higher degree than non-essential proteins. In each box plot, the horizontal bar within a box gives the median of the distribution; the two ends of the box give the 25% and 75% percentile, respectively; the two ends of the whiskers give the minimum and maximum of the degree data, respectively; and the small circles show outliers within 2-98%. The significance of the difference of the two degree distributions is measured by the Wilcoxon rank sum test.

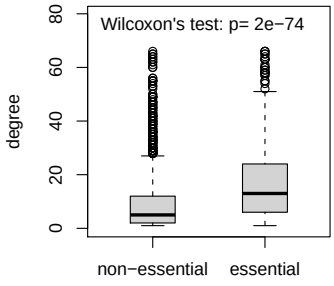


(a) correlation between degree and essentiality for proteins in BPs

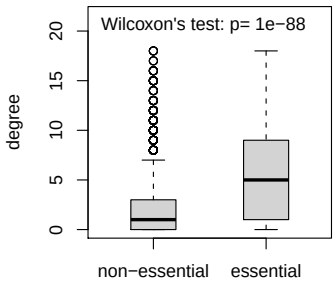
Interaction type	SRCC (p-value)
Intraprocess	0.3505 (0)
All annotated	0.3203 (0)
Interprocess	0.2232 (0)

Interaction type	Partial SRCC (p-value)
Intraprocess	0.1120 (2e-10)
Interprocess	0.2815 (0)

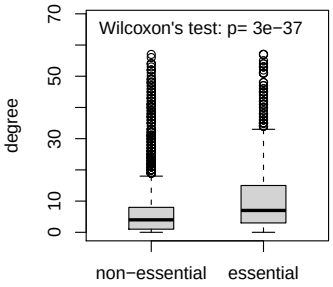
(b) Spearman's rho rank correlation coefficient (SRCC)



(c) all annotated degree

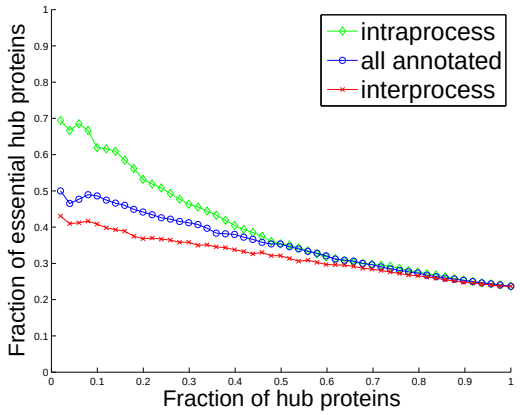


(d) intraprocess degree



(e) interprocess degree

Figure 3.3: **The intraprocess interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Pull-down* network, when interactions are categorized with specific Gene Ontology (GO) Biological Process (BP) terms, each of which annotates at most 50 proteins.** All tests as in Figure 3.2 are done in the *Pull-down* network; see the caption of Figure 3.2 for details.

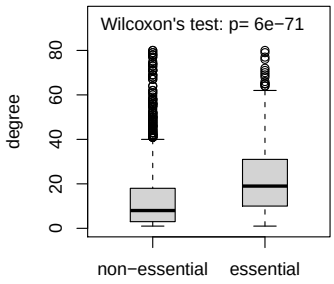


(a) correlation between degree and essentiality for proteins in BPs

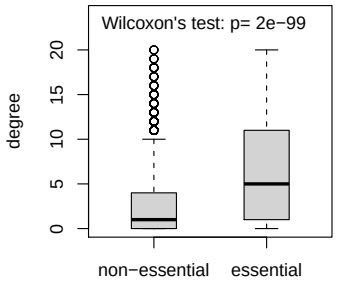
Interaction type	SRCC (p-value)
Intraprocess	0.3518 (0)
All annotated	0.2958 (0)
Interprocess	0.2109 (0)

(b) Spearman's rho rank correlation coefficient (SRCC)

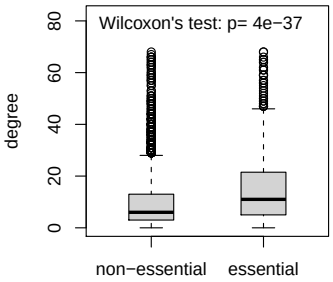
Interaction type	Partial SRCC (p-value)
Intraprocess	0.0797 (2e-6)
Interprocess	0.2906 (0)



(c) all annotated degree



(d) intraprocess degree



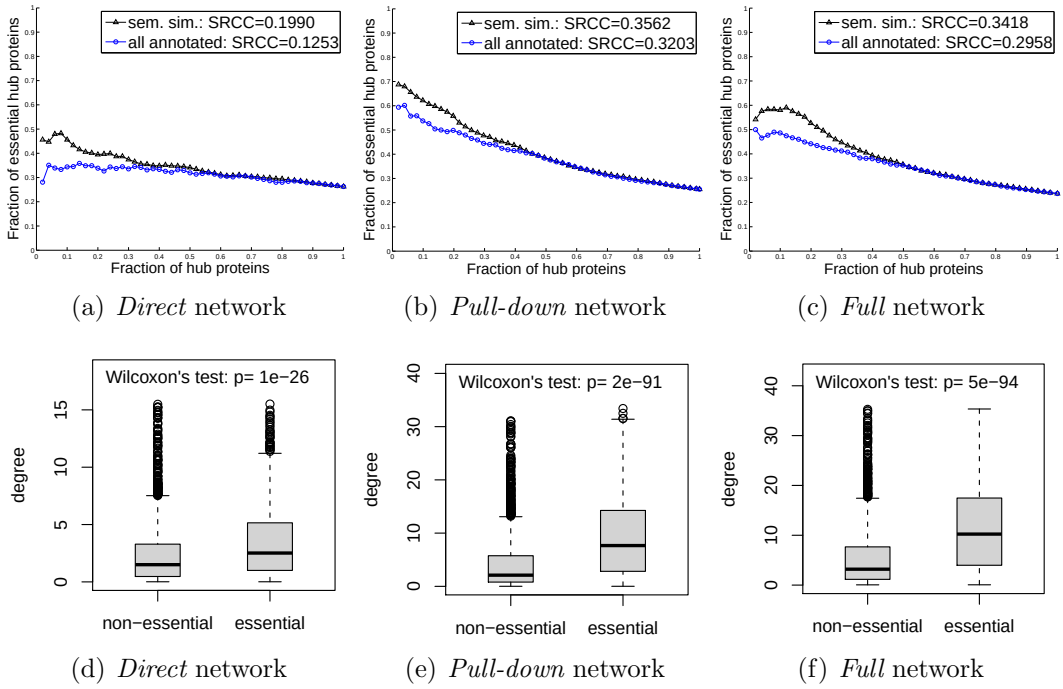
(e) interprocess degree

Figure 3.4: **The intraprocess interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Full* network, when interactions are categorized with specific Gene Ontology (GO) Biological Process (BP) terms, each of which annotates at most 50 proteins.** All tests as in Figure 3.2 are done in the *Full* network; see the caption of Figure 3.2 for details.

To further quantify the correlation between essentiality and degree, we used the Spearman's rho rank correlation coefficient (SRCC) [78], and found that the intramodular SRCC is highest for intramodular degree and near zero for the intermodular degree in the *Direct* network (Figure 3.2 (b)). We further sought to disentangle the contributions of intramodular and intermodular degree to the observed correlations, and computed partial correlations between essentiality and all annotated interactions, when controlling for intramodular and intermodular degree. For all three networks, we found that when controlling for intramodular degree, the SRCC between total degree and essentiality notably diminished, whereas when controlling for intermodular degree, the SRCC remained high (Figures 3.2, 3.3 and 3.4 (b)).

As another way of looking at the difference between intramodular and intermodular interaction degree, we compared the degree distribution of essential proteins and non-essential proteins (Figures 3.2, 3.3 and 3.4 (c)-(e)) using the Wilcoxon rank sum test. For comparing degree distributions, we included all proteins with at least one interaction; these proteins may have zero intramodular or intermodular interactions. Since the same number of proteins are considered when comparing total, intramodular, or intermodular degree (Figures 3.2, 3.3 and 3.4 (c)-(e)), the p-values given are comparable. The difference in the number of interactions between essential and non-essential proteins is much larger when only intramodular interactions are considered (Figures 3.2, 3.3 and 3.4 (d)), as compared with the case when all interactions are considered (Figures 3.2, 3.3 and 3.4 (c)) or when only intermodular interactions are considered (Figures 3.2, 3.3 and 3.4 (e)).

