Metadata of the chapter that will be visualized online

Chapter Title	Protein Function Prediction Using Protein–Protein Interaction Networks	
Chapter Sub-Title		
Chapter CopyRight - Year	Springer Science+Business Media B.V. 2011 (This will be the copyright line in the final PDF)	
Book Name	Protein Function Prediction for Omics Era	
Corresponding Author	Family Name	**Chua**
	Particle	
	Given Name	**Hon Nian**
	Suffix	
	Division	
	Organization	Institute for Infocomm Research
	Address	Singapore, Singapore, 138632
	Email	
Author	Family Name	**Liu**
	Particle	
	Given Name	**Guimei**
	Suffix	
	Division	School of Computing
	Organization	National University of Singapore
	Address	Singapore, Singapore
	Email	
Author	Family Name	**Wong**
	Particle	
	Given Name	**Limsoon**
	Suffix	
	Division	School of Computing
	Organization	National University of Singapore
	Address	Singapore, Singapore
	Email	

Abstract

Proteins perform biological functions by participating in a large number of interactions, ranging from transient interactions in signaling pathways to permanent interactions within stable complexes. Studies have shown that the immediate interaction neighborhood of a protein can be used to infer its functions. While using only such direct interactions limits prediction coverage, extending the interaction neighborhood to include indirect interaction partners reduces precision significantly, making functional inference unviable. In a series of studies, we find that the extent of partner-sharing between two non-interacting proteins makes a good estimator for their co-participation in similar function. This allows us to include indirect interactions in network-based functional inference with little compromise in precision. We also extend this idea to the related problems of protein complex prediction and interaction data cleansing.

Protein Function Prediction Using Protein–Protein Interaction Networks

AQ1

Hon Nian Chua, Guimei Liu, and Limsoon Wong

Abstract Proteins perform biological functions by participating in a large number of interactions, ranging from transient interactions in signaling pathways to permanent interactions within stable complexes. Studies have shown that the immediate interaction neighborhood of a protein can be used to infer its functions. While using only such direct interactions limits prediction coverage, extending the interaction neighborhood to include indirect interaction partners reduces precision significantly, making functional inference unviable. In a series of studies, we find that the extent of partner-sharing between two non-interacting proteins makes a good estimator for their co-participation in similar function. This allows us to include indirect interactions in network-based functional inference with little compromise in precision. We also extend this idea to the related problems of protein complex prediction and interaction data cleansing.

Introduction

AQ2

Proteins are important building blocks that contribute to key processes within cells. The elucidation of mechanisms underlying protein functionality is an active and important pursuit in biology, and remains a challenging task. Unlike protein sequences or protein-protein interactions, there is currently no systematic experimental technique that can characterize the functions of proteins in a high-throughput fashion. With various sources of biological data being made available at an unprecedented rate, efforts intensify for computational methods that can tap into this growing pool of information for reliable functional characterization of proteins. In this chapter, we summarize our efforts towards this area of research. We will describe our work on the use of protein–protein interactions for computational protein function prediction, protein complex discovery, and improving the reliability of protein–protein interactions.

AQ3

H.N. Chua (✉)
Institute for Infocomm Research, Singapore, Singapore 138632

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_13, © Springer Science+Business Media B.V. 2011

Protein–Protein Interactions

Protein–protein interactions generally refer to associations between protein molecules, which include direct physical binding and genetic interactions, amongst other definitions.

Physical Interactions

Physical binding between proteins can be detected in a high-throughput manner using a variety of assays such as co-immunoprecipitation, tandem affinity purification [1, 2], and two-hybrid systems [3–5]. In yeast two-hybrid assays, the GAL4 transcriptional activator is split into two fragments, one containing the binding domain and the other containing the activating domain. To detect an interaction (or lack thereof) between two proteins, one protein is fused to the fragment containing the binding domain (also referred to as the bait) while the other protein is fused to the other fragment (the prey). An interaction between the bait and prey proteins indirectly connects the two fragments of the transcription factor, bringing the activating domain close to the transcription start site, and results in the expression of the downstream reporter gene. In co-immunoprecipitation experiments, proteins that are suspected to interact directly or indirectly with a protein of interest are isolated together with the protein using an antibody, and subsequently identified using western blot. Tandem affinity purification involves creating fusion proteins with one end that can be bound to beads coated with a specific antibody. The modified proteins, along with the unknown proteins that they bind, are isolated over two rounds of purification and identified. The use of fusion proteins makes this technique suitable for systematic genome-wide studies [2, 6]. Datasets of large numbers of physical protein–protein interactions have been experimentally derived using two hybrid systems for a number of species, particularly for the model organisms *Saccharomyces cerevisiae* (budding yeast), *Drosophila melanogaster* (fruit fly) and *Caenorhabditis elegans* (nematode).

Genetic Interactions

Genetic interactions, on the other hand, capture functional dependency between genes from observations of phenotypes exhibited upon two or more gene deletions. The departure of observed phenotypes (usually cell viability) of double-deletion mutants from that expected of the two independent genes (based on the phenotypes of each single-deletion mutant) is used to identify such interactions. While there have been attempts to reconcile such observations with biological models such as parallel or serial pathways, these are insufficient to explain the complex relationships between genes that are reflected in these experiments. Nonetheless, genetic interactions provide great insight into the functional organization of gene products. Positive genetic interactions are often associated with proteins within complexes, while negative genetic interactions often capture redundancy between pathways [7]. Several large-scale genetic interaction experiments have been conducted for yeast [8–10] using the Synthetic Genetic Array technology [8], which allows systematic, unbiased screening for genetic relationships of a large number of array genes against a query gene in a high throughput fashion. Systematic screening for genetic interactions between essential genes is also possible using hypomorphic alleles [10]. The BioGRID database [11] is one of the largest collections of published protein–protein interactions, both physical and genetic, making it a valuable resource for researchers who are interested in studying protein–protein interactions.

Function Prediction Using Protein–Protein Interactions

A protein's functional behavior is intuitively related to its physical interactions with other proteins. Genetic interactions, on the other hand, capture functional dependencies between genes (and the proteins they encode for), such as serial genes in a biosynthesis pathway, or genes in parallel transport pathways. Hence protein–protein interactions potentially enrich for information about functional relationships between proteins that may not be obvious or detectable from other genomic data such as primary or higher level sequence structure.

Many computational approaches have been developed to utilize protein interactions for the functional characterization of proteins. One of the earliest approaches is the neighbor counting method proposed by [12]. The simple method, which assigns a protein with the function that is annotated most frequently to its interaction partners, was applied to a large-scale physical interaction dataset generated from yeast two-hybrid experiments, and performs reasonably well. The approach, however, did not take into account the background frequency of different function annotations. The mere observation of a very common functional annotation assigned to the majority of a group of proteins does not necessary suggests enrichment unless its prior probability is taken into account. Hishigaki and colleagues addressed this limitation by using the Chi-square statistic to estimate the enrichment of functional annotations in each protein's interaction neighborhood [13].

An obvious limitation in both the Neighbor Counting and Chi-square approaches is the inability to infer functional annotations to a protein that do not interact with annotated proteins. These approaches will also be biased in making inference when the majority of the proteins in the interaction neighborhood of a protein are not annotated. To overcome these limitations, some methods cleverly made assumptions along the lines that the "correct" set of functional annotations to unannotated proteins in an interacting network is the one in which functional association between adjacent proteins is best upheld. While it is unfeasible to find such a best solution in the vast space of possible configurations, many stochastic inference techniques can be used to find a reasonably good solution. Such "global" inference methods also have the advantage of being more resilient against errors in functional annotations and in the interaction network.

One such "global" inference approaches is the Markov Random Field method described in [14], which proposes that the probability of a set of inferred annotations to proteins in an interaction network is inversely related to the amount of annotation inconsistencies between interacting proteins. This probability is formally defined for each functional annotation to be a function of its prior probabilities, the number of functionally associated interactions, and the number of functionally unassociated interactions. A Gibbs sampler is then used to find a near optimal set of annotation assignments that maximizes the probability. A similar approach is used in [15]. Vazquez et al. also proposed another optimization method based on Simulated Annealing [16].

Indirect Association of Protein Function

Functional Association Between Indirect Neighbors

In 2006, we proposed the hypothesis of indirect association of protein function [17]. The motivation behind the hypothesis is the observation that many proteins do not share similar function with any of their interaction partners. In the study, we investigated the functional relationships between interacting proteins in the *Saccharomyces cerevisiae* (bakers' yeast) genome using physical and genetic interactions deposited in the BioGRID [11], as well as FunCat functional annotations from MIPS [18]. We observed that there are proteins that do not share any functional annotation with their immediate interaction partners (i.e., level-1 neighbours) and yet share some function similarity with the interaction partners of their immediate partners (i.e., level-2 neighbours). Two examples of such proteins are shown in Fig. 1. Among 4162 annotated yeast proteins in the dataset studied, only 48.0% share some function with its level-1 neighbours. 22.7% of the annotated proteins shared functional annotations with their level-2 neighbours but not their level-1 neighbours. Less than 2% of the annotated proteins share functions with level-1 neighbours without sharing functions with their level-2 neighbours. This suggested that many existing approaches to functional inference based on protein–protein interaction, whether in a local or global fashion, may be somewhat limited by making only assumptions of functional linkage between directly interacting proteins. Local inference methods will not be able annotate a protein with a function that is not observed in its direct neighbors. Global inference methods may erroneously propagate function in an indiscriminative way.

