BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

A random walk based approach for improving protein-protein interaction network and protein complex prediction

Chengwei Lei [1] and Jianhua Ruan [1]

[1]Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Recent advances in high-throughput technology have dramatically increased the availability of protein-protein interaction (PPI) data and stimulated the development of many methods for predicting protein complexes, which are important in understanding the functional organization of PPI networks. However, automated protein complex prediction from PPI data alone is significantly hindered by the high level of noise, sparseness, and highly skewed degree distribution of PPI networks. Here we present a novel network topology-based algorithm to remove spurious interactions and recover missing ones by computational predictions, and to increase the accuracy of protein complex prediction by reducing the impact of hub nodes. The key idea of our algorithm is that two proteins sharing some high-order topological similarities, which are measured by a novel random walk-based procedure, are likely interacting with each other and may belong to the same protein complex.

Results: Applying our algorithm to a yeast PPI network, we found that the interactions in the reconstructed network have higher biological relevance than in the original network, assessed by multiple types of information, including gene ontology, gene expression, essentiality, conservation between species, and known protein complexes. Comparison with existing methods shows that the network reconstructed by our method has the highest quality. Using two independent graph clustering algorithms, we found that the reconstructed network has resulted in significantly improved prediction accuracy of protein complexes. Furthermore, our method is applicable to PPI networks obtained with different experimental systems such as affinity purification, Y2H, and PCA, and evidence shows that the predicted edges are likely bona fide physical interactions.

Contact: {clei,jruan}@cs.utsa.edu

1 INTRODUCTION

Recent advances in high-throughput techniques such as yeast two-hybrid and tandem affinity purification have enabled the production of a large amount of protein-protein interaction (PPI) data (Krogan *et al*., 2006; Gavin *et al*., 2006; Yu *et al*., 2008; Tarassov *et al*., 2008). These PPI data can be modeled by networks, where nodes in networks represent proteins and edges between the nodes represent physical interactions between proteins. These networks, together with other high-throughput functional genomics data, are offering unprecedented opportunities for both biological and computational scientists to understand the cell at a systems level (Przulj, 2011). For example, global analysis of PPI networks have revealed important connections between topology and function (Jeong *et al*., 2001; Han *et al*., 2004; Yu *et al*., 2007). PPI networks have also been utilized for predicting gene functions, functional pathways, or protein complexes, with both supervised and unsupervised methods (Bader and Hogue, 2002; Wang *et al*., 2007a; King *et al*., 2004; Asthana *et al*., 2004; Wang *et al*., 2010; Ulitsky and Shamir, 2009; Chua *et al*., 2006; Sharan *et al*., 2007; Friedel *et al*., 2009; Lee *et al*., 2008). Furthermore, much effort has been devoted recently towards incorporating PPI networks to obtain a better mechanistic understanding of complex diseases and to improve the diagnosis and treatment of diseases (Chuang *et al*., 2007; Hannum *et al*., 2009; Ideker and Sharan, 2008; Kim *et al*., 2011; Hidalgo *et al*., 2009).

However, the growing size and complexity of PPI networks poses multiple challenges to biologists. First, PPI networks often have a high false positive rate and an even higher false negative rate (Huang *et al*., 2007). Second, PPI networks are typically sparse, partially due to the high false negative rate, which places a hurdle for algorithms that rely on neighbor information, e.g., in gene function prediction (Chua *et al*., 2006; Sharan *et al*., 2007). Third, PPI networks are known to have skewed degree distribution, meaning that they have more than expected quantity of hub genes. Such hub nodes can often reduce the performance of existing graph theoretic algorithms (e.g., for predicting protein complexes) which were often designed for networks with relatively uniform degree distributions.

In this paper, we present a novel idea to improve the quality of a given PPI network by computationally predicting some new interactions and removing spurious edges, utilizing the information only from the input PPI network. In bioinformatics, the only work that is related to ours is by Kuchaiev *et al*. (2009), where they embed a PPI network into a low dimensional geometric space, and assign edges to pairs of nodes that have short distances in the embedded space. In computer science, many methods have been developed to predict missing links from networks (Ruan and Zhang 2006; Radicchi *et al*. 2004; Li and Horvath 2007; Fouss *et al*. 2007; Tong *et al*. 2006, and reviewed in Lü and Zhou 2011). These methods basically fall into two categories: common neighbor-based and distance-based. The first type of methods is based on a simple yet effective idea - two nodes sharing many common neighbors are likely in the same module (Ruan and Zhang, 2006; Radicchi *et al.*, 2004; Li and Horvath, 2007; Wang *et al.*, 2007a). These methods may have limited value on PPI networks which are usually very sparse. The second type of methods measures the distance between pairs of nodes in the network by considering all alternative paths; popular examples include two algorithms based on random walks, namely, Euclidean commute time (ECT) (Fouss *et al.*, 2007) and random walk with restart (RWR) (Tong *et al.*, 2006). ECT measures the number of steps needed for a random walker to travel between two nodes as the distance between them, while RWR computes the probability for a random walker starting from node i to reach another node j. Performance of this type of methods may be significantly affected by hub nodes. Furthermore, nodes that are not directly connected but are otherwise topologically similar / identical (e.g., those that are connected to the same set of hub nodes) may be biologically relevant, but may have very low similarity by such distance-based measurement. Our idea can be considered as a hybrid of both types of methods. It can be considered as an extension of the simple common neighbor-based methods. Basically, we consider two nodes similar if they are topologically similar - i.e., having similar distances to all other nodes in the network (instead of only their direct neighbors). The core of our algorithm is a novel random walk procedure that reduces the impact of hub nodes.