As an alternative to categorizing all interactions as either intermodular or intramodular, we also considered the case where interactions are weighted according to the semantic similarity [50] between the functional terms annotating the two proteins. This weight is in the range of 0 and 1 with proteins sharing highly specific functional terms getting higher scores (see section 3.2, **Materials and methods**,



(a) *Direct* network (b) *Pull-down* network (c) *Full* network

(d) *Direct* network (e) *Pull-down* network (f) *Full* network

Figure 3.5: **The semantic similarity degree is more correlated with protein essentiality than the overall interaction degree in all three networks. (a)-(c)** The fraction of essential proteins among hub proteins decreases as more proteins are considered hub proteins; this is done by adding proteins in a non-increasing order of the semantic similarity degree. For each network, the Spearman's rho rank correlation coefficient (SRCC) is computed between protein essentiality and either semantic similarity or all annotated degree; these values are boxed in each panel. **(d)-(f)** The semantic similarity weighted degree distribution of non-essential proteins is compared to that of essential proteins for the Direct, Pull-down and Full networks.

for more details). Thus, the semantic similarity between two interacting proteins is a continuous measure of the "intramodularity" of the interaction. Then, the semantic similarity degree of a protein is defined as the sum of the semantic similarity of the interactions. Across all three networks, we find that there is stronger correlation with essentiality when all interactions are weighted with semantic similarity than when they are just counted (Figure 3.5). In other words, proteins having many interactions within a similar functional context are more likely to be essential than proteins having many interactions. Thus, a range of computational analyses shows that much of the relationship between essentiality and interaction degree can be explained when considering just intraprocess interactions.

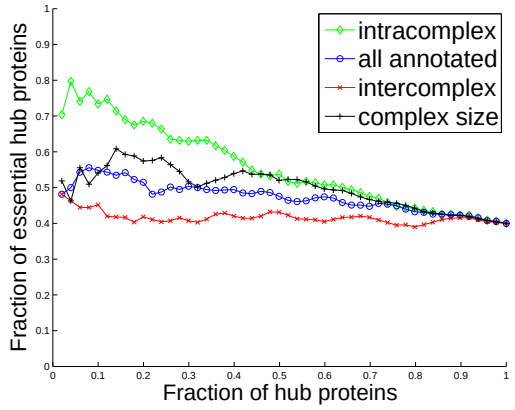
3.3.3 The correlation between intramodular degree and protein essentiality is largely due to complexes, not processes.

Having shown the strong correlation between intraprocess interaction degree and essentiality, we sought to characterize the contribution of intracomplex interactions. In particular, previously, it had been observed that essentiality is a modular property and that essential proteins tend to be clustered together within essential protein complexes [38, 96]. Thus, we hypothesized that having intracomplex physical interactions for a protein is more important for predicting its essentiality than having other types of physical interactions. That is, as we have defined them, functional modules can be comprised either of protein complexes or biological processes corresponding to Gene Ontology (GO) Biological Process (BP) terms. Here, we focus on modules derived from protein complexes. We begin by observing that complexes as a whole are enriched in essential proteins. In particular, 18.60% (or 1049/5640) of proteins are essential in the yeast genome, whereas 34.50% (or 622/1803) of proteins are es-

sential among proteins involved in any of a set of protein complexes, and 37.54% (or 598/1593) are essential when excluding the four complexes comprising the ribosome. In fact, 59.29% (or 622/1049) of all essential proteins are involved in protein complexes, even though only 31.97% (or 1803/5640) of proteins take part in our set of complexes. Thus, any conclusions arising from the analysis of protein complexes is based on the interaction properties of approximately 60% of essential proteins.

In a manner similar to how we obtained a subnetwork for GO BP terms in the previous section, we derive a subnetwork from each of the three networks where nodes represent proteins involved in any protein complex and edges represent interactions between these proteins. Repeating the analysis we performed for intraprocess vs. interprocess interactions, we found that intracomplex physical interactions are more correlated with protein essentiality than all physical interactions (Figures 3.6, 3.7 and 3.8 (a)).

It has been previously observed that there is a strong correlation between complex size and essentiality [89]. In our dataset, there is a positive correlation between complex size and the fraction of essential proteins within the complex at a complex level (SRCC (Spearman's rho rank correlation coefficient): 0.2394, p-value: 2e-6). At a protein level as well as a complex level, there is a strong correlation between essentiality and the size of a complex to which a protein belongs (black curve in Figures 3.6, 3.7 and 3.8 (a)). We found, however, this correlation is not stronger than the correlation between essentiality and intracomplex degree (black vs. green curve in Figures 3.6, 3.7 and 3.8 (a)). In our dataset, there is, however, also a positive correlation between complex size and the fraction of essential proteins within the complex at a complex level (SRCC (Spearman's rho rank correlation coefficient): 0.2394, p-value: 2e-6).

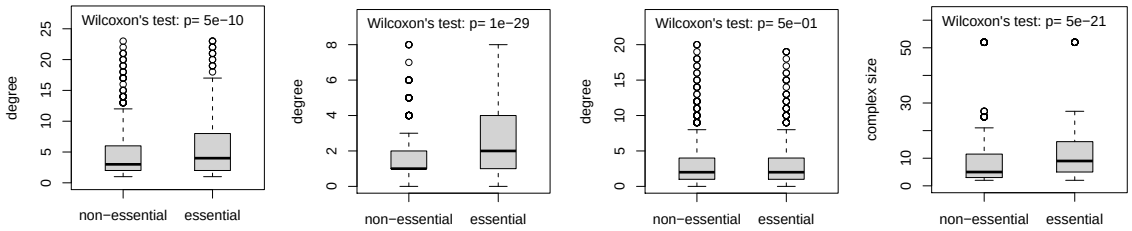


(a) correlation between degree and essentiality for proteins in complexes

Interaction type	SRCC (p-value)
Intracomplex	0.3055 (0)
All annotated	0.1664 (7e-10)
Intercomplex	0.0028 (0.9*)
Complex size	0.2541 (0)

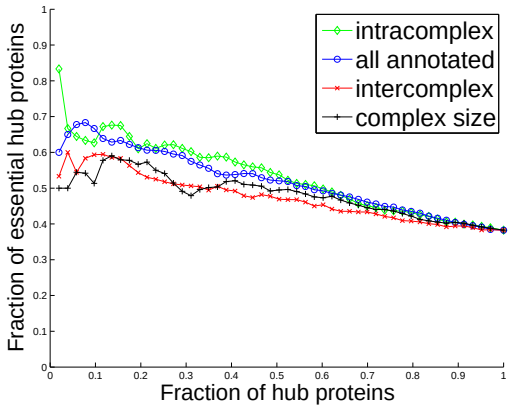
Interaction type	Partial SRCC (p-value)
Intracomplex	-0.0149 (0.6*)
Intercomplex	0.2891 (0)
Complex size	0.1220 (7e-6)

(b) Spearman's rho rank correlation coefficient (SRCC)



(c) all annotated degree (d) intracomplex degree (e) intercomplex degree (f) complex size

Figure 3.6: **The intracomplex interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Direct* network, when interactions are categorized with protein complexes.** **(a)** The fraction of essential proteins among hub proteins decreases as more proteins are considered hub proteins; this is done by adding proteins in a non-increasing order of the interaction degree or the complex size. The correlation between protein essentiality and interaction degree is shown in green (intracomplex), blue (all) and red (intercomplex). The correlation between protein essentiality and the complex size where a protein belongs is also shown (black). If a protein belongs to multiple complexes, the largest complex is considered. **(b)** Spearman's rho rank correlation coefficient (SRCC) is given to measure the correlation between degree and protein essentiality. The partial correlation is also computed between all annotated degree and essentiality when controlling for either intracomplex degree, intercomplex degree, or the size of the largest complex to which the protein belongs. * indicates a non-significant p-value > 0.05. **(c)-(f)** The degree distribution of non-essential proteins is compared to that of essential proteins within complexes for– (c) all annotated degree, (d) intracomplex degree, (e) intercomplex degree, and (f) complex size. Outliers within 2-98% are shown. The significance of the difference of the two degree distributions is measured by the Wilcoxon rank sum test. The four ribosomal complexes were not included in this analysis.