The observation left us pondering if it is possible to make predictions for more proteins by explicitly taking into account the functional annotations of the level-2 neighbors of proteins. Hishigaki et al. [13] studied the use of larger interaction neighborhoods (which they termed *n*-neighbouring proteins, analogous to our definition of *n*-level neighbors) by using their Chi-square based method on the functional classification used in the Yeast Proteome Database (YPD), and concluded that the value of *n* for the best prediction performance is small (1 for cellular role and subcellular localization, and 2 for biochemical function). Such observation is

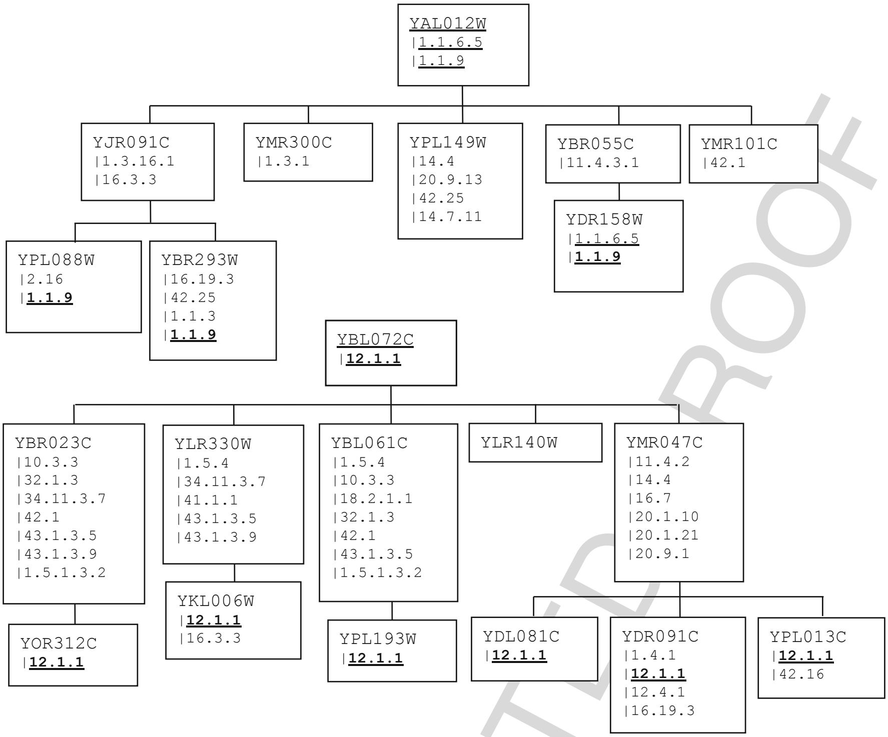


Fig. 1 Two examples of proteins that do not share functional annotations with their direct interaction neighbor, but share functional annotations with their indirect (level-2) neighbors (indirect neighbors that share no annotation are not shown). Figure from [17]

expected as we expect functional relationship to diminish with the interaction distance. The number of neighboring proteins also often increases quickly with the size of the neighborhood, and the predictive powers of the closer (and more functionally related) neighbors tend to be diminished as a result. Moreover, errors in the lower level interaction neighborhood will spill over and propagate to the higher levels, resulting in more errors introduced in each level. Hence the number of functionally irrelevant interactions is expected to be higher when more levels of interactions are used.

Estimating Function Similarity Between Interacting Proteins

To be able make use of the indirect neighbors for increasing prediction coverage without severely affecting precision, some form of filtering has to be employed to avoid including functionally unrelated neighbors in the prediction process. At that time, there have already been some studies that observe functional similarity between proteins with overlapping interaction neighborhood [19, 20]. These independent observations motivated us to study the possibility of using the observation of common interaction partners as a way to identify functionally related protein pairs from the large number of indirectly interacting proteins. We initially adopted the Czekanowski-Dice distance (CD distance) used in [20] for this purpose. The CD-Distance is defined as:

$$D\left(u,v\right)=\frac{\left|N_u \Delta N_v\right|}{\left|N_u\cup N_v\right|+\left|N_u\cap N_v\right|} \tag{1}$$

where N_p refers to the set that contains p and proteins that interact with it, and X Δ Y refers to the symmetric difference between two sets X and Y. $D(u,v) < 1$ if proteins u and v interacts with each other, or with at least one common protein. If $N_u = N_v, D(u,v)$ will be 0. On the other extreme, if $N_u \cap N_v = \emptyset, D(u,v)$ will be 1. This distance function can be trivially converted into its corresponding similarity function:

$$S_{CD}\left(u,v\right)=\frac{\left|N_u\cap N_v\right|}{\left|N_u\cup N_v\right|+\left|N_u\cap N_v\right|} \tag{2}$$

The similarity function captures the overlap between two sets reasonably when the sets N_u and N_v are not very different in size. However, when one set is greater than the other, $S_{CD}(u,v)$ will be small even when $N_u \cap N_v$ is a large or complete subset of the smaller set. Since the sets represent interaction neighbors in this case, this means that the similarity score between a protein with low degree and one that is well connected will always be low. As protein interactions are subjected to systematic biases due to experimental design and incomplete coverage, this similarity function is likely to underestimate functional relationships in such cases. Hence we proposed a variant of the similarity function, which we refer to as the Functional Similarity weight (FS-weight) to place greater weight on the overlap between the two sets:

$$S_{FS}\left(u,v\right)=\frac{2\left|N_u\cap N_v\right|}{\left|N_u-N_v\right|+2\left|N_u\cap N_v\right|+\lambda_{u,v}}\times\frac{2\left|N_u\cap N_v\right|}{\left|N_v-N_u\right|+2\left|N_u\cap N_v\right|+\lambda_{v,u}} \tag{3}$$

$\lambda_{u,v} = \max\left(0, n_{avg} - (|N_u - N_v| + |N_u \cap N_v|)\right)$ where n_{avg} is the average number of interactions that a protein participates in.

Functional Association and Experimental Assays

As described earlier, protein-protein interactions can be observed in a variety of experimental assays. While the different assays are capable of identifying interactions between proteins (and genes), they often rely very diverse mechanisms. Consequently, each assay comes with its limitations. In yeast two-hybrid systems, false positives (interactions observed that are non-existent) can arise due to a wide number of factors such as background transcriptional activity of baits, mutation of the host yeast strain, bait proteins that binds directly to the DNA upstream of the reporter genes, and "sticky" bait or prey proteins that easily binds a large number of proteins in a non-specific manner [21]. In tandem affinity purification experiments, false negatives (interactions that exist but not observed) may arise due to the TAP tag interfering with interaction, and not all proteins within the complex may bind tightly enough to be detected [22]. While there is no simple way to take into account such differences in the nature and limitations of different experimental assays, we can moderate the impact of such differences to the function prediction process by estimating the confidence we have in each type of experiment with regard to their ability to associate proteins with similar functions. For each type of experiment, this can be a simple estimate of the prior probability that protein interactions observed by such experiments involve protein pairs that share some function:

$$r_i = \frac{\sum\limits_{(u,v)\in E_i} \delta(u,v)}{|E_i|} \tag{4}$$

E_i is the set of interactions observed in experiment i; $\delta(u,v)$ is 1 when protein u and v share some function, 0 otherwise.

For interactions that are observed in multiple experiments, we would expect the confidence to be much higher since it is reproducible and less likely to be a false positive due to random experimental errors. Taking into account the confidence of individual experimental types, as well as reproducibility over multiple experiments of the same or different nature, we can naively combine the prior probabilities for each experimental type to compute the probability that an observed interaction is associated with sharing of function:

$$r_{u,v} = 1 - \prod_{i\in E_{u,v}} (1 - r_i)^{n_{i,u,v}} \tag{5}$$

r_i is the estimated reliability of experimental type i; $E_{u,v}$ is the set of experiments in which interaction between u and v is observed;
$n_{i,u,v}$ is the number of times interaction (u,v) is observed from experimental type i.

With a quantifiable estimate of the confidence of different experimental sources of interaction data, we can incorporate this information into the FS-weight formulation:

$$S_{FS}(u,v) = \frac{2\sum\limits_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum\limits_{w\in(N_u-N_v)} r_{u,w} + \sum\limits_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right) + 2\sum\limits_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w} + \lambda_{u,v}} \times \frac{2\sum\limits_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum\limits_{w\in(N_v-N_u)} r_{v,w} + \sum\limits_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right) + 2\sum\limits_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w} + \lambda_{v,u}} \tag{6}$$

We find the FS-weight measure to correlate positively with function similarity between interacting proteins (pearson's correlation coefficient = 0.53). The measure also exhibit a positive, abeit weaker correlation with function similarity between level-2 interaction neigbors (correlation coefficient = 0.38).