We evaluated our algorithm extensively. We tested it on three yeast PPI networks obtained with different experimental systems, using heterogeneous information sources, including gene ontology annotations, gene expression data, protein complexes, list of essential genes, conservation between species, and a large collection of known physical interactions in the BioGRID database. We also compared our algorithm with three existing methods. The remainder of the paper is organized as follows. In Section 2 we described our algorithm. We present the evaluation results in Section 3 and conclude in Section 4.

2 METHODS

Let $G(V, E)$ be an undirected graph representing a PPI network, with V the set of nodes and E the set of edges. For $v \in V$, let $N(v) = \{u \in V \mid (v, u) \in E\}$ be the set of neighbors of v and $d(v) = |N(v)|$ the degree of v.

The $simple\ random\ walk$ for one node on a graph G is a walk on G where the next node is chosen uniformly at random from the set of neighbors of the current node, i.e., when the walk is at node v, the probability to move in the next step to the neighbor u is $P_{vu} = 1/d(v)$ for $(v, u) \in E$ and 0 otherwise. Assume that a random walk is initiated at an unspecified node v. Let $q_i^{(k)}$ be the probability for a random walker sitting at node i at a discrete time point k. Then, at time point $k + 1$, the probability for the random walker taking the path from node i to node j can be calculated as

$$f_{ij}^{(k+1)} = q_i^{(k)} P_{ij}, \qquad (1)$$

and the probability for the random walker to reach node j at time point $k + 1$ can be calculated as

$$q_j^{(k+1)} = \sum_i f_{ij}^{(k+1)}. \qquad (2)$$

It is important to note that, with this simple random walk, the final (stationary) probability vector converges to the same values regardless of the starting point. Therefore, the stationary probability vectors generated from a simple random walk cannot be used to measure similarity between nodes.

Below we describe an extension to the simple random walk. The key idea is to superimpose a small amount of resistance at each step of a random walk, which will cause the stationary probability vector to be slightly different for each different starting node. This difference is magnified, and the resulting vector can be used as a topological profile of the node. Similarities between pairs of nodes can then be computed based on their topological profiles.

2.1 Random walk with resistance (RWS)

In our algorithm we introduce two types of resistance into the simple random walk model. We replace Equation (1) above by

$$f_{ij}^{(k+1)} = \begin{cases} \max(0, q_i^{(k)} P_{ij} - \epsilon), \text{ if } q_j^{(k)} > 0; \\ \max(0, q_i^{(k)} P_{ij} - \epsilon), \text{ if } q_j^{(k)} = 0 \ \&\ \max_i (q_i^{(k)} P_{ij}) \geq \beta; \\ 0, \quad \text{otherwise.} \end{cases} \qquad (3)$$

The first parameter ϵ is introduced to ensure that the final probability vectors for different starting node will be different. This can be considered as if each edge has some friction resistance and consumes energy. Therefore, whenever a random walker takes a path, the probability f_{ij} will be deducted by a small value, ϵ. The probability will be reset to zero if it is smaller than 0.

The second parameter, β, is introduced to ensure that whenever the random walker is exposed to a new node that she has never visited before, the probability must be large enough for her to actually visit that node. The motivation comes from fluid dynamics where resistance can be caused by surface tension. In order to overcome a surface tension of a fluid, an additional force is required to get expansion. Here, we use this parameter to effectively control the depth of a random walk and reduce the impact from the hub nodes, which tend to reduce the performance of predicting new edges.

In our experiment, ϵ is set to $|V| / |E|^2$ and β is set to $1/ |E|$. This choice is based on an analysis of the minimum and average flow on each edge. Empirically we have found that these two values perform well on multiple, both biological and non-biological, networks. Variations of these two values within a constant multiple do not significantly change the results.

The probability of reaching node j at time point $k + 1$ is then calculated by adding up the probabilities to enter j from all paths, and re-normalized so that the probability vector sums to 1:

$$q_j^{(k+1)} = \sum_i f_{ij}^{(k+1)} / \sum_{ij} f_{ij}^{(k+1)} \qquad (4)$$

The above procedure is applied to each node individually. A random walk is considered to have reached its stationary distribution when the change of its probability vector is less than a small cutoff value. We then stop the procedure for this node and start the next one until all the nodes finish the procedure. In our experiment, all nodes converged in between 5 to 20 iterations.

2.2 Network reconstruction

After applying the above random walk procedure to the network, we have a probability vector for each node. For node i, the probability vector is denoted as ψ_i , which is a $1 \times |V|$ vector, and the whole group of probability vectors can be denoted as a $|V| \times |V|$ matrix Ψ.

To magnify the difference between probability vectors from different nodes, we first obtain the median vector H from all the vectors, where the j-th element of H is defined as H_j = median ($\psi_{i=1\sim|V|,j}$), and calculate the $|V| \times |V|$ offset matrix Θ, where $\Theta_{ij} = \Psi_{ij}$ - H_j.

Then, we calculate the Pearson correlation coefficient between each column of the offset matrix as a measurement of similarity between nodes: C_{ij} = pcc($\Theta_{1\sim|V|,i}$, $\Theta_{1\sim|V|,j}$). Empirically we have found that using each column of the offset matrix as a topological profile of a node works slightly better than if we had used each row as a topological profile. Informally speaking, a row vector represents the information passed from a node to all nodes in the network, while a column vector represents the information that a node receives from the network; therefore, the latter is a more accurate way of describing the position of the node in the network.

Finally, a network is reconstructed from the correlation matrix by connecting pairs of nodes whose similarity is above a certain threshold. Although more sophisticated methods are possible (e.g. Ruan 2009), in this paper we choose to implement a very simple strategy for easy evaluation and fair comparison to other methods: we simply pick a cutoff value so that the number of edges can be kept the same as in the original network. We will show that this simple strategy served as well. In Section 4 we discuss some future plans in improving cutoff selection which should further improve the quality of the reconstructed network, and particularly, reduce the false negative rate of PPI networks.