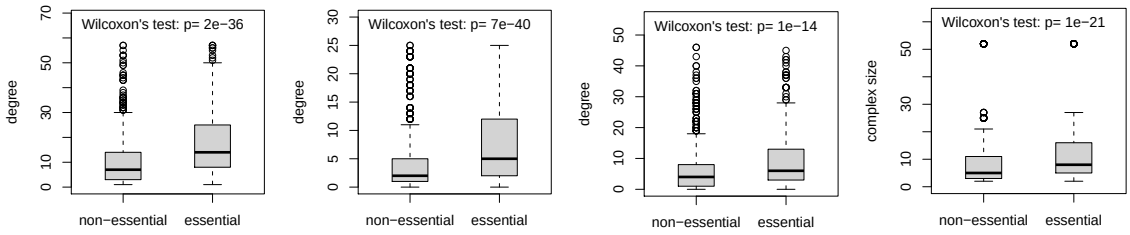


(a) correlation between degree and essentiality for proteins in complexes

Interaction type	SRCC (p-value)
Intracomplex	0.3351 (0)
All annotated	0.3191 (0)
Intercomplex	0.1935 (2e-14)
Complex size	0.2414 (0)

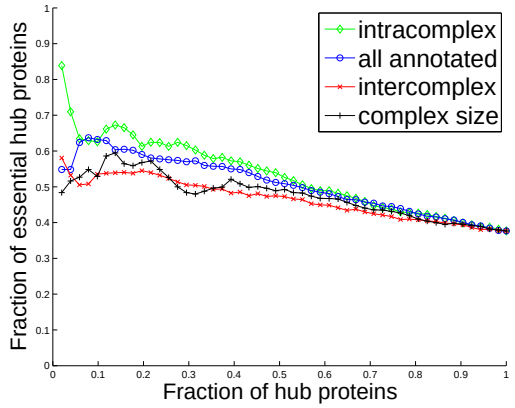
Interaction type	Partial SRCC (p-value)
Intracomplex	0.1343 (1e-7)
Intercomplex	0.2858 (0)
Complex size	0.2401 (0)

(b) Spearman's rho rank correlation coefficient (SRCC)



(c) all annotated degree (d) intracomplex degree (e) intercomplex degree (f) complex size

Figure 3.7: **The intracomplex interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Pull-down* network, when interactions are categorized with protein complexes.** All tests as in Figure 3.6 are done in the *Pull-down* network; see the caption of Figure 3.6 for details.

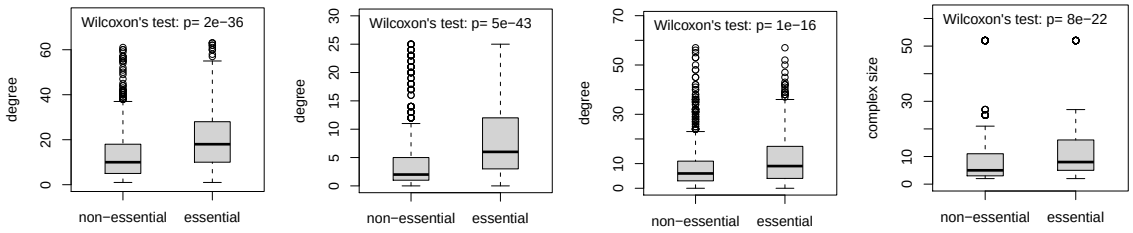


(a) correlation between degree and essentiality for proteins in complexes

Interaction type	SRCC (p-value)
Intracomplex	0.3452 (0)
All annotated	0.3162 (0)
Intercomplex	0.2060 (1e-16)
Complex size	0.2400 (0)

Interaction type	Partial SRCC (p-value)
Intracomplex	0.1393 (3e-8)
Intercomplex	0.2783 (0)
Complex size	0.2402 (0)

(b) Spearman's rho rank correlation coefficient (SRCC)



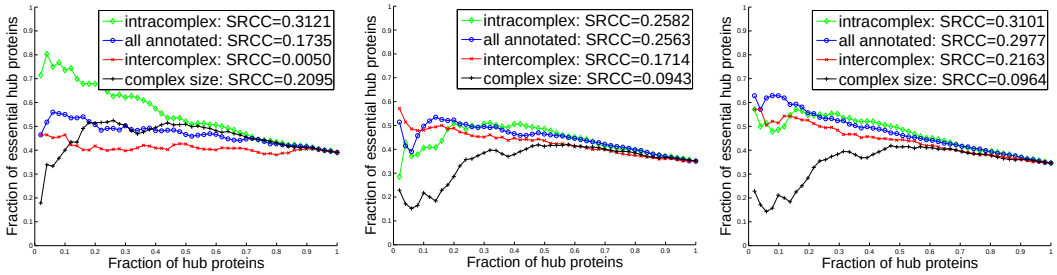
(c) all annotated degree (d) intracomplex degree (e) intercomplex degree (f) complex size

Figure 3.8: **The intracomplex interaction degree is more correlated with protein essentiality than the overall interaction degree for proteins in the *Full* network, when interactions are categorized with protein complexes** All tests as in Figure 3.6 are done in the *Full* network; see the caption of Figure 3.6 for details.

The four complexes pertaining to the ribosome tend to be removed from analysis in the literature [38, 89] because they have a relatively large number of member proteins (ranging from 32 to 79) yet have a low fraction of essential proteins ranging from 4.55% to 15.19%. Due to the total number of proteins, computational analysis can be largely affected by these four complexes. We still observe a large difference in the correlation between essentiality and intracomplex interaction degree, as well as essentiality and all interaction degree when the four ribosomal complexes are included (Figure 3.9 (a)). Including these four complexes does not affect our finding much in the *Direct* network, but it has a larger effect in the two other networks (Figures 3.9 (b),(c) vs. Figures 3.7 and 3.8 (a)). This may be due to the fact that the other networks contain indirect interactions, and non-essential proteins within large complexes can have a larger number of intracomplex interactions than essential proteins within small complexes. In particular, if there are many indirect interactions, the large complexes may have a higher chance to have many indirect intracomplex interactions than small complexes. Throughout this Chapter, we removed these ribosomal complexes for the reported complexes, unless otherwise noted.

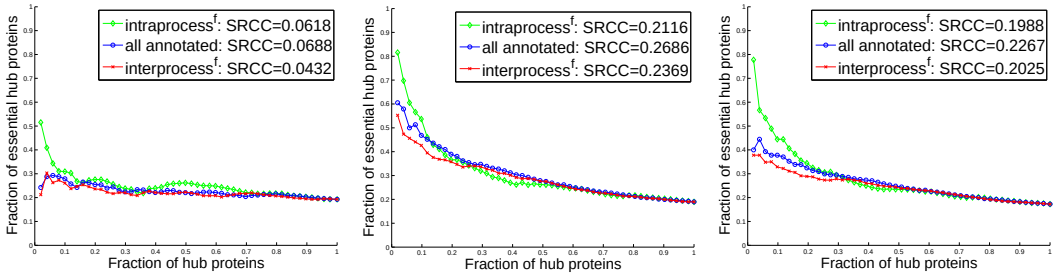
We also computed partial correlations between essentiality and all annotated interactions, when controlling for intracomplex degree, intercomplex degree, or complex size. For all three networks, we found that when controlling for intracomplex degree, the SRCC between total degree and essentiality notably diminished, whereas when controlling for intercomplex degree or complex size, the SRCC remained high (Figures 3.6, 3.7 and 3.8 (b)).

We looked at the difference in degree distribution between essential and non-essential proteins (Figures 3.6, 3.7 and 3.8 (c)-(e)). For intracomplex degree, we see the most significant difference between the mean degrees of essential and non-essential proteins (p-value: 1e-29 for the *Direct* network), and the least significant difference for intercomplex degree.



(a) *Direct* network (b) *Pull-down* network (c) *Full* network

Figure 3.9: **We show the correlations between interaction degree and essentiality for proteins in all complexes including ribosomal complexes for all three networks.** The fraction of essential proteins among hub proteins decreases as more proteins are considered hub proteins; this is done by adding proteins in a non-increasing order of the interaction degree or the complex size in all three networks. We did the same analysis as in Figures 3.6, 3.7 and 3.8 (a), respectively, but interactions are categorized with all protein complexes including the four ribosomal complexes.



(a) *Direct* network (b) *Pull-down* network (c) *Full* network

Figure 3.10: **For a set of filtered biological processes, the intraprocess interaction degree is not more correlated with protein essentiality than the overall interaction degree for proteins in all three networks.** The fraction of essential proteins among hub proteins as a function of an increasing number of proteins considered as hub proteins; this is done by adding proteins in a non-increasing order of the interaction degree. The correlation between protein essentiality and interaction degree is shown in green (intraprocess), blue (all) and red (interprocess). Spearman's rho rank correlation coefficient (SRCC) is given. For filtered biological processes, see section 3.2 **Materials and methods**.