Function Prediction Using Indirect Association

With an appropriate function to estimate the strength of functional relationships between directly and indirectly interacting proteins, it is now more plausible to include the level-2 neighborhood for functional prediction. We proposed the FS-weighted averaging function that uses the weighted frequency of a function x in both the direct (N_u) and indirect (N_w) neighbors of a protein u to compute a normalized score to estimate the likelihood of protein u to participate in function x:

$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\text{int}}\pi_x + \sum_{v \in N_u}\left(S_{FS}(u,v)\delta(v,x) + \sum_{w \in N_v} S_{FS}(u,w)\delta(w,x)\right)\right]$$

Z is the sum of all weights:

$$Z = 1 + \sum_{v \in N_u}\left(S_{FS}\,(u,v) + \sum_{w \in N_v} \max(S_{FS}(u,v)S_{FS}(v,w), S_{FS}(u,w))\right) \qquad (7)$$

Evaluation on Function Prediction

The FunCat annotation scheme is a tree-like structure with each child term being a more specific form of its parent. Some fuctional aspects of proteins tend to be better studied than others, and hence some annotation branches tend to be deeper and annotated to a larger number of proteins. To minimize biases when evaluating prediction performance, we want to avoid evaluating redundant annotations (e.g. a functional term and its parent function, as well as more distant ancestor terms). A simple way to achieve this would be to decide on an arbitary level of annotation (e.g. all nodes with a depth of 5), but due to large variations in the depth of different branches, we may end up evaluating very general functions of some branches and very specific functions of others. To overcome this problem, we adopt the informative functional classes approach proposed in [23]. A functional term is designated as *informative* if it is annotated to n or more proteins (we use $n = 30$), and does not have a child term that is annotated to n or more proteins. In this way, an informative term will be the only informative term among all its ancestors or descendants. By using only informative terms, we can ensure that there is no redundancy between the functions that are used for evaluation. Moreover, since these informative terms are annotated to a sufficiently large number of proteins, we will avoid evaluating functional terms that are too rare to be inferred practically. Using a 10-fold cross validation procedure, we benchmarked our proposed method against several published approaches at the time of the study on the prediction of informative FunCat terms using protein-protein interactions from BioGRID and showed that it performed significantly better (Fig. 2). We also benchmarked our method against other approaches using a dataset compiled in an earlier study comprising YPD functional categories

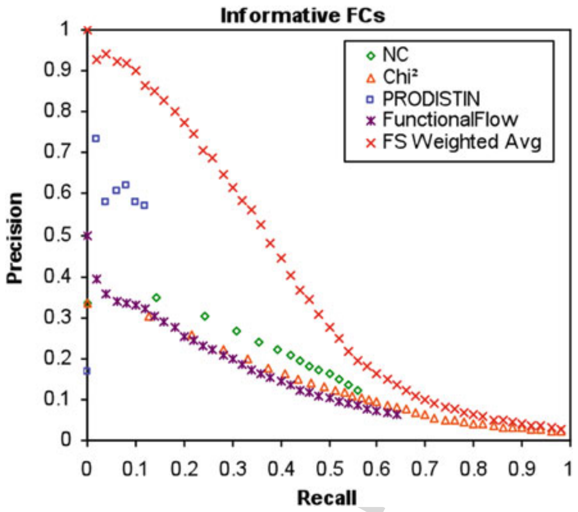


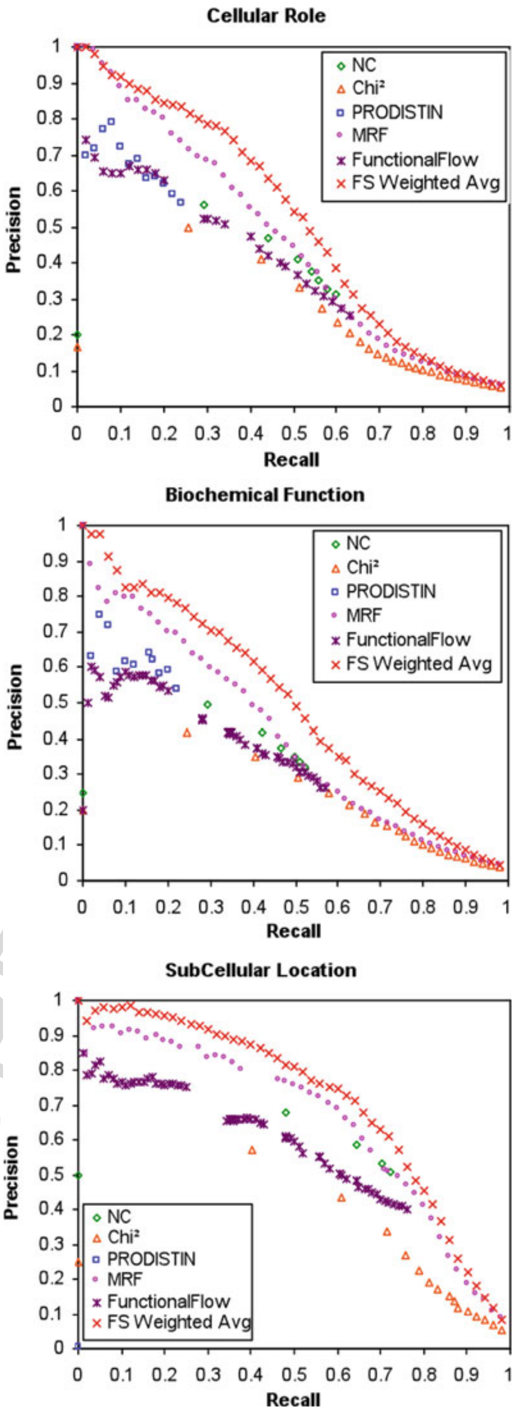
Fig. 2 Precision–recall curves for prediction of FunCat functions for proteins from *S. cerevesiae* from BioGRID interactions using various approaches. Figure from [17]

and protein–protein interactions from MIPS [14], and showed that the predictions made using our method achieved a better precision at nearly all levels of recall for the three YPD categories (Fig. 3).

Prediction of Gene Ontology Functional Annotations on Multiple Species

While we had some success showing that indirect association of FunCat functional annotations are abundant between non-interacting proteins, the annotation scheme that was, and still is most widely adopted is the Gene Ontology (GO). Similar to FunCat, this comprehensive functional annotation scheme organizes functional annotations into a hierarchical structure that explicitly describes parent-child relationships between annotations, where the children of an annotation are more specific annotations that fall under it. The hierarchy structure adopted by GO, however, is a Directed Acyclic Graph (DAG), instead of the tree structure used by Funcat. The main implication of this is that a GO term can have more than one parent term. The GO annotation scheme constitute a DAG structure for each of the 3 namespaces *molecular_function*, *biological_process*, and *cellular_component*, that provide different aspects of biological characterization of a gene and its protein product. To study if the use of indirect functional association is general enough to be beneficial for functional prediction based on the GO scheme, and for species other than *S. cerevisiae*, we performed a follow-up computational study in 2007 on 7 species [24]. The objective of the study was to answer 3 key questions about using protein-protein interactions and indirect functional association for protein function prediction: (1) Does the use of protein-protein interactions provide any additional coverage over the

Fig. 3 Precision–recall curves for prediction of YPD functions for proteins from *S. cerevesiae* from MIPS protein–protein interactions using various approaches. Figure from [17]



conventionally accepted use of sequence homology for protein function prediction; (2) Does the use of indirect functional association provide~~s~~ any additional enhancement in coverage over direct guilt by association; and (3) Are the conclusions made for indirect functional association on FunCat terms applicable to function prediction using GO terms over different species with differences in quantity and even quality of data?

Data Availability

At the time of study, protein-protein interaction data was available for 7 species in the BioGRID database: *S. cerevisiae*, *D. melanogaster*, *A. thaliana*, *H. Sapiens*, *M. Musculus*, *R. norvegicus* and *C. elegans*. Gene On~~o~~tology annotations were also available for these species. The amount of interaction data available to perform the study is summarized in Table 1. As we can only evaluate prediction performance on annotated proteins, we present the number of interactions that involve annotated proteins as a proxy for data availability.