3 RESULTS AND DISCUSSION

For evaluation, we applied our algorithm to three yeast PPI networks obtained from different technologies: tandem-affinity purification (Krogan *et al*., 2006), yeast two-hybrid (Yu *et al*., 2008) and protein-fragment complementation assay (Tarassov *et al*., 2008). In Section 3.1 – 3.3, we discuss results on the Krogan data set, which is the largest, and in Section 3.4 and 3.5, we present some comparative analysis of the three datasets.

3.1 Reconstructed PPI network has better functional relevance

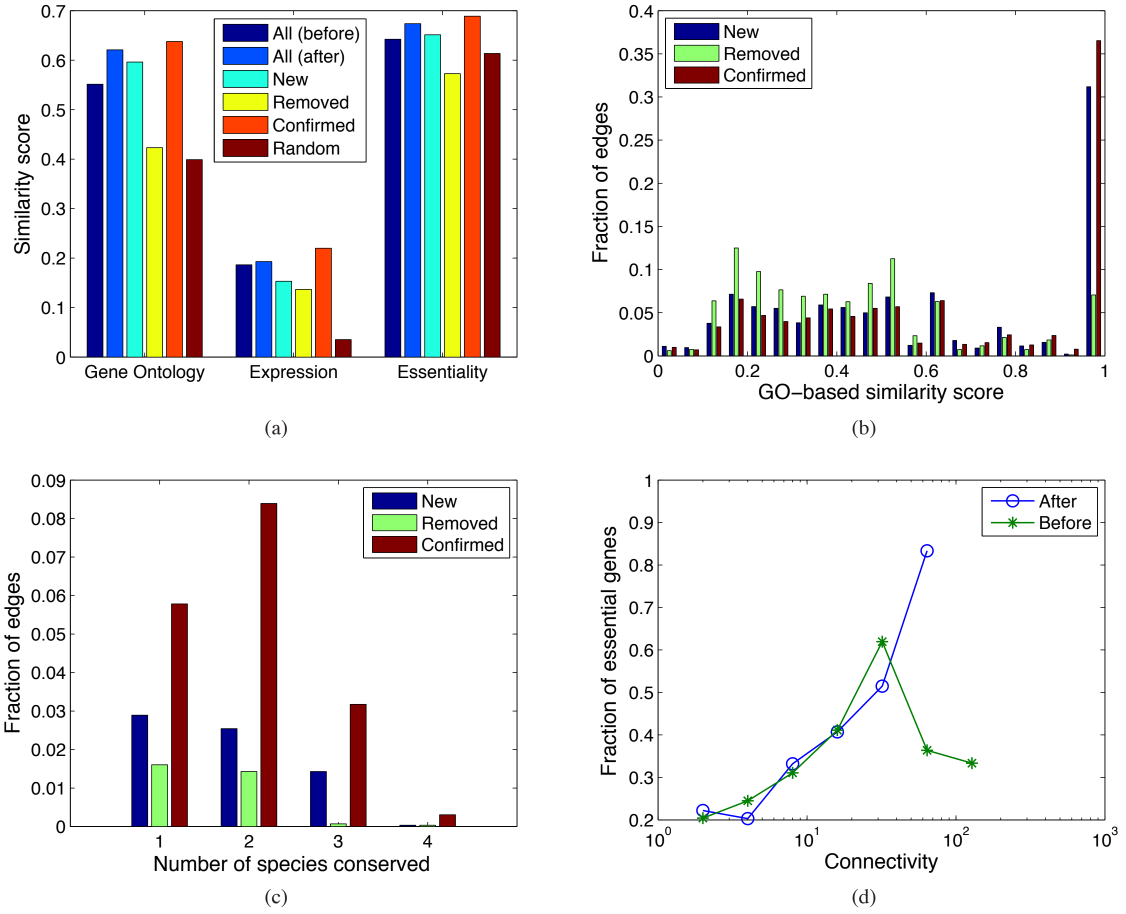
We performed a random walk on the Krogan PPI network, which covers 2708 genes with 7123 edges, and derived a modified PPI network by choosing 7123 potential connections with the highest similarities (see Methods). Within the modified network, 2870 (40%) edges are new (and the same number of edges in the original network have been removed). To evaluate the functional relevance of the newly predicted edges, we resort to several types of sources, including gene ontology, gene expression, essentiality, known protein complexes and conservation of interactions in other species. To facilitate discussion, we call the group of edges present in the original network “before” group, and that in the modified network “after” group. Furthermore, “new” edges designate the edges that are in “after” but not “before” group, “removed” edges are “before” but not “after”. Finally, those present in both “before” and “after” are called “confirmed”. We also generated random networks that have the same number of edges with a procedure that preserves the degree of each node.

Since interacting proteins are likely involved in similar biological processes, they are expected to have similar functional annotations in gene ontology and similar gene expression patterns across diverse conditions. Therefore, we measure the functional relevance between any pair of genes that are connected by an edge using the semantic similarity between the GO terms annotated with the proteins, using a popular method (Wang *et al*., 2007b; Yu *et al*., 2010). Results shown are based on the “Molecular Function” branch of Gene Ontology. Using “Biological Process” yielded very similar values, and “Cellular Localization” resulted in slightly lower but consistent values (data not shown). We also measured the Pearson correlation coefficient between the gene expression profiles of every pair of genes, using the yeast stress response microarray data (Gasch *et al*., 2000). We used the average similarity of the pairs of nodes connected by an edge in a certain group to represent the functional relevance of that edge group. As shown in Fig. 1 (a), the after group has a higher functional relevance than the before group based on both GO and gene expression. Moreover, the confirmed group has the highest functional similarity compared to the other groups, and the removed group is far lower than the new group. The standard error of these average measurements are all below 1e-5 and therefore these differences are highly significant. Further investigation showed that the GO-based similarity is $>$ 0.95 for the 36% and 32% of edges in the confirmed and new groups, respectively. In contrast, only 7% of removed edges have a GO-based similarity > 0.95 (Fig. 1(b)).