Thus far, we have found that a stronger correlation between essentiality and intramodular degree than between essentiality and all annotated degree for all functional modules holds when we focus on either biological process or protein complex derived modules. What happens if we consider intraprocess interactions when excluding those that are intracomplex? That is, some biological processes consist of a single protein complex or several protein complexes; in these cases the intraprocess interactions are more specifically intracomplex interactions within complexes that are also annotated with the process. To focus on interactions that are not intracomplex, we filtered biological processes to remove these interactions. See section 3.2, **Materials and methods**, for more detail. Among the proteins that are annotated with any filtered biological process, 16.52% (or 424/2567) proteins are essential, which is slightly less than that when considering all proteins in the genome. In a subnetwork for the set of filtered biological processes from each of three interaction networks, there is a weaker correlation between interaction degree and essentiality as compared to the correlation for complexes, and the intraprocess degree is not a better predictor of essentiality than all annotated degree (Figure 3.10). The correlations are especially weak in the *Direct* network. One possible reason may be that these biological processes are mostly metabolic processes where proteins do not physically bind to each other, but rather through small molecules, and there are metabolic interactions within processes.

3.3.4 Essential proteins are more central within essential protein complexes.

From the observation that essential proteins tend to have more intracomplex interactions than non-essential proteins in protein complexes, we hypothesized that, for each essential protein complex, its essential proteins are more central or have a higher intracomplex degree than its non-essential proteins. We tested this hypothesis for a

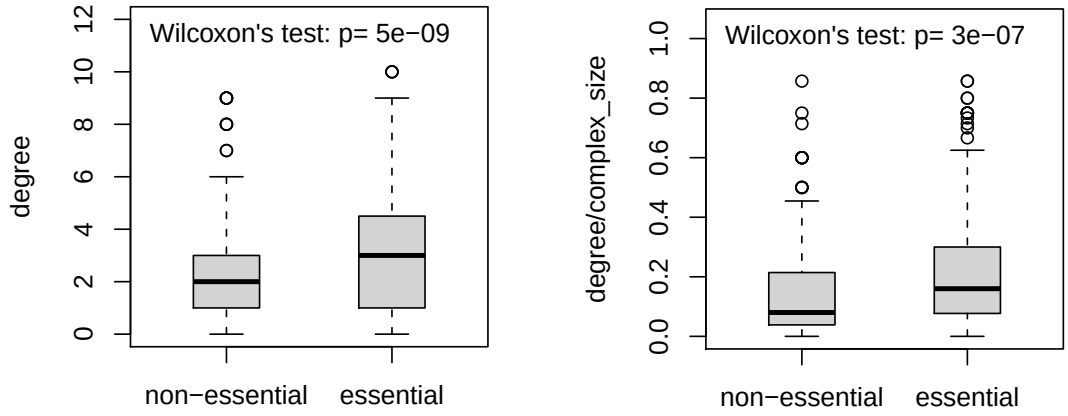
subset of protein complexes with enough member proteins and intracomplex interactions. In particular, we included a complex in our test if it had at least two essential proteins and at least two non-essential proteins, each of which has intracomplex interactions. Table 3.2 shows that complexes tend to have a higher intracomplex degree on average for essential proteins than for non-essential proteins in all three physical interaction networks. In particular, in the *Direct* network, more than 70% of complexes have essential proteins with a higher intracomplex degree on average. In the *Pull-down* or the *Full* network, the fraction of complexes with a higher degree on average for essential proteins is lower than in the *Direct* network since these networks include indirect intracomplex interactions. In fact, there are seven "clique" complexes in which every protein has intracomplex interactions with all other member proteins in both the *Pull-down* and the *Full* networks, whereas there are no such complexes in the *Direct* network. Without these clique complexes, the complex percentage goes up to 68.18% and 71.11% for the *Pull-down* and the *Full* networks, respectively.

Network	**# Tested complexes**	**# Complexes with higher degree**	**Complex percentage**
Direct	38	29	76.32%
Pull-down	51	30	58.82%
Full	52	32	61.54%

Table 3.2: **Within each essential protein complex, essential proteins tend to have a high intracomplex degree on average. Network** gives the three physical interaction networks considered. **# Tested complexes** gives the number of complexes considered; each has at least two essential proteins and at least two non-essential proteins with intracomplex interactions. **# Complexes with a higher degree** gives the number of complexes among the tested complexes where essential proteins have a higher intracomplex degree on average than non-essential proteins. **Complex percentage** gives the percentage of complexes with a higher average degree for essential proteins.

By considering each complex individually, this analysis better handles proteins involved in multiple complexes. Although we removed highly overlapping complexes (See section 3.2, **Materials and methods**), 14% (or 223/1593) of proteins in some

complex still belong to at least one other complex. Moreoever, these proteins tend to be essential; that is, the fraction of essential proteins is 53.81% (or 120/223) among proteins in more than one complex, whereas for all proteins in complexes this is 37.54% (or 598/1593). Thus, it was possible that a main cause of our earlier finding that essential proteins tend to have a high intracomplex degree is that essential proteins tend to belong to multiple complexes and intracomplex degree of an essential protein is summed over the complexes which it belongs to; however, looking at a complex one at a time alleviates this problem.

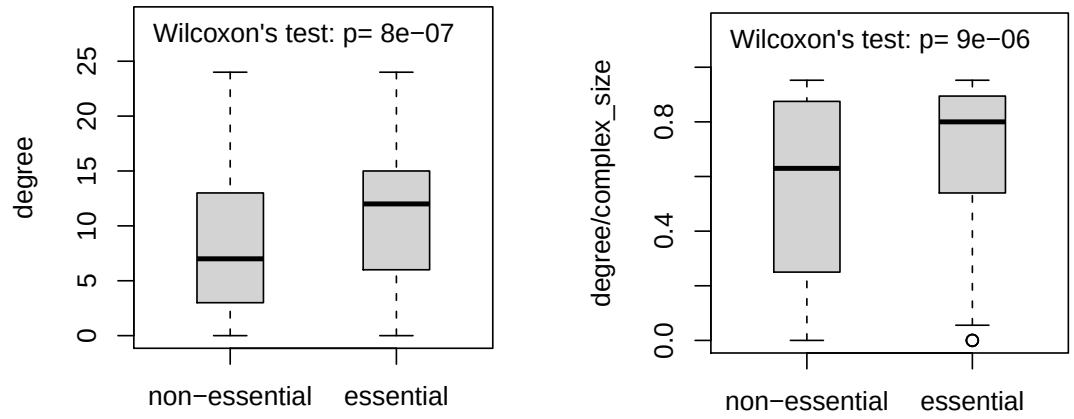


(a) Degree

(b) Degree/complex size

Figure 3.11: **Essential proteins tend to have a higher intracomplex degree than non-essential proteins within protein complexes in the *Direct* network. (a)** The intracomplex degree or **(b)** the normalized intracomplex degree of essential proteins is significantly greater than that of non-essential proteins. Only protein complexes that have at least two essential proteins and at least two non-essential proteins with intracomplex interactions are tested. Outliers within 2-98% are shown. The significance of the difference of the two degree distributions is determined by the Wilcoxon rank sum test.

As another way of addressing possible bias due to proteins in multiple complexes, for each protein in multiple complexes, we put it only in the complex with which it has a maximum intracomplex degree in the *Direct* network. Next, we compared proteins within complexes altogether. We see that there is a significant difference in degree distribution between essential and non-essential proteins (Figure 3.11 (a)). In



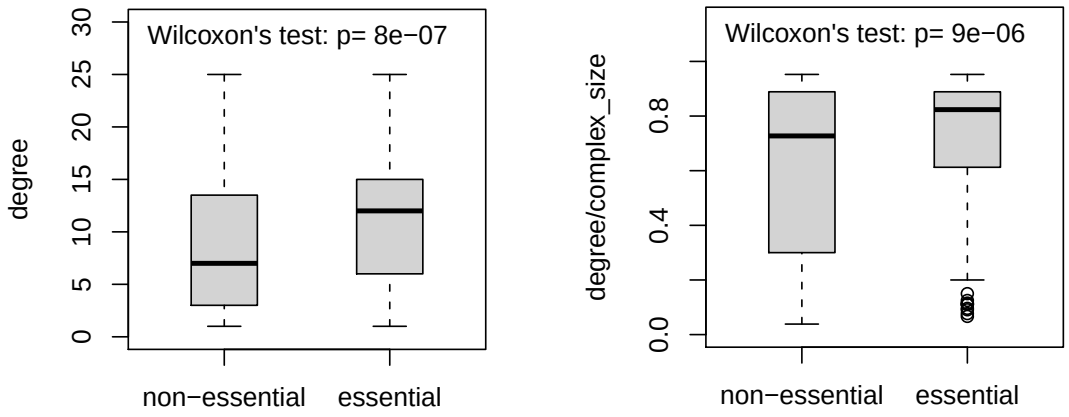
(a) Degree

(b) Degree/complex size

Figure 3.12: **Essential proteins tend to have a higher intracomplex degree than non-essential proteins within protein complexes in the *Pull-down* network. (a)** The intracomplex degree or **(b)** the normalized intracomplex degree of essential proteins is significantly greater than that of non-essential proteins. Only protein complexes that have at least two essential proteins and at least two non-essential proteins with intracomplex interactions are tested. Outliers within 2-98% are shown. The significance of the difference of the two degree distributions is determined by the Wilcoxon rank sum test.

the two other networks, we see the same results (Figures 3.12 and 3.13 (a)).