Table 1 Annotation and interaction data statistics for different species at time of study. Table from [24]

Genome	Interactions involving annotated proteins	Annotated proteins	Avg. no. of annotated neighbours per protein
S. cerevisiae	50, 434	4005	21.6654
D. melanogaster	24, 991	2763	4.2823
A. thaliana	909	382	1.8386
H. Sapiens	5784	5784	1.6761
M. Musculus	1892	1892	1.3595
R. norvegicus	590	590	0.9803
C. elegans	4349	382	0.7382

Protein–Protein Interactions vs. Sequence Homology

To answer our first question on the usefulness of protein–protein interaction data as an additional source of data to complement conventional sequence homology for protein function inference, we examine the number of known functional annotations can already be inferred using the top hits of a BLAST search against all sequences from the Gene Ontology Database. The analysis is only done for *S.cerevisiae* and *D. melanogaster* as the amount of protein–protein interaction data is too little for meaningful analysis on the other species. The fraction of known annotations that can be annotated in this way for each species is computed using E-value cut-offs between 1 and 1e–10, and summarized as white bars in Fig. 4. As one would expect, coverage decrease with more stringent E-value cut-offs, possibly in exchange for better precision (not shown). For each E-value cut-offs, we next compile the number of additional functional annotations that can be transfer in a guilt-by-association fashion based on protein–protein interactions as a fraction of the total number of known annotations (light blue bars in Fig. 4). We find that protein–protein

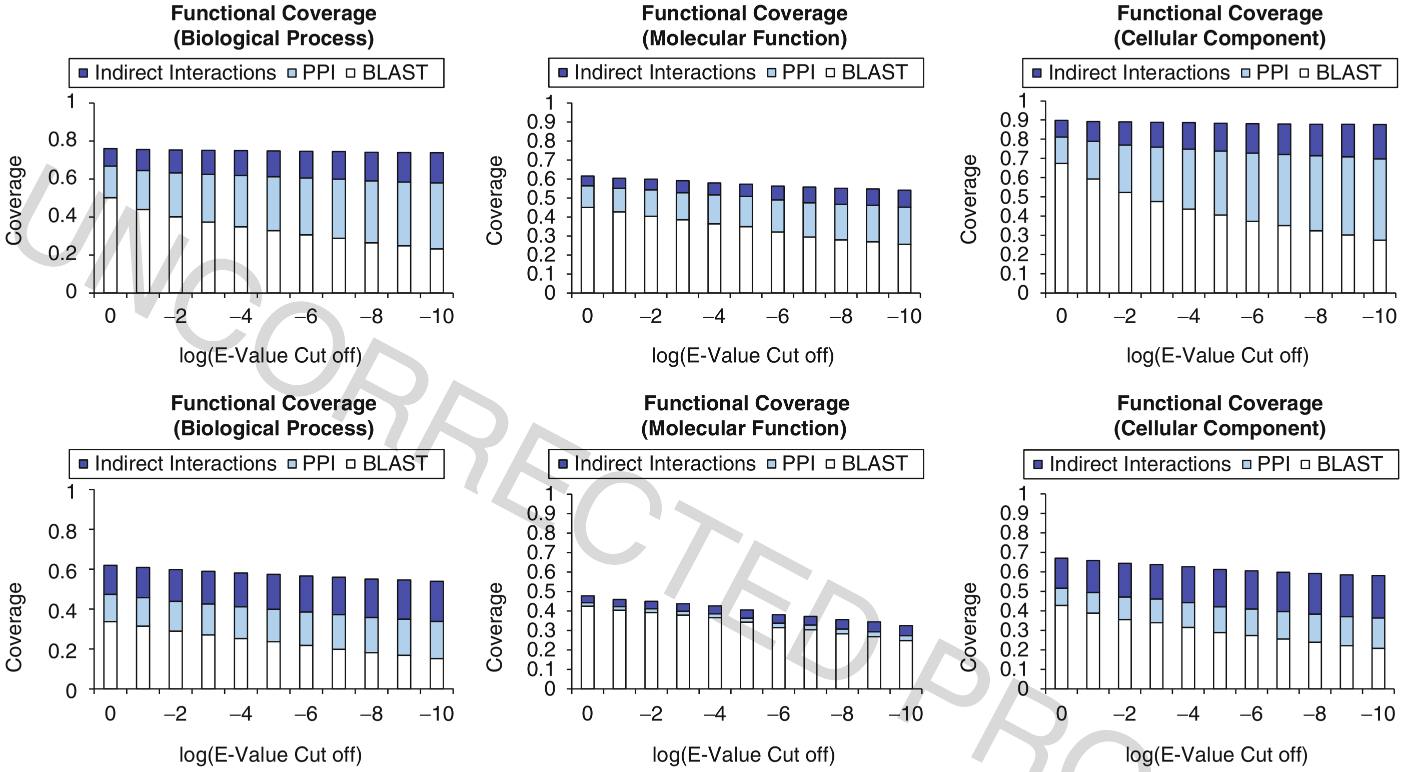


Fig. 4 Fraction of known functional annotations that can be suggested through BLAST homology search; and the additional annotations that can be suggested through: (1) direct-protein interactions (PPI); and (2) indirect-protein interactions. A range of BLAST E-value cut-off between 1 and 1e–10 is used. BLAST is performed on sequences from the gene ontology database. Proteins with very close homologs (E-value <= 1e–25) are excluded from analysis. The *top row* shows the results from *S. Cerevisiae* and the *bottom row* shows the results from *D. melanogaster.* The three columns depict results on the biological process (*left*), molecular function (*centre*) and cellular component (*right*) categories of the Gene Ontology. Figure from [24]

interactions provided some additional coverage (around 20% for *S.cerevisiae* and 10% for *D. melanogastor*) even at relaxed BLAST E-value cutoffs of >=0.01 for inferring *biological_process* and *cellular_component* annotations. Finally, we also compute any further coverage that may be gleaned if we also allow functional inference using indirect functional associations between level-2 interaction neighbors. We found that there is substantial additional coverage that may be gained in this way (dark blue bars in Fig. 4) for both species. This analysis addressed the first two questions we seek to answer, that is: (1) There are a fair number of GO annotations that cannot be inferred through simple sequence homology, but can potentially be predicted from protein-protein interactions; and (2) Extending functional predictions to level-2 neighbors helps to further increase coverage by including functional annotations that cannot be associated to a protein via sequence homology or direct protein–protein interactions.

Function Prediction Performance

Finally, we investigate if the function prediction method that we proposed earlier can be used to make better predictions for GO terms for the seven species by using functional association with indirect interaction neighbours. Again, we used the informative functional classes concept to identify informative GO terms to be used for evaluation for each species. Comparing FS-weighted averaging with the Neighbor-Counting and Chi-Square approaches, we found that FS-weighted averaging achieved superior precision–recall performance in all seven species (Fig. 5).

Indirect Functional Association and Complex Discovery

Protein Complex Discovery

Proteins often perform function by aggregating into complexes to perform sophisticated biological tasks. Many well-conserved protein complexes perform key biological functions such as transcription, splicing, mRNA export and protein synthesis. Through complex formation, the primary molecular functions of individual proteins (such as the ability to bind DNA or RNA, shuttle between membranes, transport certain materials and interact with particular proteins) are recruited in a coordinated fashion to perform highly specialized functions. RNA polymerases, ribosomes and spliceosomes are some examples of widely studied protein complexes with well-understood functionalities. Therefore to better understand the higher-level biological processes in which proteins participate, it is necessary to look beyond individual protein features such as sequences and structures and observe how proteins form larger functional units. While experimental assays such as tandem affinity purification and co-immunoprecipitation can be used to identify protein complexes, these are usually suitable for capturing stable complexes. Many weak or transient complexes are likely to be missed.