Next, we used essential genes to compare different edge groups. The list of essential genes in yeast is retrieved from the Saccharomyces Genome Database (Dwight *et al*., 2004). As two interacting proteins may belong to the same protein complex, they tend to have the same essentiality. In other words, if one is (not) essential, the other is also expected to be (not) essential. As shown in Fig. 1 (a), the percentage of the removed edges that share the same essentiality is actually lower than that of the randomly generated edges, which suggests that the removed edges are probably connecting genes in different complexes (also see Section 3.2). In contrast, the measurement for the new edges is close to that of the confirmed PPIs.

We also looked at the conservation of the edges in other species. We downloaded conserved PPIs between yeast and four species including C. elegans, fly, mouse, and human from InteroLogFinder (`http://www.interologfinder.org/`)(Wiles *et al*., 2010). As shown in Fig. 1(c), a considerable fraction of confirmed edges are conserved in at least two other species. While a small fraction of the removed edges are conserved in one or two species, they are rarely conserved in more than two species. In comparison, the new edges tend to be more conserved than the removed edges, although not as much as the confirmed ones. The conservation analysis also suggests that the predicted edges are bona fide physical interactions rather than functional links (see also Section 3.5).

Finally, it has been shown that genes with high connectivity in the PPI network tend to be more essential, but it is also known that connectivity and essentiality (percentage of genes that are essential) are only weakly correlated (Jeong *et al*., 2001). For



(a)

(b)

(c)

(d)

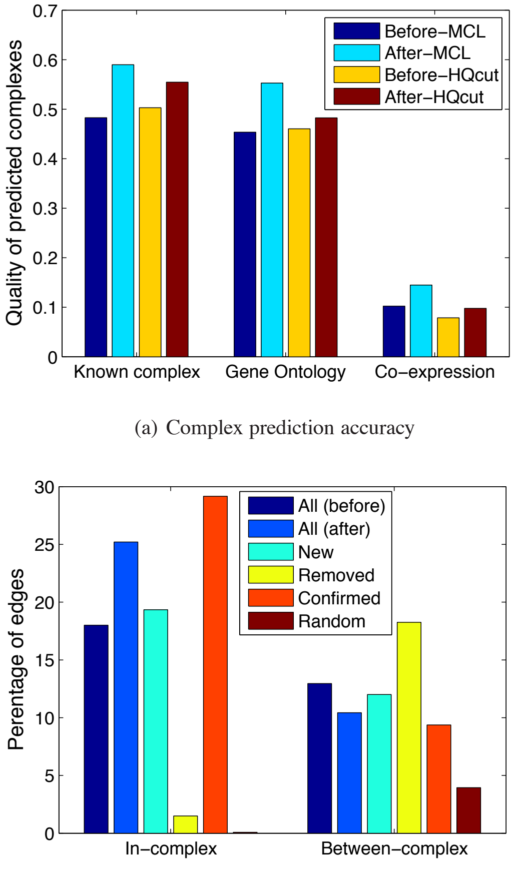
Fig. 1. Functional relevance of different edge groups

example, Fig. 1(d) shows that although the essentiality of genes are generally increasing for genes with low to intermediate degrees, the essentiality of genes with the highest degrees are relatively lower than expected by their degrees. Besides several possible explanations, it may be that some of the proteins with the highest degrees may be "sticky" in the sense that they may appear to interact with many proteins under the experimental protocol, but these interactions do not exist in reality because the protein is generally poorly expressed or not co-expressed with the proteins they potentially interact with. It can be seen that in the reconstructed network, essentiality and degree have a much better correlation.

In summary, using multiple independent sources of evidence, we have shown that compared to the removed edges, the new edges have higher functional relevance. These results suggest that our algorithm can indeed reduce the noise in PPI network and improve the network quality.

3.2 Reconstructed PPI network improves accuracy of protein complex prediction

We investigated whether the improved PPI network can also improve the prediction accuracy of protein complexes. We applied two network clustering algorithms to the original and modified PPI networks, and compared the predicted complexes with the MIPS known protein complexes (Mewes *et al.*, 2006), which included 767 proteins in 170 known complexes after intersecting with the PPI network. MCL is a well-known graph clustering algorithm and has been shown to outperform other protein complex prediction algorithms in two independent evaluation studies (Brohee and van Helden, 2006; Vlasblom and Wodak, 2009). HQcut is a community discovery algorithm developed by one of the co-authors, based on the optimization of a so-called modularity function. For MCL, we set the inflation parameter to 1.8 as suggested by others (Brohee and van Helden, 2006). HQcut does not require any user-tuned parameters. To measure the accuracy of the prediction, we used the Fowlkes-Mallows index for comparing clustering (Meila, 2005; Fowlkes and Mallows, 1983). Formally, let A be the list of gene pairs that fall into the same complex in the set of predicted complexes and B that in the set of known complexes, the prediction accuracy is measured by $|A \cap B| / \sqrt{|A| \times |B|}$, where $|A|$ denotes the cardinality of the set A. As shown in Fig. 2(a), the prediction accuracy is significantly improved for both MCL and HQcut, demonstrating that the improvement is general. Moreover, as the MIPS database of known protein complexes only covers $< 30\%$ of the proteins in the PPI network, we measured the average pairwise functional similarity using gene ontology semantic similarity and co-expression (see Section 3.1) between every pair of nodes that are