Since there is a strong correlation between the complex size and the fraction of essential proteins within the complex [89], and complex size is also correlated with the intracomplex degree of its member proteins, it is possible that the correlation between intracomplex degree and essentiality comes from the correlation between the complex size and essentiality. To address this, we normalize the degree by the complex size; that is, the normalized intracomplex degree of a protein is the number of intracomplex interactions divided by the complex size. We found that the normalized degree of essential proteins tends to be significantly greater than that of non-essential proteins (Figures 3.11, 3.12 and 3.13 (b)).



(a) Degree

(b) Degree/complex size

Figure 3.13: **Essential proteins tend to have a higher intracomplex degree than non-essential proteins within protein complexes in the *Full* network. (a)** The intracomplex degree or **(b)** the normalized intracomplex degree of essential proteins is significantly greater than that of non-essential proteins. Only protein complexes that have at least two essential proteins and at least two non-essential proteins with intracomplex interactions are tested. Outliers within 2-98% are shown. The significance of the difference of the two degree distributions is determined by the Wilcoxon rank sum test.

3.3.5 A large number of interactions are across different functional modules.

As we have just shown, essential proteins tend to have many intramodular interactions and these interactions connect essential proteins with other proteins within a functional module; presumably these interactions are important for the module to accomplish a task, particularly so in the case of protein complexes. Then, what role do intermodular interactions play in the yeast network? There are a significant number of intermodular physical interactions (i.e, interactions between different functional modules); see Table 3.3. In the subnetwork consisting of protein complexes in the *Direct* network, the fraction of intercomplex interactions is over 60% and in a subnetwork for a subset of specific GO BP terms, the fraction is almost 80%. Even when we consider GO BP terms that are quite general, the fraction of intermodular interactions is over 50%. For example, when we consider interactions within a GO BP term

which annotates up to 500 proteins in the yeast as intraprocess interaction, 50.49% of physical interactions are still interprocess interactions where the two interacting proteins participate in different biological processes. Thus, it is important to understand how physical intermodular interactions connect different functional modules in the network.

We hypothesized that proteins that are associated with a same functional module might have similar patterns of physical interactions in the sense of connecting functional modules. To test this, we can find systematic patterns of physical intermodular interactions between functional modules; this gives us an idea of how functional information "flows" through physical interactions at a modular level. From this, we can build a "module network" where nodes are modules and edges are cross-talks between modules having a significant number of intermodular interactions. We wish to relate essential functional modules to cross-talk degree in the module network. To do this, at a modular level, we need to define degree and essentiality at the module level. The interaction degree of a module is defined as the number of inferred cross-talk interactions that this module has, and the normalized interaction degree of a module is defined as the interaction degree of the module divided by the number of proteins in the module (i.e., module size). The essentiality of a module is the fraction of essential proteins within the module. If there is at least one protein in the module is essential, we say the module is essential.

In this way, a module network from each physical interaction network is generated for either protein complexes or filtered biological processes. (Table 3.4). The number of cross-talks for processes is $2.6 - 5.9$ fold higher than that for complexes because a relatively high number of interactions for processes are intermodular rather than intramodular. The intermodular percentage for processes is 86.98% whereas for complexes it is 64.34% (Table 3.3).

Modules	**# annotated interactions**	**# intramodular interactions**	**#intermodular interactions**	**Intermodular percentage**
Protein complexes	3825	1364	2461	64.34%
BP terms of size$\leq$ 50	10119	2530	7589	75.00%
BP terms of size$\leq$ 100	11757	3706	8051	68.48%
BP terms of size$\leq$ 300	13350	5780	7570	56.70%
BP terms of size$\leq$ 500	13573	6689	6884	50.72%
Filtered biological processes	4415	575	3840	86.98%

Table 3.3: **The substantial fraction of physical interactions are intermodular. Modules** gives the set of functional modules considered. These are: 1) protein complexes, 2) a subset of specific Gene Ontology (GO) Biological Process (BP) terms each of which annotate at most 50,100,300,or 500, respectively, proteins in the yeast genome, or 3) a subset of filtered biological processes (as decribed in section 3.2 **Materials and methods**). **# annotated interactions** gives the number of interactions in the subnetwork, which is generated from the *Direct* network where nodes represent proteins in the modules and edges represent between proteins in the modules. **# Intramodular interactions** gives the number of interactions in the subnetwork where the two interacting proteins belong to the same module. **# Intermodular interactions** gives the number of interactions in the subnetwork where the two interacting proteins belong to different modules. **Intermodular percentage** gives the percentage of intermodular interactions among annotated interactions.

(a) Protein complexes

Network	# cross-talks	# modules	Fraction of essential modules
Direct	194	143	0.68
Pull-down	535	242	0.60
Full	727	279	0.56

(b) Filtered biological processes

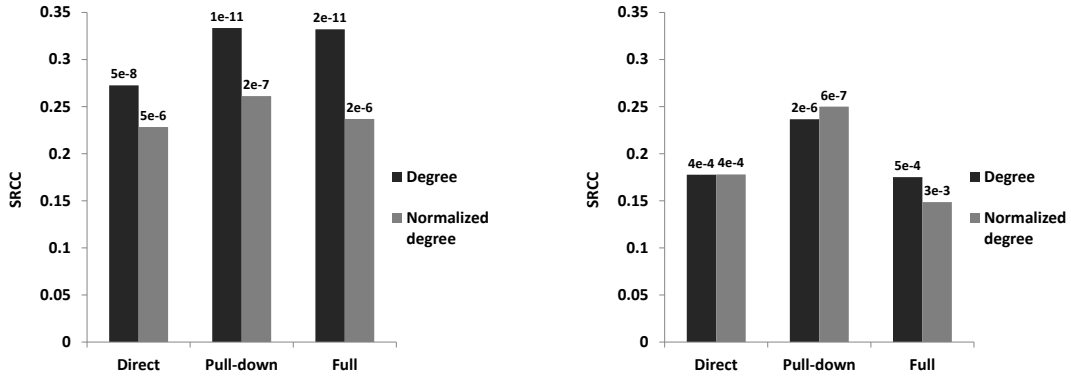
Network	# cross-talks	# modules	Fraction of essential modules
Direct	1149	307	0.79
Pull-down	1409	321	0.77
Full	2306	371	0.74

Table 3.4: **The number of cross-talks and the number of modules and the fraction of essential modules in the three module networks. Network** gives the network for which the cross-talks are inferred. **# cross-talks** gives the number of cross-talks with the likelihood ≥ 2 and at least two independent interactions. **# modules** gives the number of modules, either protein complexes or filtered biological processes, with at least one cross-talk. **Fraction of essential modules** gives the fraction of essential modules (i.e. modules having at least one essential protein) among modules with at least one cross-talk. A module network is built for (a) protein complexes or (b) filtered biological processes.

3.3.6 Essential functional modules tend to have a high cross-talk degree in the module network.

For each module network, we find that there is a correlation between cross-talk degree and the fraction of essential proteins in the module, for both protein complexes and filtered biological processes (Figure 3.14). Complexes or processes which have at least one cross-talk tend to associate with essential proteins.

Figures 3.15 and 3.16 show the module networks of protein complexes and filtered biological processes, respectively, from the *Direct* network. A functional module is essential if it has at least one essential protein. For complexes, 48% (or 189/390) of complexes have at least one essential protein but, in the complex network from the *Direct* network, 68% (or 97/143) of complexes with cross-talks have at least one essential protein (Table 3.4 (a)). Also, while 72% (or 281/391) of processes have



(a) Protein complexes

(b) Filtered biological processes

Figure 3.14: **The correlation between cross-talk degree and the fraction of essential proteins in a module is computed using SRCC (Spearman's rho rank correlation coefficient) for the module network inferred in the *Direct*, *Pull-down*, and *Full* networks for (a) protein complexes and (b) filtered biological processes.** The value above each bar gives the SRCC p-value.

at least one essential protein, for processes in the process network from the *Direct* network, 79% (or 241/307) of processes with cross-talks have at least one essential protein (Table 3.4 (b)).