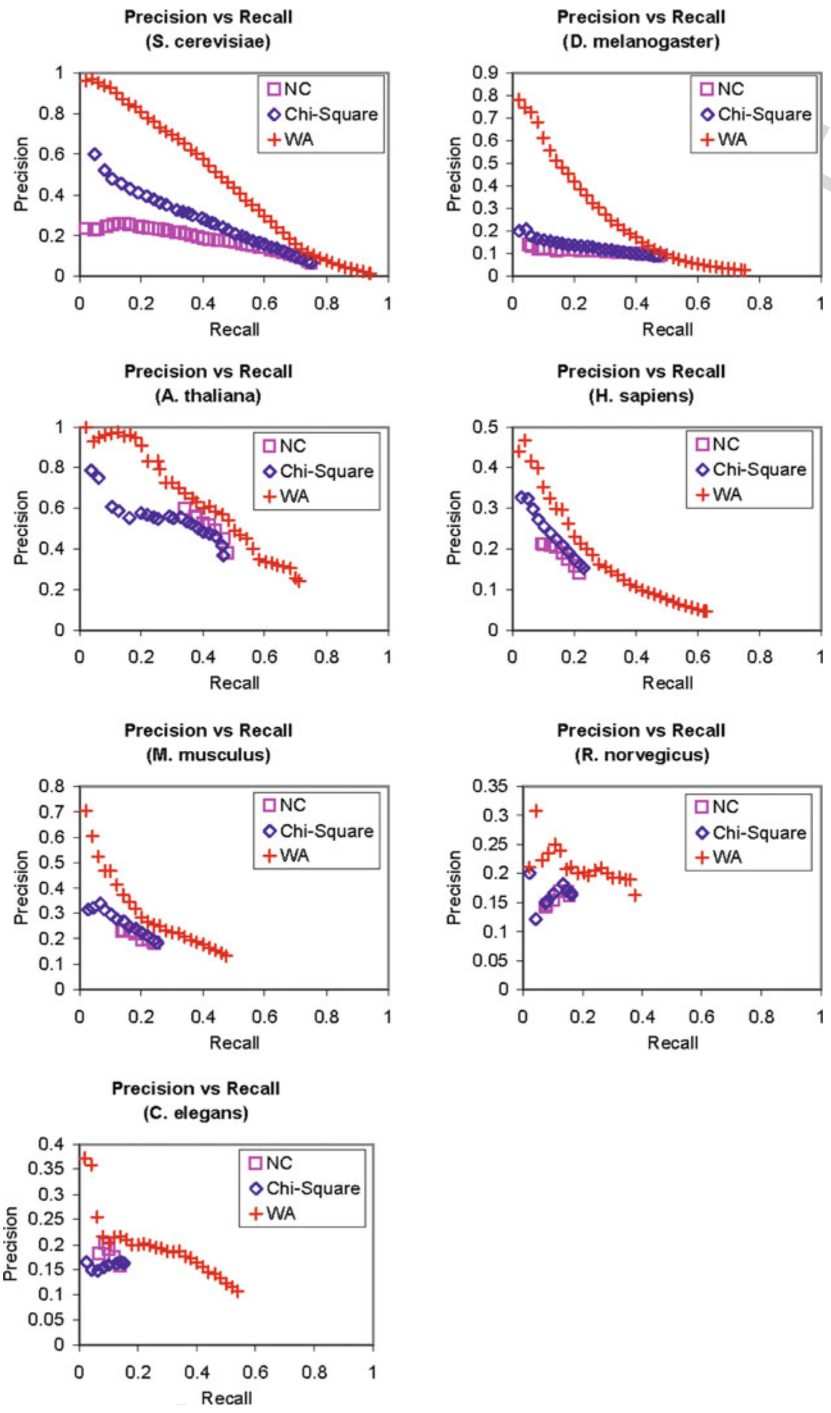


Fig. 5 Precision vs. recall graphs of the predictions of informative GO terms from the gene ontology biological process category using (1) *Neighbour Counting*(*NC*); (2) *Chi-Square*; and (3) *FS-Weighted Averaging*(*WA*), for seven genomes. Figure from [24]

The importance of identifying protein complexes motivated many bioinformatics approaches to identify protein complexes computationally from protein–protein interactions. Several insightful studies contributed significantly in motivating research in this area. Spirin and Mirny [25] investigated highly connected proteins in a physical protein–protein interaction network, and found functionally related proteins to be highly connected with each other, but sparsely connected with the rest of the network. Some of these densely connected proteins coincide with known stable protein complexes, while many others are found to be related to dynamic functional units involved in activities such as signaling cascades and cell cycle regulation. Bu and colleagues studied topological structures (quasi-cliques and quasi-bicliques) in protein–protein interactions and found that many of these structures involved functionally related proteins [26]. Bader and Hogue [27] proposed a computational method of protein complex discovery from protein–protein interaction networks by "growing" complexes from "seed proteins" with dense local network. The algorithm, MCODE, was subsequently implemented as a plug-in for the popular bioinformatics visualization software Cytoscape [28]. The recurring theme among these studies is that function modularity in biological systems may be encoded in protein–protein interactions, and identifying such functional modules allows us to better understand how proteins function together.

Protein Complexes with Limited Interactions

From our earlier studies, we found that many indirectly interacting proteins share functional annotations from different schemes including YPD, FunCat and GO. These indirectly interacting proteins that perform similar biological functions could in reality be forming protein complexes, with their common interacting proteins acting as adaptors that bring them into close proximity. This is especially likely for larger complexes since proteins have limited binding pockets and usually have reasonably high binding specificity. Since these proteins do not interact, there may not be sufficient overlap between their local interaction neighborhoods for conventional clustering approaches based on network density to associate them. As the FS-weight measure has been demonstrated to provide some estimation to functional similarities between two indirectly interacting proteins, we are interested to see whether including indirect interactions with high FS-weight scores into the protein interaction network can help improve discovery of complexes that involve~~s~~ less physical inter-connections. On the other hand, since the FS-weight can also provide some estimation of functional similarity between proteins that interact, we may be able to remove possibly spurious interactions that are likely to be functionally unrelated from the interaction network. We explore these ideas in a subsequent work [29, 30] that study how complex prediction performance is affected by (1) applying existing clustering methods on modified physical protein–protein interactions; and (2) proposing a clustering algorithm that implicitly take FS-weight into account.

Approaches for Protein Complex Prediction

At the time of the study there are two general approaches to protein complex prediction from protein-protein interactions. The first approach, which we refer to as *clique finding*, imposes a stringent requirement on what constitutes a protein complex. A *clique* is a fully connected subgraph in which each node is connected to all other nodes in the subgraph. Spirin and Mirny [25] explored two methods of finding densely connected subgraphs in a protein interaction network, one of which is to ~~r~~enumerate all cliques in the network. The strict constraint imposed by clique finding keeps false positives low and makes the approach robust to noise in the interaction network. However, sensitivity is likely to be severely limited. Bu and colleagues used a more relaxed constraint for complex prediction by looking for *quasi-cliques*, which are dense subgraphs that are almost complete [26]. The other general approach to complex prediction, which we refer to as clustering, involves the use of heuristic algorithms to find groups of densely connected proteins, usually based on network properties such as network density. Brohee and colleagues [31] evaluated some of these clustering methods, namely the Restricted Neighborhood Cost-Based Clustering (RNSC) [32], MCODE, Markov Clustering (MCL) [33], and Super Paramagnetic Clustering (SPC) [34] for protein complex prediction from protein–protein interaction networks. Using 6 protein–protein interaction networks from [2, 5, 35–38] and cataloged complexes from MIPS [39], the authors optimized the parameters for each clustering algorithm and benchmarked them over several performance metrics.

Modifying the Interaction Network with FS-Weight

Given a input interaction network, FS-weight is applied to assign a score to all interactions as well as level-2 indirect interactions. By applying a threshold $\mathit{FS\text{-}Weight}_{\mathrm{min}}$, we include indirect interactions that surpass this threshold into the original interaction network. On the other hand, direct interactions in the original interaction network that does not meet this threshold are removed from the interaction network. Since the FS-Weight measure exhibit positive correlation with functional similarity, we expect connected proteins in the modified network to be more functionally related than that of the original network. In the study we performed experiments using the 6 protein–protein interaction networks studied in [31], which comprises 2 datasets derived from large-scale yeast two-hybrid studies, and 4 datasets from affinity purification and mass spectrometry. We refer to this combined network as the “combined” dataset. We also used a larger dataset comprising all physical protein-protein interactions from BioGRID which is a superset of the 6 networks.

As a preliminary study of the feasibility of this approach, we compute the fraction of all interactions that involve a pair of proteins that belong to some common complex for the two interaction networks, as well as the modified versions of these networks. We find that if we introduce level-2 indirect interactions indiscriminately, the fraction of interactions that involve co-complex proteins decreases drastically (Fig. 6, L1 & L2). However, if we only include level-2 interactions with high FS-weight scores, we are able to maintain these fractions at similar levels (L1 & Filtered L2) as that for the original interaction networks (L1). Finally, when we also remove direct interactions with low FS-weight after including level-2 interactions with high

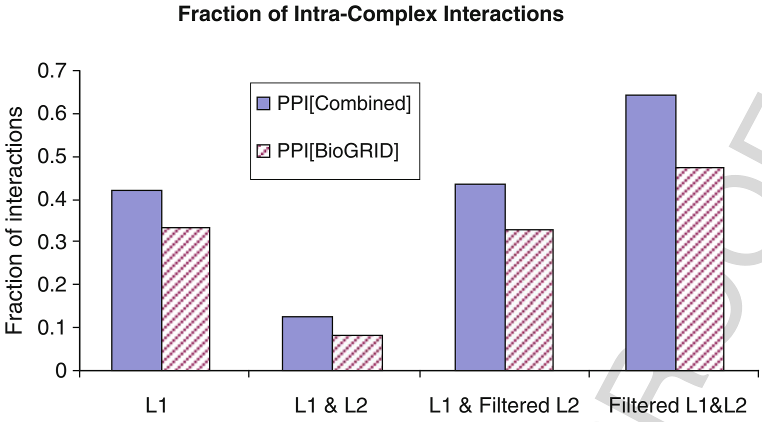


Fig. 6 Fraction of intra-complex interactions with nodes sharing some complex membership for different PPI networks. Figure from [30]

FS-weight, the fractions of the interactions that involve proteins from common complex increased significantly (Filtered L1 & L2). These observations are encouraging and suggest that we could possibly make the network more amenable to complex discovery in this manner.

A New Complex Prediction Approach

Since the FS-weight can provide a decent estimate of the functional relatedness of an interaction, we may be able to exploit this information in the complex prediction process. Taking this idea into consideration, we proposed a novel complex prediction approach and benchmark it alongside with the 4 existing clustering algorithms evaluated in [31]. Our approach, PCP (Protein Complex Prediction), is a heuristic algorithm that involves a three-step iterative process:

Maximal Clique Finding

The first step involves finding all maximal cliques of at least size 2 from the network. This can be done efficiently on a sparse graph using the algorithm described in [40]. For nodes that belong to multiple cliques, we assign them to only one clique using a heuristic method to maximize the average FS-Weight scores of the edges in each non-overlapping clique. Since this is also the performance bottleneck of the algorithm, we also proposed an alternative heuristic method for finding non-overlapping cliques as a replacement for this step which did not have any significant impact on prediction performance.