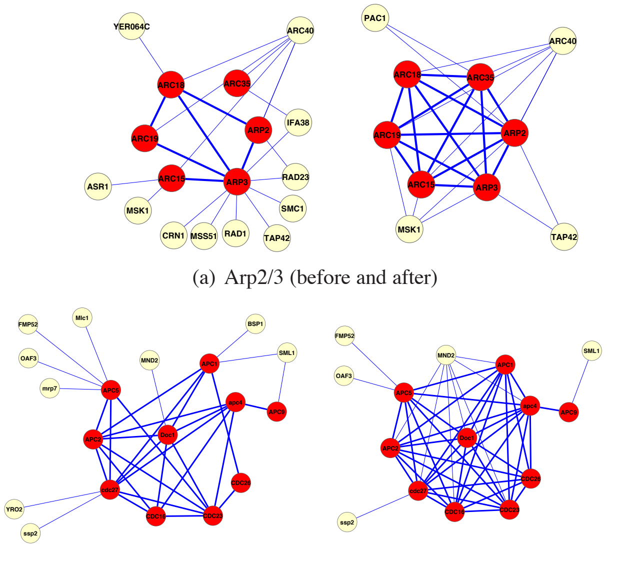
(a) Complex prediction accuracy

(b) In- vs. between-complex edges

Fig. 2. Evaluation using known protein complexes

predicted to be in the same complex. Again, it is shown that the results are improved significantly in the modified network for both MCL and HQcut (Fig. 2 (a)).

To further investigate why the reconstructed network can result in better prediction accuracy of protein complexes, we directly compared different edge groups for the fraction of edges that are connecting genes in the same known complex (in-complex) versus those that are in different known complexes (between-complex). Indeed, as shown in Fig. 2(b), the new edges have much higher in-complex probability and lower between-complex probability compared to the removed edges, while the confirmed edges have the highest in-complex probability and lowest between-complex probability. Therefore, it is likely that the reconstructed PPI network can be combined with any existing protein complex prediction algorithm and improve its accuracy. Fig. 3 shows two known protein complexes. The prediction for the Arp2/3 complex is improved in the after network, because connectivity is increased within the complex and many between-complex edges are removed. Interestingly, for the APC complex, our algorithm not only removed several external edges and added many in-complex edges, but also predicted interactions between a non-member protein, MND2, and the complex members. It turns out that MND2 is indeed a member of the APC complex (Hall *et al*., 2003).



(a) Arp2/3 (before and after)

(b) APC (before and after)

Fig. 3. Example complexes. Red nodes are known members of the complexes.

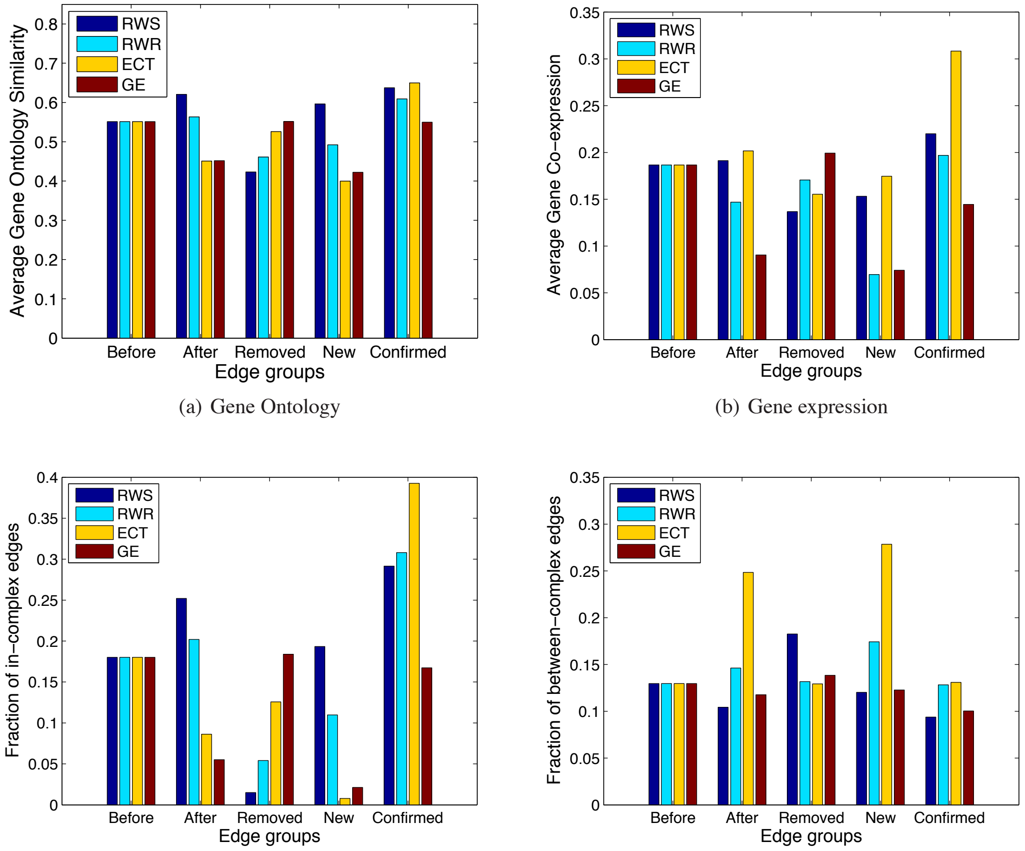
3.3 Comparison with previous network link prediction methods

We compared our algorithm with three existing methods, including two random walk-based algorithms, namely Euclidean commute time (ECT) (Fouss *et al*., 2007) and random walk with restart (RWR) (Tong *et al*., 2006), and a geometric embedding method (GE) (Kuchaiev *et al*., 2009). The ECT and RWR methods are well-known in data mining and network analysis communities, while the GE method was proposed for essentially the same purpose of our study - to improve the quality of PPI networks (see Introduction). As all three algorithms give some topology-based similarity measure of pairs of network nodes, for each algorithm we took the top 7123 pairs of genes having the highest similarities as the predicted PPIs. We then compared the functional relevance of the PPIs falling in different edge groups as in Section 3.1. As shown in Fig. 4, our algorithm outperforms the existing algorithms according to GO and known complexes, in that the "after" network produced by our method has the highest GO similarity, highest fraction of in-complex edges, and lowest fraction of between-complex edges. Analysis of the other three edge groups (removed, new, and confirmed) also shows consistent results. In fact, our algorithm is the only one that shows consistent improvement over the before network using all measurements.