When we categorize modules as having essential proteins vs. not having any essential proteins, the correlation between cross-talk degree and this binary essentiality measure is stronger than that between cross-talk degree and the fraction of essential proteins in a module.

We observed that many cross-talks occur between functional modules that are functionally related and both belong to more general biological processes. These types of cross-talk can be interpreted as intraprocess interactions at a broader level of functional similarity. To see if essential functional modules have many cross-talks with functional modules in representing truly different biological processes, we ignored cross-talks between functional modules if they are annotated with the same Gene Ontology (GO) Biological Process (BP) term among biological expert selected terms that annotate at most 500 proteins [58]. Given the selected terms, a functional

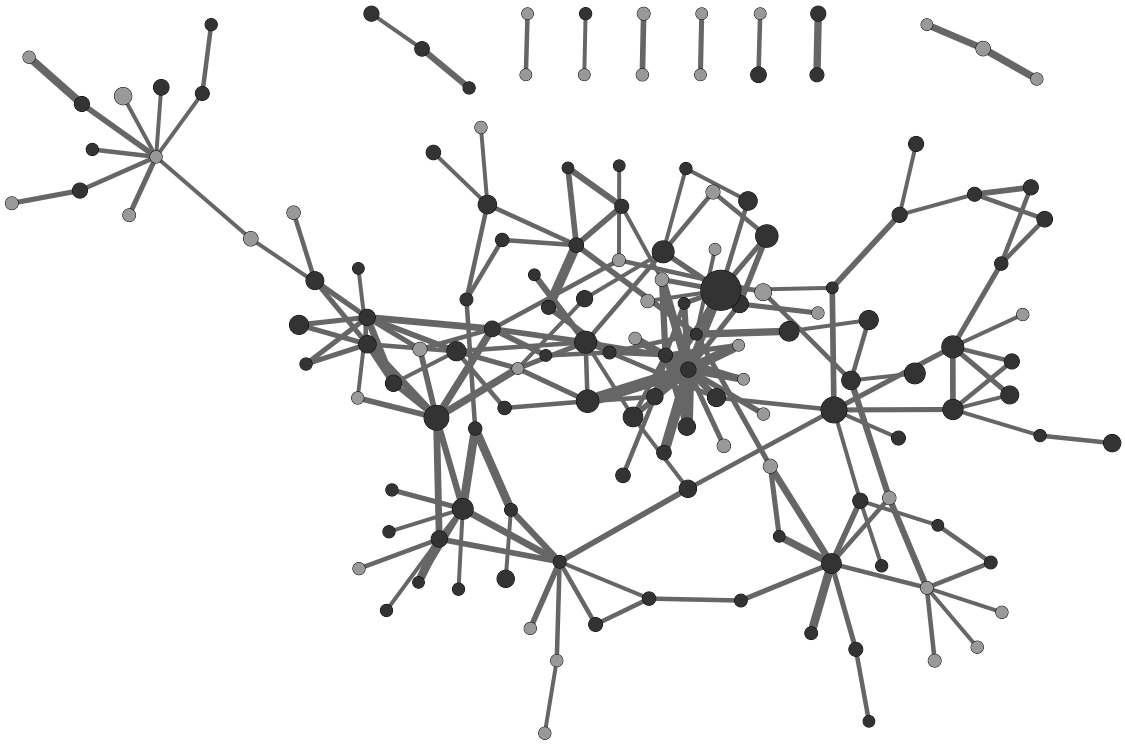


Figure 3.15: **The module network for protein complexes from the *Direct* interaction network.** Each node represents a complex and an edge represents an uncovered cross-talk between complexes. Node size is proportional to the number of proteins that belong to the corresponding complex and node color represents the essentiality of the complex; that is, the color is dark grey if a complex has at least one essential protein, and light grey otherwise. Also, the edge width is proportional to the number of interactions between two complexes. There are 97 essential and 46 non-essential complexes with cross-talks, and 92 essential and 155 non-essential complexes without cross-talks (the latter are not shown).

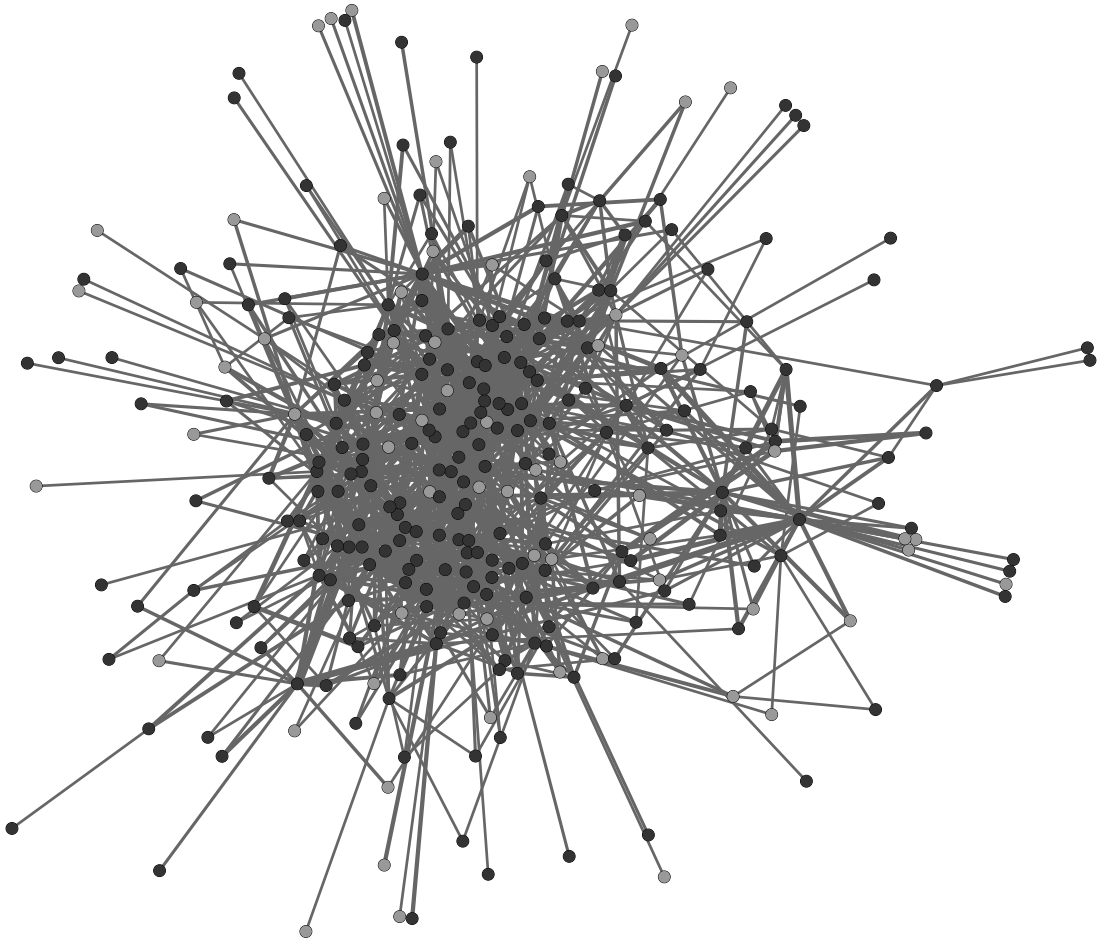
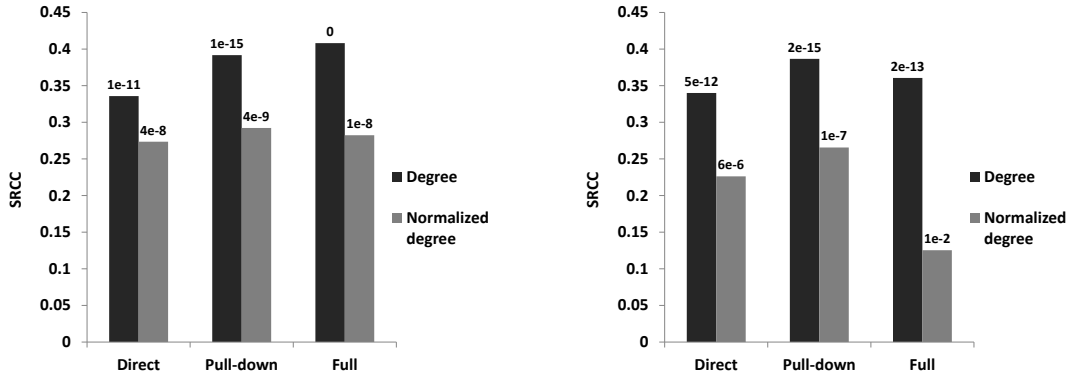


Figure 3.16: **The module network for filtered biological processes from the *Direct* interaction network.** Each node represents a filtered biological process and an edge represents an uncovered cross-talk between processes. Node size is proportional to the number of proteins that belong to the corresponding process and node color represents the essentiality of the process; that is, the color is dark grey if a process has at least one essential protein, and light grey otherwise. Also, the edge width is proportional to the number of interactions between two processes. There are 241 essential and 66 non-essential processes with cross-talks, and 40 essential and 44 non-essential processes without cross-talks (the latter are not shown).