Computing InterClusterDensity

The clique finding step will return very dense subgraphs that are completely connected. A clique is unlikely to represent a complete real complex, but rather a densely-connected subset of it. To associate less densely connected parts of the complex, we can merge cliques that are well-connected. To provide a quantitative measure of interconnectedness between a pair of subgraphs (S_a, S_b), we define the *InterClusterDensity* (ICD) as follows:

$$ICD(S_a, S_b) = \frac{\sum S_{FS}(i,j)|i \in V_a, j \in V_b, (i,j) \in E}{|V_a| \cdot |V_b|} \tag{8}$$

where V_x is the set of vertices of subgraph S_x. This is simply the weighted sum of all edges between members of the two subgraphs, divided by the maximum number of possible edges between them.

Subgraphs Merging

Using the ICD measure, we can now imagine each clique as a node in a new graph, and insert an edge between two nodes that has a ICD score greater than an arbitary threshold $\mathrm{ICD}_{\mathrm{min}}$. We can now perform the maximal clique finding step again on the new graph. The nodes in the cliques found will no longer be proteins, but rather groups of proteins. By reiterating this process, smaller groups of proteins will gradually be merged into large groups in a hierarchical manner. To allow the better connected nodes to be merged first, we start from a high $\mathrm{ICD}_{\mathrm{min}}$ threshold, and gradually reduce the threshold whenever no further merging can be made.

Performance Evaluation

Known protein complexes from MIPS is used as the gold standard against which performance is evaluated. In order to see if novel predictions are indeed made, we also used MIPS complexes released 2 years apart, in 2004 and 2006. Unlike function prediction, the practical usefulness of complex prediction lies in the ability to predict a set instead of a pair. Therefore to make quantitative evaluation meaningful, we must first define what constitute a correct prediction, that is, the critia for a predicted cluster to be considered as matching a known complex. We adopt the overlap measure from [27]:

$$Overlap(S, C) = \frac{|V_s \cap V_c|}{|V_s| \cdot |V_c|} \tag{9}$$

In [27], and overlap score of 0.2 or more is considered a match. We used a slightly higher threshold of 0.25 in our study. Since there may be more than one cluster matching a complex and vice versa, we used a slightly modified version of the conventional precision and recall measure. We defined precision here as the number of predicted clusters that matched a complex:

$$\text{Precision} = \frac{\text{matched}_{\text{clusters}}}{\text{predicted}_{\text{clusters}}} \tag{10}$$

Similarly, we defined recall as the number of known complex that matched a cluster:

$$\text{Recall} = \frac{\text{matched}_{\text{complexes}}}{\text{known}_{\text{complexes}}} \tag{11}$$

Complex Prediction Performance

We performed protein complex prediction using RNSC, MCL, MCODE and PCP on the original interaction networks as well as the modified networks. For the RNSC, MCL and MCODE algorithms we used the optimal parameters that are derived by the authors in [31]. We determined optimal parameters for PCP empircally. Compared to predictions made on the orignal network (Fig. 7 top row), we found that the precision–recall performance for MCL, MCODE and PCP improved significantly after the networks are augmented and filtered using FS-weight (Fig. 7 middle row) for both the combined and BioGRID datasets. The performance of RNSC, however, did not changed significantly. PCP performed the best among the clustering algorithms studied for both interaction datasets. We also evaluated the predictions made for the modified network against the newer 2006 MIPS complex dataset (Fig. 7 bottom row), and found that precision–recall performance has generally improved for all the algorithms, which suggested that some of the predictions made which are "novel" based on the 2004 complex dataset were indeed identified to be real complexes a couple of years later.

Improving the Reliability of Interactions

Efforts in computational protein function prediction and protein complex discovery are plagued by the common challenges of false positives, and perhaps more seriously, false negatives in protein–protein interactions. Much work has been done to assess the error rates of interaction data [41–44], and estimates based on overlaps in datasets indicated yeast two-hybrid datasets to contain false positives as high as 50%. More recent work [45] suggested that such estimation are likely to be flawed, and a more recent estimate [46] placed the false discovery rate of yeast two-hybrid interactions at around 10% and false negative rate at around 50% for *S.cerevisiae*. Nonetheless, false positives and false negatives is an important concern, and much effort has been made to improve the quality of interaction data by computationally assessing the confidence of individual interactions. Some of these methods involve using independent, biologically relevant data such as gene expression and sequence homology [43, 47], while others solely used topological properties inherent in the network [48–51].

For methods that derive confidence for each interaction using a topological measure, the weighted interactions can be seen as a being more representative of the underlying "real" network. Hence intuitively it would make sense to use this weighted network to re-compute the confidence for each interaction. We showed in

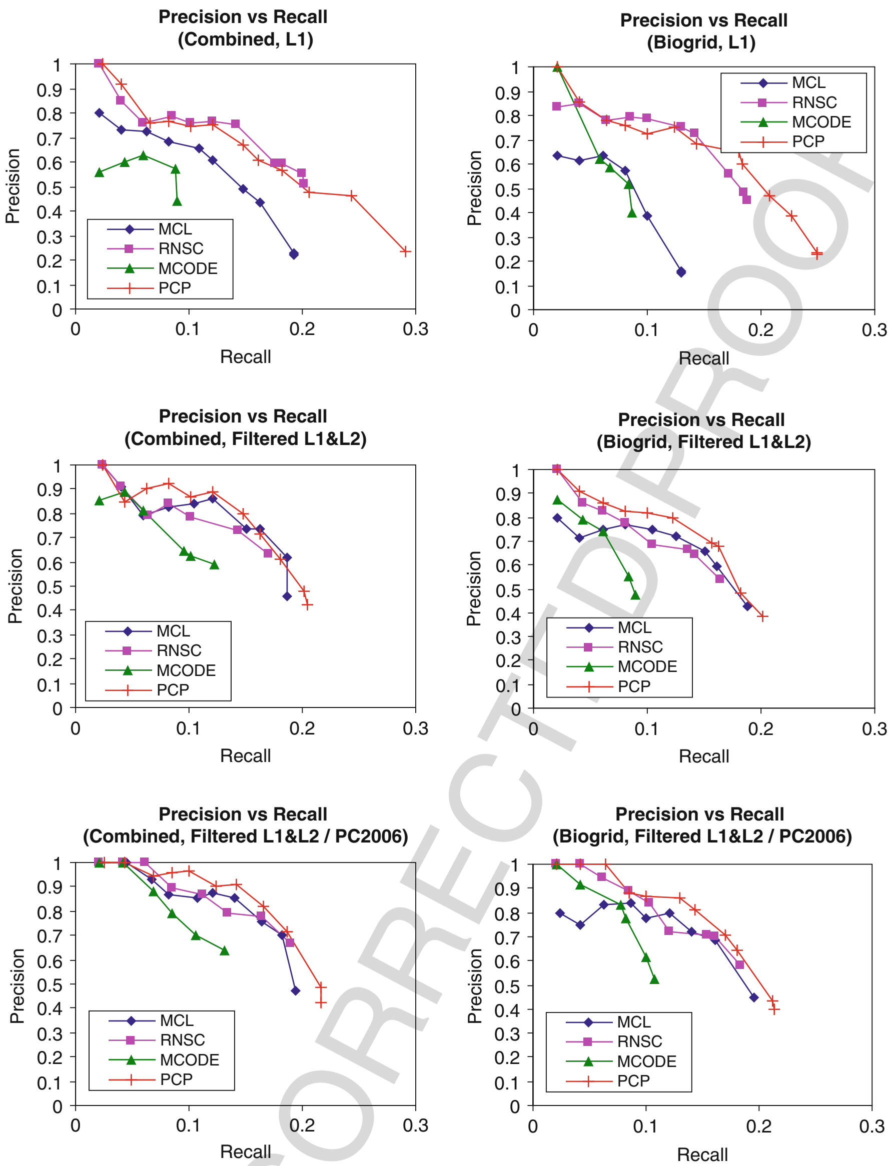


Fig. 7 Precision–recall curves for complex predictions using MCL, RNSC, MCODE and PCP for the combined (*left column*) and BioGRID (*right column*) datasets. Predictions are made using the original networks (*top row*) and the modified networks (*middle row*) and evaluated against complexes from the 2004 MIPS dataset. Predictions made using the modified networks are also evaluated against complexes from the 2006 MIPS dataset (*bottom row*). Figure from [30]

two recent studies that this concept can be used to improve upon local topological measures such as the CD-Distance or FS-Weight in identifying functionally-related interactions and improve complex prediction performance [52, 53].