Interestingly, it appears that the confirmed edges by ECT have a high co-expression and a large fraction of in-complex edges (Fig. 4 (b)). This may be partially explained by the fact that ECT predicted (and removed) significantly more edges than RWR and RWS (Table 1) and as a result has a much smaller number of confirmed edges. Nevertheless, it may also suggest that the ECT algorithm performs well in preserving PPIs that are not only having similar functions, but are also highly coexpressed. The geometric embedding method appears to perform well in keeping a low fraction of between-complex edges.

Table 1. Changes to network statistics by different algorithms

	RWS	RWR	ECT	GE
Number of nodes / edges before	2708 / 7123	2708 / 7123	2708 / 7123	2708 / 7123
Number of nodes / edges after	2549 / 7123	2708 / 7123	2016 / 7123	2241 / 7123
Number of replaced edges	2870 (40.3%)	2795 (39.2%)	5671 (79.6%)	5468 (76.8%)



(a) Gene Ontology

(b) Gene expression

(c) In-complex edges

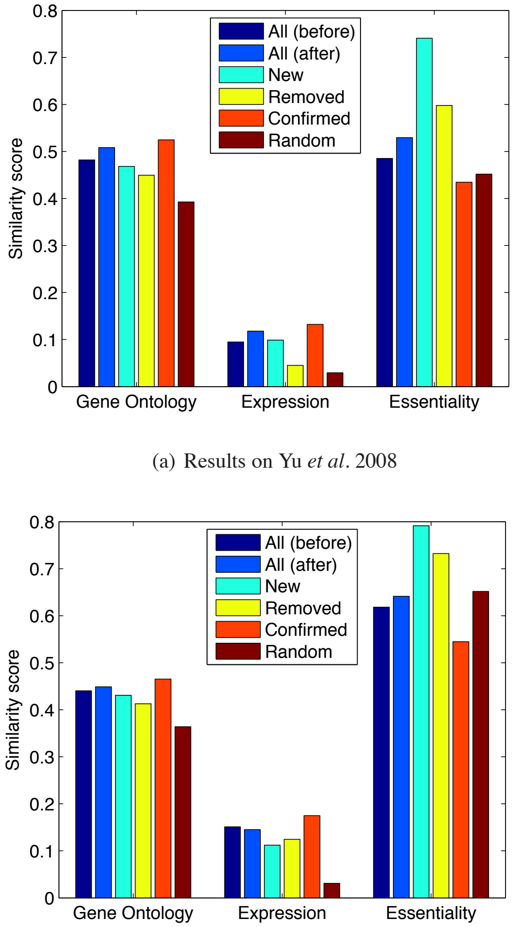
(d) Between-complex edges

Fig. 4. Comparison with other algorithms

3.4 Applicability to other types of PPI networks

We also applied our method to two other data sets, obtained by Yu *et al*. (2008) and Tarassov *et al*. (2008), using yeast two-hybrid (Y2H) and protein-fragment complementation assay (PCA), respectively. While the affinity purification (AP) technique used in Krogan *et al*. (2006) is designed to capture co-complex memberships, Y2H and PCA directly detect binary interactions, and were shown to have higher false negative rate but lower false positive rate than AP (Yu *et al*., 2008; Tarassov *et al*., 2008).

The original Yu and Tarassov networks cover 1278, and 1124 genes with 1641 and 2513 edges, respectively, excluding self-interacting edges. With our method, we were able to replace (predict and remove) 706 (43.0%) and 1203 (47.9%) of the edges for the two networks, respectively. As shown in Fig. 5, the evaluation results based on Gene Ontology for these two data sets and the evaluation results based on co-expression for the Yu data set are consistent with those for the Krogan data set, confirming the general applicability of our method to PPI networks regardless of the experimental systems used to infer them. On the other hand, the differences between the predicted and removed edges are smaller in Yu/Tarassov data than in Krogan data. In fact, for the Tarassov data, the removed edges have slightly better co-expression than the predicted ones; neverthless both are lower than the confirmed edges and significantly higher than the randomly predicted ones. These deviations from the Krogan data set likely reflect the lower false positive rate of the Y2H and PCA data compared to AP data (Yu *et al*., 2008; Tarassov *et al*., 2008). As our results consistently showed that the removed edges have much higher functional relevance than random predictions, chances are that many of the removed edges are not really false positives - they just tend to contain more false positive edges than the confirmed group of edges. To reiterate, we chose to keep the original number of edges in the predicted networks to facilitate an



(a) Results on Yu *et al*. 2008

(b) Results on Tarassov *et al*. 2008

Fig. 5. Results on two additional data sets

unbiased comparison of different approaches, as otherwise changing parameters may signficantly alter and bias the evalution outcomes. In practice, because of the high false negative rate in PPI networks, especially for Y2H and PCA based networks, one would prefer to choose a lower similarity threshold to make more predictions and remove fewer edges than we have done here. As mentioned in Methods and later in Conclusions, we are aware of and are developing better ways to select cutoffs in order to improve the coverage of PPI networks, which however is out of the scope in this paper.