(a) Protein complexes

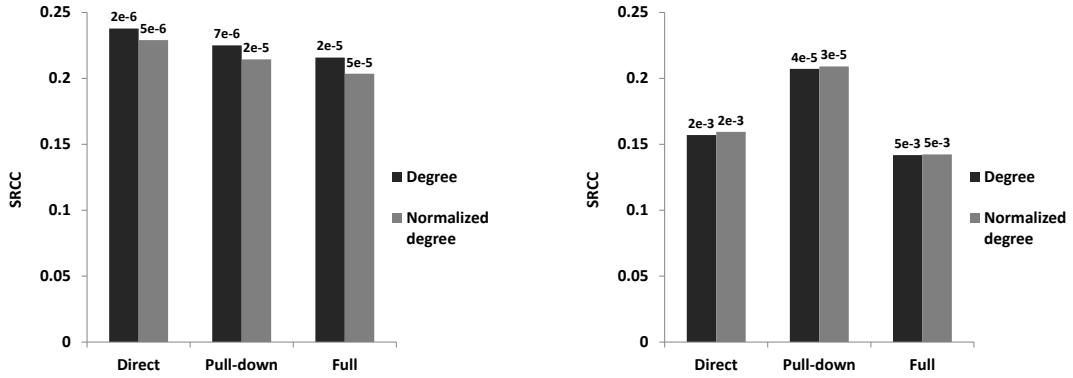
(b) Filtered biological processes

Figure 3.17: **The correlation between cross-talk degree and binary essentiality for modules is computed using SRCC (Spearman's rho rank correlation coefficient) for the module network inferred in the *Direct*, *Pull-down*, and *Full* networks for (a) protein complexes and (b) Filtered biological processes.** The value above each bar gives the SRCC p-value. The binary essentiality for a module is defined as 1 if the module has at least one essential protein, and 0 otherwise.

module is annotated with one of these terms if $\geq 70\%$ of proteins in the module are annotated with it. Without such cross-talks, essential functional modules are still correlated with cross-talk degree (Figure 3.18).

3.4 Discussion and conclusions

We incorporated functional information into network topology analysis in order to better understand protein essentiality. Using this functional information, physical interactions were categorized into intramodular interactions within functional modules and intermodular interactions across functional modules. Previously, it had been shown in several analyses that hub proteins in physical interaction networks tend to be essential. Our analysis revealed that essential proteins tend to have many intramodular interactions, and these are more predictive of essentiality just any interactions. In addition, essential proteins tend to be organized in a modular manner and interact with each other within essential functional modules, and especially within protein



(a) Protein complexes

(b) Filtered biological processes

Figure 3.18: **The correlation between cross-talk degree and the fraction of essential proteins in a module after removing functionally similar cross-talks is computed using SRCC (Spearman's rho rank correlation coefficient) for the module network inferred in the *Direct*, *Pull-down*, and *Full* networks for (a) protein complexes and (b) Filtered biological processes.** We removed cross-talks between modules that are annotated with a shared general biological process. The value above each bar gives the SRCC p-value.

complexes. We found that if biological processes annotate essential proteins that are densely interacting with each other, these essential proteins tend to overlap complexes. Moreover, if we remove the association of proteins that belong to complexes from a process, then intraprocess interactions do not correlate with essentiality any better than all interactions for the rest of proteins associated with the process. Further, within essential protein complexes, we showed it is more likely that essential proteins have a higher intracomplex degree than non-essential proteins.

Therefore, protein complexes, which include many essential proteins and thereby many intracomplex interactions within complexes, to a large extend, explain the correlation between physical interactions and protein essentiality. However, there are still a significant number of intermodular interactions. We looked at interactions at a modular level and found systematic relationships between functional modules. We found that essential functional modules tend to have many cross-talks with other functional modules. From this, we showed that there is correlation with essentiality

both at a protein level and at a modular level. Further, we observed that functionally related modules are likely to interconnect to each other, thereby revealing the hierarchical structure of physical interaction networks.

There are mainly two types of physical interactions– direct or binary interactions and indirect interactions indicating co-membership of proteins in complexes. To ensure that our findings were not biased towards any of the experimental techniques for detecting physical interactions, we tested our hypotheses on three different networks– Direct, Pull-down, and Full. We saw consistency in our results regardless of what networks are used for testing.

Overall this work has advanced our understanding of the relationship between essentiality and network topology. We have shown the importance of intramodular interactions, especially intracomplex interactions, and demonstrated that essential modules tend to have a higher cross-talk degree than non-essential modules. In the future, it would be interesting to characterize the network properties of essential proteins that are not central in protein physical interaction networks. Based on our current findings, we can speculate that some of these proteins are important for the functioning of specific essential modules, but perhaps their interactions with other proteins in the module may be better represented with other types of interactions (e.g., metabolic, or regulatory). Further research is necessary to uncover whether this is indeed the case.

Chapter 4

Conclusions

In this thesis, we developed computational methods for analyzing protein-protein physical interaction networks in order to better understand protein function and cellular organization. In Chapter 2, we attempted to characterize how to best utilize clustering approaches for protein functional analysis. We demonstrated that the performances of clustering algorithms in recapitulating functional modules depend strongly upon topological features of networks, and that these features should guide algorithm choice in real-world applications. We evaluated six diverse network clustering algorithms on *S. cerevisiae*. We found significant differences in the performances of these algorithms when run on the same network, and a dependence of these performances based upon the topological features of the underlying networks. Moreover, for the specific task of function prediction, surprisingly, our analysis uncovered that for well-annotated genomes such as *S. cerevisiae*, a commonly-used network clustering approach is less accurate than a very simple, local, non-clustering guilt-by-association approach. Finally, since computational biologists continue to develop novel network clustering algorithms, our work established guidelines for justifying and evaluating these approaches for interaction networks.

In Chapter 3, we looked at protein topology and its effect on functioning in a

slightly different setting. We focused on the centrality-lethality rule at different scales of organization. It was previously known that protein essentiality is correlated with physical interaction degree but we found more specifically that it is correlated with physical intra-complex or intra-process interaction degree. From this, it is more likely that essential proteins play an important role in connecting proteins within a complex or process rather than in mediating different functional modules. Within an essential complex, we found that essential proteins tend to have a larger number of intra-complex interactions than non-essential proteins. Not only do essential proteins tend to have many interactions within complexes or processes, but also essential complexes and processes tend to have higher cross-talk degrees in a module-level network. In other words, we see strong evidence that centrality-lethality rule is true both at a protein level and in a large scale module level.

To conclude, this thesis has investigated the relationship between the topological features of cellular networks and the overall functioning of the cell.

Bibliography

[1] Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics 22*, 8 (2006), 1021.

[2] Aittokallio, T., and Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics 7*, 3 (2006), 243–255.

[3] Alm, E., and Arkin, A. P. Biological networks. *Current Opinion in Structural Biology 13*, 2 (2003), 193–202.

[4] Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., and Kanaya, S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics 7* (2006), 207.

[5] Arnau, V., Mars, S., and Marin, I. Iterative cluster analysis of protein interaction data. *Bioinformatics 21* (2005), 364–378.

[6] 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.

[7] Asthana, S., King, O., Gibbons, F., and Roth, F. Predicting protein complex membership using probabilistic network reliability. *Genome Res. 14* (2004), 1170–1175.

[8] Bader, G. D., and Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics 4* (2003), 2.

[9] Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology 22*, 1 (2004), 78–85.