Iterative Scoring

We define the iterative scoring function from a base topological score function. In the study we used a variant of the CD-Distance as the base measure:

$$AdjustCD\,(u,v) = \frac{2\,|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v} \tag{12}$$

λ_u and λ_v are pseudo counts used to penalize proteins with few neighbors, and are defined similarily as $\lambda_{u,v}$ used in FS-weight. The iterative version of AdjustCD is defined as:

$$w^k(u,v) = \frac{\sum_{x\in N_u\cap N_v}\left(w^{k-1}(x,u) + w^{k-1}(x,v)\right)}{\sum_{x\in N_u} w^{k-1}(x,u) + \lambda_u^k + \sum_{x\in N_v} w^{k-1}(x,v) + \lambda_v^k} \tag{13}$$

where $w^{k-1}(u,v)$ is the weight of the edge (*u*,*v*) at the (*k*–*1*)-th iteration. At the initial stage ($k = 0$), $w^0(u,v) = 1$ if the edge (*u,v*) exists and $w^0(u,v) = 0$ otherwise.

$$\begin{aligned} \lambda_u^k &= \max\left\{0, \frac{\sum_{x\in V}\sum_{y\in N_x} w^{k-1}\,(x,y)}{|V|} - \sum_{x\in N_u} w^{k-1}\,(x,u)\right\} \\ \lambda_v^k &= \max\left\{0, \frac{\sum_{x\in V}\sum_{y\in N_x} w^{k-1}(x,y)}{|V|} - \sum_{x\in N_v} w^{k-1}(x,v)\right\} \end{aligned} \tag{14}$$

are the weighted variants of λ_u and λ_v at the *k*-th iteration and *V* is the set of all nodes in the network. At iteration $k = 1, w^k(u,v) = \text{AdjustCD}(u,v)$. We refer to the *k*-iteration version of this scoring function as AdjustCD$^{\text{k}}$.

We showed in [52], that the use of this iterative scoring function reaches best performance at $k = 2$. The weights assigned to interactions using the score function were significantly more predictive of functional similarity and co-localization than FS-Weight and CD-Distance. The weights assigned to indirect level-2 interactions with the iterative function are also more relevant to functional homogenity and localization coherence. These observations suggested that the iterative weighting function may be used to improve the protein complex prediction approach we visited in the previous section.

Complex Discovery Using AdjustCD^k Weighted Interactions

In [53] we conducted a detailed analysis on protein complex finding using interactions that are weighted using AdjustCD$^{\mathrm{k}}$. Two reference sets of protein complexes are used. The first set is the set of hand-curated complexes from MIPS [39]. The other set of complexes are modeled from three-dimensional structures that were screened using electron microscopy by Aloy et al. [54]. Using the 6 physical protein-protein interaction datasets used in [30, 31], we study how the performance of MCL, MCODE, CFinder [55] and a new clustering algorithm, which we called CMC (Clustering Based on Maximal Cliques), is affected when the input interaction is weighted using AdjustCD$^{\mathrm{k}}$.

The CMC Algorithm

Like the PCP algorithm, the CMC algorithm starts by finding all maximal cliques in the network using the algorithm described in [40]. However, unlike PCP, CMC do not iteratively merge cliques through building higher-level abstract networks. Instead, a heuristic procedure is used to quickly merge well overlapping cliques into larger clusters. Each clique C is first scored based on its weighted network density:

$$score(C) = \frac{\sum_{u \in C, v \in C} w(u,v)}{|C| \cdot (|C| - 1)} \tag{15}$$

where $w(u,v)$ is the weight of edge (u,v) scored using AdjustCD$^{\mathrm{k}}$. The cliques are then sorted into a list based on their score in a decreasing order. Each clique C_i is in turn examined beginning from the top of the sorted list. For every other clique C_j in the list which overlaps with C_i above a predefined threshold (i.e. $\left|C_i \cap C_j\right| / \left|C_j\right| \geq overlap\_thres$) and $score(C_j) < score(C_i)$, C_j is removed from the list. A weighted inter-connectivity score is then computed between C_i and C_j to decide if C_j should be merged with C_i:

$$inter\!-\!score\,(C_1, C_2) = \sqrt{\frac{\sum_{u \in (C_1 - C_2)} \sum_{v \in C_2} w(u,v)}{|C_1 - C_2| \cdot |C_2|} \cdot \frac{\sum_{u \in (C_2 - C_1)} \sum_{v \in C_1} w(u,v)}{|C_2 - C_1| \cdot |C_1|}} \tag{16}$$

If $inter\text{–}score(C_i, C_j) \geq merge\_thres$, then C_j will be merged with C_i, otherwise it is discarded. *merge_thres* is a pre-defined parameter. The parameters *overlap_thres* and *merge_thres* are empirically determined.

Performance Evaluation

In this study we considered a predicted cluster to match a protein complex if the Jaccard index between them is at least 0.5. To ensure that random matches are unlikely, we randomly swapped complex members to see if the resulting random complexes match with any predicted clusters from the CMC algorithm. We found no matches over 1000 such runs. Precision and recall are defined similarly as described in the previous section of this chapter. We found that all 4 clustering methods achieved significant improvement in precision when using weighted networks compared to unweighted networks. Using k=2 in the AdjustCD^k weighting function result in the best performance among most of the clustering algorithms that are evaluated, and further increase in k to 20 showed little change in performance for CMC and Cfinder.

Robustness Against Noise in the Interaction Network

Perhaps the most interesting observation we made from this study is the robustness of the weighted network to random additive noise. By randomly adding edges to the original network, we examine the impact of additive noise on the prediction performance of CMC using k=1, k=2 and k= 20 for AdjustCD^k weighted versions of the interaction network. Evaluating against the complex dataset from [54], we find that when k=1, the performance of the CMC algorithm degrades significantly when random interactions amounting to 50% of the original network ~~is~~ added, and continues to degrade quickly with higher levels of noise (Fig. 8, top). When k=2, however, the performance of CMC show~~ed~~ only a slight decrease when 50% random interactions are added, and only exhibited significant degradation when added random interactions ~~is greater than~~ 300% of the original network. At k=20, the performance of CMC only show~~ed~~ signs of degradation when the number of added random interactions is 5 times that of the original network. These observations suggest~~s~~ that the iterative scoring approach can potentially be used to benefit downstream analyses that make~~s~~ use of protein-protein interaction data by accentuating the biologically relevant subset of interactions within noisy datasets.

Conclusions

In this chapter, we briefly review some of the works we have done on using protein-protein interactions for computational approaches related to protein function discovery. The key concepts introduced here include~~s~~ indirect functional association between proteins that do not interact directly, the use of topological weights such as FS-weight to identify functionally relevant interactions so that such indirect interactions can be feasible for practical use, and the impact of using topological weighting techniques (such as FS-weight and the iterative AdjustCD^k) to improve interaction data quality on protein complex prediction. It is noteworthy that while protein-protein interaction data is highly relevant to understanding and inferring protein functions, it captures a limited aspect of protein functionality. Greater success in computational function prediction is likely to be achievable through the use of a multitude of biological data such as expression profiles, sequence homology and more. Such holistic approaches are actively being researched on [56–59], and

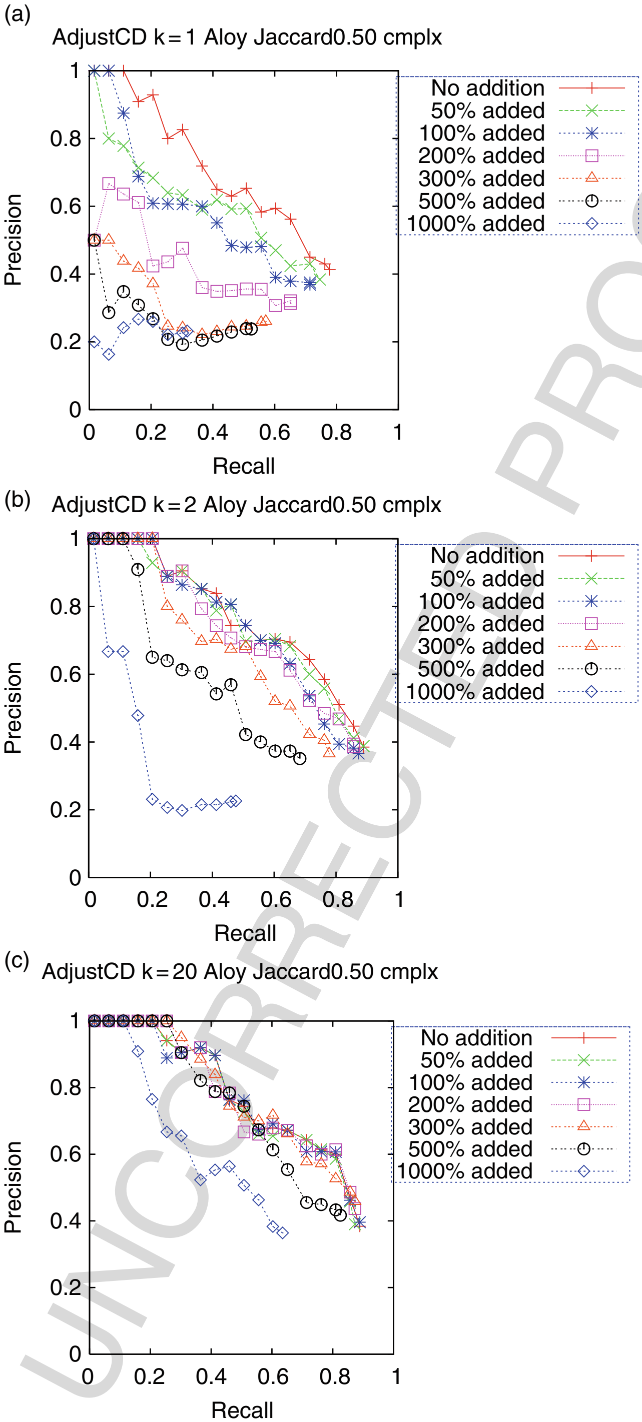


Fig. 8 Precision–recall curves for Aloy reference set when different amount of interactions are randomly added. Overlap thres=0.5, match thres=0.5. Figure from [53]

hold promise for the eventual goal of reliable characterization of protein functionality in a high-throughput fashion. Protein-protein interaction data is an important source of data for these approaches, and research on the analysis and processing of protein–protein interactions will continue be a key area of research in protein function prediction.