Similarly as in Krogan data, the predicted edges in Yu/Tarassov data have higher co-essentiality than the removed edges, suggesting that the predicted ones are more likely to be within-complex than the removed ones. On the other hand, while the confirmed edges in Krogan data have very high co-essentiality, those in the Yu/Tarassov data have much lower co-essentiality, indicating that a large portion of between-complex edges in these two data sets are preserved by our algorithm. As shown in Yu *et al*. (2008), Y2H and PCA usually detect more between-complex edges than AP. Our algorithm nicely preserved this property.

3.5 Predicted edges are bona fide physical interactions

Finally, it is interesting to ask whether the edges predicted by our algorithm are bona fide physical interactions or simply functional interactions. To answer this question, we validated the edges predicted by our algorithm using all physical interactions present in the BioGRID database (Stark *et al*., 2011). Table 2 shows the number of edges in different groups validated in BioGRID, both for all physical interactions and for edges categorized according to experimental systems. When considering all physical interactions together, Krogan data has the highest percentage of validated predictions (36.0%, or 1033/2870). For Yu and Tarassov data, 9.2% (65/706) and 11.2% (135/1203) predicted edges are present in BioGRID. In comparison, the validation rate for random predictions are below 1%. For Krogan data, the predicted edges have a much higher validation rate than the removed ones. In contrast, the validation rate for the predicted edges is similar as or lower than that for the removed ones for Tarassov and Yu data. As discussed in Section 3.4, this confirms that Y2H and PCA data have lower false positive rate than AP; therefore for these two data sets a lower cutoff may be used to improve the coverage. On the other hand, the relatively low validation rate of the confirmed edges for the Yu and Tarassov data may also reflect a bias caused by the imbalanced data in the BioGRID database (see next paragraph). Furthermore, as the BioGRID PPI data may still have a high false negative rate, it is likely that the real validation rate is higher.

It is known that the three experimental systems (AP, Y2H and PCA) have different characteristics and produce complementary, largely disjoint, results (Yu *et al*., 2008; Tarassov *et al*., 2008). We therefore grouped the edges in BioGRID into four categories: affinity purification-based (AP), yeast two-hybrid (Y2H), PCA, and Other. We then validated our results using the BioGRID interactions within each specific category. As shown in Table 2, except for case of Yu data validated by Y2H, the predicted edges always have a validation rate that is higher than the removed ones. In all cases, the validation rates for confirmed / removed / predicted edges are much higher than random, suggesting that the predicted edges are bona fide physical interactions, and reconfirming that the removed ones are not necessarily false positives, especially for Y2H and PCA-based data. Interestingly, it seems that the different characteristics in different experimental systems are carried over to the predicted edges. For example, the predicted edges for Krogan data is mostly validated by the AP-based interactions. Moreover, while Y2H only covers 12.6% of edges in BioGRID, it accounts for 53.9% (35/65) of the validated predictions in Yu data. Similarly, PCA only contributes 7.1% of the edges in BioGRID, but accounts for 17.0% of the validated predictions in Tarassov data. Therefore, the relatively low validation rate for the Yu and Tarassov data may be partially explained by the insufficient presence of Y2H and PCA data in the database.

4 CONCLUSIONS

In this paper we have presented a novel network topology-based algorithm to improve the quality of PPI networks which in turn can improve the prediction accuracy of protein complexes. The key idea of our algorithm is that two proteins sharing some high-order topological similarities, which are measured by a novel random walk-based procedure, are likely interacting with each other and

Table 2. Validation by BioGRID, breaking down according to experimental systems. Values in parentheses are fractions of edges validated.

		BioGRID Physical Interactions				
	# edges	All	AP	Y2H	PCA	Other
BioGRID	73929	73929 (1.00)	52842 (0.71)	9303 (0.12)	5237 (0.07)	13080 (0.18)
Krogan						
Predicted	2870	1033 (0.36)	983 (0.34)	134 (0.05)	35 (0.01)	180 (0.06)
Removed	2870	412 (0.14)	393 (0.14)	48 (0.02)	16 (0.01)	67 (0.02)
Confirmed	4253	2732 (0.64)	2682 (0.63)	595 (0.14)	165 (0.04)	770 (0.18)
Random	7123	54 (0.01)	33 (0.00)	13 (0.00)	1 (0.00)	13 (0.00)
Yu						
Predicted	706	65 (0.09)	52 (0.07)	35 (0.05)	4 (0.01)	20 (0.03)
Removed	706	114 (0.16)	53 (0.07)	85 (0.12)	3 (0.00)	19 (0.03)
Confirmed	935	405 (0.43)	304 (0.33)	290 (0.31)	35 (0.04)	169 (0.18)
Random	1641	7 (0.00)	3 (0.00)	1 (0.00)	0 (0.00)	3 (0.00)
Tarassov						
Predicted	1203	135 (0.11)	96 (0.08)	39 (0.03)	23 (0.02)	42 (0.03)
Removed	1203	109 (0.09)	81 (0.07)	22 (0.02)	15 (0.01)	26 (0.02)
Confirmed	1333	352 (0.26)	301 (0.23)	143 (0.11)	34 (0.03)	121 (0.09)
Random	2536	24 (0.01)	18 (0.01)	2 (0.00)	1 (0.00)	6 (0.00)

Note: prior to evaluating a particular data set, all interactions in BioGRID from the corresponding publication were removed.

may belong to the same protein complex. Overall, the reconstructed yeast PPI network has much higher biological relevance than the original network, and better quality than those reconstructed by several existing algorithms, assessed by multiple types of information, including gene ontology, gene expression, essentiality, and conservation between species. The reconstructed network has also resulted in significantly improved protein complex prediction accuracy using two different algorithms. Furthermore, our method is applicable to PPI networks obtained with different experimental systems such as yeast two-hybrid, affinity purification based, and protein-fragment complementation assay, and evidence shows that the predicted edges are likely bona fide physical interactions.