[10] Banks, E., Nabieva, E., Chazelle, B., and Singh, M. Organization of physical interactomes as uncovered by network schemas. *PLoS Computational Biology 4*, 10 (2008), e1000203.

[11] Barabási, A. L., and Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet. 5* (2004), 101.

[12] Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J., Ou, J., San Luis, B., Bandyopadhyay, S., Hibbs, M., Hess, D., Gingras, A., Bader, G., Troyanskaya, O., Brown, G., Andrews, B., Boone, C., and Myers, C. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods 7*, 12 (2010), 1017–1025.

[13] Batada, N. N., Hurst, L., and Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Computational Biology 2*, 7 (2006), e88.

[14] Blatt, M., Wiseman, S., and Domany, E. Superparamagnetic clustering of data. *Physical Review Letters 76*, 18 (1996), 3251–3254.

[15] Boone, C., Bussey, H., and Andrews, B. Exploring genetic interactions and networks with yeast. *Nature Review Genetics 8* (2007), 437–449.

[16] Brady, A., Maxwell, K., Daniels, N., and Cowen, L. Fault tolerance in protein interaction networks: stable bipartite subgraphs and redundant pathways. *PLos One 4* (2009), e5364.

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

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

[19] Chen, J., and Yuan, B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics 22* (2006), 2283–2290.

[20] Colak, R., Hormozdiari., F., Moser., F., Schönhuth, A., Holman, J., Ester, M., and Sahinalp, S. Dense graphlet statistics of protein interaction and random networks. In *Pacific Symposium on Biocomputing* (2009), vol. 14, pp. 178–189.

[21] Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C. P., Weissman, J. S., and Krogan, N. J. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & Cellular Proteomics 6* (2007), 439–450.

[22] Datta, S., and Datta, S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics 7* (2006), 397.

[23] Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology 10* (2003), 947.

[24] Dezso, Z., Oltvai, Z., and Barabasi, A.-L. Bioinformatics analysis of experimentally determined protein complexes in yeast. *Genome Research 13* (2003), 2450–2454.

[25] Dunn, R., Dudbridge, F., and Sanderson, C. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics 6* (2005), 39.

[26] Enright, A. J., Dongen, S. V., and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res. 30*, 7 (2002), 1575.

[27] Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. Large-scale mapping of human proteinprotein interactions by mass spectrometry. *Molecular Systems Biology 3* (2007), 89.

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

[29] Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B.,

Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault1, J.-A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.-P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rossé, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. Protein interaction mapping: A drosophila case study. *Genome Research 15* (2005), 376–384.

[30] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. Evolutionary rate in the protein interaction network. *Science 296* (2002), 750–752.

[31] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature 440*, 7084 (2006), 631.

[32] Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster,

B., Neubauer, G., and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415*, 6868 (2002), 141.

[33] Georgii, E., Dietmann, S., Uno, T., Pagel, P., and Tsuda, K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics 25*, 7 (2009), 933–940.

[34] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., Bakkoury, M. E., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., ldener, U. G., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kötter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., yun Wang, C., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. Functional profiling of the saccharomyces cerevisiae genome. *Nature 418*, 25 (2002), 387–391.

[35] 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.

[36] Hahn, M. W., and Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution 22*, 4 (2005), 803–806.

[37] Handl, J., Knowles, J., and Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics 21*, 15 (2005), 3201–3212.

[38] Hart, G. T., Lee, I., and Marcotte, E. M. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics 8* (2007), 236.

[39] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. From molecular to modular cell biology. *Nature 402*, 6761 Suppl (1999).

[40] He, X., and Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genetics 2*, 6 (2006), e88.

[41] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D., and Tyers, M. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature 415*, 6868 (2002), 180.

[42] 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 98*, 8 (2001), 4569–4574.

[43] 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*, 3 (2000), 1143.

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

[45] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. The large-scale organization of metabolic networks. *Nature 407*, 6804 (2000), 651.

[46] Jiang, P., and Singh, M. Spici: a fast clustering algorithm for large biological networks. *Bioinformatics 26*, 8 (2010), 1105–1111.

[47] Kelley, R., and Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology 23* (2005), 561–566.

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

[49] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine,

A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., Onge, P. S., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature 440*, 7084 (2006), 637.

[50] Lin, D. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning* (San Francisco, CA, USA, 1998), Morgan Kaufmann, pp. 296–304.

[51] Loewenstein, Y., Portugaly, E., Fromer, M., and Linial, M. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics 24* (2008), i41–i49.

[52] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing 8* (2003), 601–612.

[53] Luo, F., Yang, Y., Chen, C., Chang, R., Zhou, J., and Scheuermann, R. Modular organization of protein interaction networks. *Bioinformatics 23* (2007), 207–214.

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

[55] Mayer, M. L., and Hieter, P. Protein networks–built by association. *Nature Biotechnology 18* (2000), 1242–1243.

[56] Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., Warfsmann, J., and Ruepp, A. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res. 32(Database issue)* (2004), D41–D44.

[57] 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.

[58] Myers, C. L., Barrett, D. R., Hibbs, M. A., Huttenhower, C., and Troyanskaya, O. G. Finding function: evaluation methods for functional genomic data. *BMC Genomics 7* (2006), 187.

[59] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics 21 Suppl. 1* (2005), i302–i310.

[60] Navlakha, S., Schatz, M., and Kingsford, C. Revealing biological modules via graph summarization. *Journal of Computational Biology 16*, 2 (2009), 253–264.

[61] Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA 103*, 23 (2006), 8577.

[62] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E 64* (2001), 026118.

[63] Palla, G., Derényi, I., Farkas, I. J., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 7043 (2005), 814.

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

[65] Poyatos, J., and Hurst, L. How biologically relevant are interaction-based modules in protein networks? *Genome Biol. 5* (2004), R93.

[66] Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research 37* (2009), 825–831.

[67] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA 101*, 9 (2004), 2658–2663.

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

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

[70] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. Towards a

proteome-scale map of the human proteinprotein interaction network. *Nature 437* (2005), 1173–1178.

[71] Samanta, M., and Liang, S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA. 100* (2003), 12579–12583.

[72] 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.

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

[74] The Saccharomyces Genome Database (SGD). http://www.yeastgenome.org.

[75] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA 102*, 6 (2005), 1974.

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

[77] Song, J., and Singh, M. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics 25*, 23 (2009), 3143–3150.

[78] Spearman, C. The proof and measurement of association between two things. *American Journal of Psychology 15*, 1 (1904), 72–101.

[79] Spirin, V., and Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA 100*, 21 (2003), 12123.

[80] Stanyon, C. A., Liu, G., Mangiola, B. A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., and Finley, R. L. A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biology 5* (2004), R96.

[81] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res. 34(Database Issue)* (2006), D535–D539.

[82] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. A human protein-protein interaction network: A resource for annotating the proteome. *Cell 122* (2005), 957–968.

[83] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., and Knight, J. R. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature 403*, 6770 (2000), 623.

[84] Vidal, M., Cusick, M. E., and Barabási, A.-L. Interactome networks and humun disease. *Cell 144* (2011), 986–998.

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

[86] von Mering, C., Zdobnov, E., Tsoka, S., Ciccarelli, 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.

[87] Wagner, A., and Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B 268* (2001), 1803–1810.

[88] Wang, C., Ding, C., Yang, Q., and Holbrook, S. R. Consistent dissection of the protein interaction network by combining global and local metrics. *Genome Biol. 8* (2007), R271.

[89] Wang, H., Kakaradov, B., Collins, S. R., Karotki, L., Fiedler, D., Shales, M., Shokat, K. M., Walther, T. C., Krogan, N. J., and Koller, D. A complex-based reconstruction of the saccharomyces cerevisiae interactome. *Molecular and Cellular Proteomics 8* (2009), 1361–1381.

[90] Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., Bakkoury, M. E., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J. L., andChristopher J. Roberts, L. R., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Véronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., and Davis, R. W. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science 285* (1999), 901–906.

[91] Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H., and Gerstein, M. Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem. 73* (2004), 1051–1087.

[92] Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. High-quality binary protein interaction map of the yeast interactome network. *Science 101*, 16 (2008), 5934–5939.

[93] Yu, H., Greenbaum, D., Lu, H. X., Zhu, X., and Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends in Genetics 20*, 6 (2004), 227–231.

[94] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology 3*, 4 (2007), e59.

[95] Zhu, X., Gerstein, M., and Snyder, M. Getting connected: analysis and principles of biological networks. *Genes development 21*, 9 (2007), 1010–1024.

[96] Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Computational Biology 4*, 8 (2008), e1000140.