References

1. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Séraphin, B. The Tandem Affinity Purification (TAP) Method: a general procedure of protein complex purification. Methods **24**: 218–229 (2001).
2. Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C., Remor, M., Hofert, 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., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**: 141–147 (2001).
3. Fromont-Racine, M., Rain, J., Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat. Genet. **16**: 277–282 (2001).
4. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., 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**: 1143–1147 (2001).
5. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature **403**: 623–627 (2001).
6. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., J.M. Peregrín-Alvarez, 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., St Onge, P., Ghanny, S., Lam, M.H.Y., 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., Greenblatt, J.F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440**: 637–643 (2001).
7. Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B., Boone, C. Systematic mapping of genetic interaction networks. Annu. Rev. Genet. **43**: 601–625 (2001).
8. Tong, A.H.Y., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W.V., Bussey, H., Andrews, B., Tyers, M., Boone, C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science **294**: 2364–2368 (2001).
9. Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Menard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H., Boone, C. Global mapping of the yeast genetic interaction network. Science **303**: 808–813 (2001).

10. Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y., Cheng, X., Brown, G.W., Boone, C., Andrews, B.J., Hughes, T.R. The synthetic genetic interaction spectrum of essential genes. Nat. Genet. **37**: 1147–1152 (2001).
11. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34**: D535 (2001).
12. Schwikowski, B., Uetz, P., Fields, S., others. A network of protein–protein interactions in yeast. Nat. Biotechnol. **18**: 1257–1261 (2001).
13. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T. Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast **18**: 523–531 (2001).
14. Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F. Prediction of protein function using protein–protein interaction data. J. Comput. Biol. **10**: 947–960 (2001).
15. Letovsky, S. Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics **19**: i197–204 (2001).
16. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A. Global protein function prediction from protein–protein interaction networks. Nat. Biotechnol. **21**: 697–700 (2001).
17. Chua, H.N., Sung, W.K., Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics **22**: 1623–1630 (2001).
18. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.W. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. **32**: 5539–5545 (2001).
19. Samanta, M.S, Liang, P. Predicting protein functions from redundancies in large-scale protein interaction networks, Proc. Natl. Acad. Sci. USA **100**: 12579–12583 (2001).
20. Brun, C., Chevenet, F., Martin, D., Wojcik, J., A. Guénoche, Jacq, B. Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. Genome Biol. **5**: 6–6 (2001).
21. AQ4 Serebriiskii, I.G., Golemis, E.A. *Two-hybrid system and false positives. Two-hybrid systems* (2001).
22. Friedel, C.C., Zimmer, R. Identifying the topology of protein complexes from affinity purification assays. Bioinformatics **25**: 2140–2146 (2009).
23. Zhou, X., Kao, M.J., Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. Proc. Natl. Acad. Sci. USA **99**: 12783–12788 (2009).
24. Chua, H., Sung, W.K., Wong, L. Using indirect protein interactions for the prediction of gene ontology functions. BMC Bioinformatics **8**: S8 (2009).
25. Spirin, V., Mirny, L.A. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA **100**: 12123–12128 (2009).
26. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., Chen, R. Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res. **31**: 2443–2450 (2009).
27. Bader, G.D., Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics **4**: 2 (2009).
28. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**: 2498–2504 (2009).
29. AQ5 Chua, H.N., Ning, K., Sung, W.K., Leong, H.W., Wong, L. Using indirect protein–protein interactions for protein complex prediction. *Computational systems bioinformatics: proceedings of the CSB 2007 conference*. London: Imperial College Press, p. 97 (2009).
30. Chua, H.N., Ning, K., Sung Wing-Kin, Leong, H.W., Wong, L. Using indirect protein–protein interactions for protein complex prediction. J. Bioinform. Comput. Biol. **6**: 435–466 (2009).
31. Brohee, S., van Helden, J. Evaluation of clustering algorithms for protein–protein interaction networks. BMC Bioinformatics **7**: 488 (2009).

32. King, A.D., Przulj, N., Jurisica, I. Protein complex prediction via cost-based clustering. Bioinformatics **20**: 3013–20 (2009).

AQ6 33. Van Dongen, S. Graph clustering by flow simulation. PhD thesis (2000).

34. Blatt, M., Wiseman, S., Domany, E. Superparamagnetic clustering of data. Phys. Rev. Lett. **76**: 3251–3254 (1996).

35. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**: 4569–74 (1996).

36. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dmpelfeld, 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., Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. Nature **440**: 631–636 (1996).

37. 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., Tyers, M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415**: 180–183 (1996).

38. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrn-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.Y., 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., Greenblatt, J.F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440**: 637–643 (2006).

39. Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., Ruepp, A. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res **32**: D41–44 (2004).

40. Tomita, E., Tanaka, A., Takahashi, H. The worst-case time complexity for generating all maximal cliques and computational experiments. Theor. Comput. Sci. **363**: 28–42 (2006).

41. Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell Proteomic. **1**: 349–356 (2002).

42. Deng, M., Sun, F., Chen, T. Assessment of the reliability of protein–protein interactions and protein function prediction. Biocomputing 2003: Proceedings of the Pacific Symposium Hawaii, USA, 3–7 January 2002, p. 140 (2003).

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

44. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J. Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnol. **22**: 78–85 (2004).

45. Huang, H., Jedynak, B.M., Bader, J.S. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. PLoS Comput. Biol. **3**: e214 (2007).

46. Huang, H., Bader, J.S. Precision and recall estimates for two-hybrid screens. Bioinformatics **25**: 372–378 (2009).

47. Gilchrist, M.A., Salter, L.A., Wagner, A. A statistical framework for combining and interpreting proteomic datasets. Bioinformatics **20**: 689–700 (2004).
48. Saito, R., Suzuki, H., Hayashizaki, Y. Construction of reliable protein–protein interaction networks with a new interaction generality measure. Bioinformatics **19**: 756–763 (2003).
49. Goldberg, D.S., Roth, F.P. Assessing experimentally derived interactions in a small world. Proc. Natl. Acad. Sci. USA **100**: 4372–4376 (2003).
50. AQ7 Chen, J., Chua, H.N., Hsu, W., Lee, M.L., Ng, S.K., Saito, R., Sung, W.K., Wong, L. Increasing confidence of protein–protein interactomes. Genome Inform. Ser. **17**: 284 (2006).
51. Chen, J., Hsu, W., Lee, M.L., Ng, S.-K. Discovering reliable protein interactions from high-throughput experimental data using network topology. Artif. Intell. Med. **35**: 37–47 (2005).
52. AQ8 Liu, G., Li, J., Wong, L. Assessing and predicting protein interactions using both local and global network topological metrics. Proc. GIW (2008).
53. Liu, G., Wong, L., Chua, H.N. Complex discovery from weighted PPI networks. Bioinformatics **25**: 1891–1897 (2009).
54. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A., Bork, P., Superti-Furga, G., Serrano, L., Russell, R.B. Structure-based assembly of protein complexes in yeast. Science **303**: 2026–2029 (2004).
55. Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I., Vicsek, T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics **22**: 1021–1023 (2006).
56. Chen, Y., Xu, D. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. Nucleic Acids Res. **32**: 6414–6424 (2004).
57. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc. Natl. Acad. Sci. USA **100**: 8348–8353 (2003).
58. Chua, H.N., Sung, W.K., Wong, L. An efficient strategy for extensive integration of diverse biological data for protein function prediction. Bioinformatics **23**: 3364 (2007).
59. Tian, W., Zhang, L., Tasan, M., Gibbons, F., King, O., Park, J., Wunderlich, Z., Cherry, J.M., Roth, F. Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biol. **9**: S7 (2008).

Chapter 13

Q. No.	Query
AQ1	Please provide email address for “Hon Nian Chua, Guimei Liu, and Limsoon Wong”.
AQ2	Please check all heading levels.
AQ3	“Hon Nian Chua” has been set as a corresponding author. Please check is this ok.
AQ4	Please provide publisher name and location for Ref. 21.
AQ5	Please provide editor name for ref. [29].
AQ6	Please provide organization name for Ref. 33, if applicable.
AQ7	Please provide journal name for Ref. 50.
AQ8	Please provide volume and page number for ref. [52].