Our method may be improved in several directions. For example, in order to derive a network from the topology-based similarity matrix, we have used a simple cutoff-based strategy to maintain the number of edges in the original network. We made this choice to facilitate a fair evaluation of different network reconstruction / clustering algorithms. In fact, we have found that many edges with similarity slightly below the cutoff also have higher biological relevance than those removed. This is also biologically understandable - the original PPI network has a higher false negative rate than false positive rate (Huang *et al*., 2007; Kuchaiev *et al*., 2009). In future work, it may be worthwhile to develop methods that can guide the selection of a more appropriate cutoff which would allow more functionally relevant edges being included without introducing too many false positive edges. One possible way is to examine the distribution of the similarity scores of the original edges and non-edges and determine what edge weights might represent a good separation.

It is worth noting that our focus in this paper is a method to improve the quality of a PPI network purely based on the topology of the network, with no additional biological information involved. This ensures that our algorithm can be easily combined with other algorithms that have already been developed for predicting protein complexes or performing PPI-based studies. For example, recently a fairly sophisticated algorithm has been developed for predicting complexes by repeatedly running the RWR algorithm to obtain neighbor information of some seed proteins, and it has been shown that the method significantly outperformed the MCL algorithm (Macropol *et al*., 2009). Given the superior performance of RWS compared to RWR, we believe that their algorithm performance could be further improved by replacing the RWR algorithm with our algorithm. There are also several studies that attempt to combine additional biological information such as gene ontology and gene expression with PPI network for protein complex prediction (Asthana *et al*., 2004; Ulitsky and Shamir, 2009; Wang *et al*., 2010). It should be straightforward to utilize our method in these approaches, by replacing the PPI network with our modified PPI network.

ACKNOWLEDGMENT

Funding: SC3GM086305, U54CA113001, R01CA152063

REFERENCES

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

Bader, G. and Hogue, C. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, **20**, 991–7.

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

Chua, H. N., Sung, W. K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**(13), 1623–1630.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3**, 140–140.

Dwight, S., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Engel, S., Feierbach, B., Fisk, D., Hirschman, J., Hong, E., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Weng, S., Botstein, D., and Cherry, J. (2004). Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform*, **5**, 9–22.

Fouss, F., Pirotte, A., Renders, J.-M., and Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, **19**(3), 355 –369.

Fowlkes, E. and Mallows, C. (1983). A method for comparing two hierarchical clusterings. *J. Amer. Statist. Assoc.*, **78**, 553–569.

Friedel, C., J., K., and R., Z. (2009). Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. *Journal of Computational Biology*, **16**(8), 1–17.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**, 4241–4257.

Gavin, A., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J., Kuster, B., Bork, P., Russell, R., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–6.

Hall, M. C., Torres, M. P., Schroeder, G. K., and Borchers, C. H. (2003). Mnd2 and swm1 are core subunits of the saccharomyces cerevisiae anaphase-promoting complex. *J Biol Chem*, **278**(19), 16698–16705.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995), 88–93.

Hannum, G., Srivas, R., Guenole, A., van Attikum, H., Krogan, N., Karp, R., and Ideker, T. (2009). Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet*, **5**, e1000782.

Hidalgo, C., Blumm, N., Barabasi, A., and Christakis, N. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, **5**, e1000353.

Huang, H., Jedynak, B. M., and Bader, J. S. (2007). Where have all the interactions gone? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol*, **3**(11), e214.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res*, **18**, 644–52.

Jeong, H., Mason, S., Barabasi, A., and Oltvai, Z. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41–2.

Kim, Y., Wuchty, S., and Przytycka, T. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*, **7**, e1001095.

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

Krogan, N., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N. a nd Tikuisis, A., Punna, T., Peregrn-Alvarez, J., Sha les, M., Zhang, X., Davey, M., Robinson, M., Paccana ro, A., Bray, J., Sheung, A., Beattie, B., Richards, D., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M., Vlasblom, J., Wu, S., Or si, C., Collins, S., Chandran, S., Haw, R., Rilstone, J., Gandi, K., Thompson, N., Musso, G., St Onge, P., Ghanny, S., Lam, M., Butland, G., Altaf-Ul, A.M. a nd Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C., Hughes, T., Parkinson, J., Gerstein, M., Wodak, S., Emili, A., and Greenblatt, J. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009). Geometric de-noising of protein-protein interaction networks. *PLoS Comput Biol*, **5**, e1000454.

Lee, K., Chuang, H., Beyer, A., Sung, M., Huh, W., Lee, B., and Ideker, T. (2008). Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res*, **36**, e136.

Li, A. and Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics*, **23**, 222–31.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, **390**(6), 1150 – 1170.

Macropol, K., Can, T., and Singh, A. (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, **10**(1), 283.

Meila, M. (2005). Comparing clusterings: an axiomatic view. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 577–584, New York, NY, USA. ACM Press.

Mewes, H., Frishman, D., Mayer, K., Munsterkotter, M., Noubibou, O., Pagel, P., Rattei, T., Oesterheld, M., Ruepp, A., and Stumpflen, V. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, **34**, D169–172.

Przulj, N. (2011). Protein-protein interactions: making sense of networks via graph-theoretic modeling. *BioEssays*, **33**(2), 115–123.

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

Ruan, J. (2009). A fully automated method for discovering community structures in high dimensional data. In *Proc. of IEEE International Conference on Data Mining (ICDM-09)*, Miami, FL, USA. IEEE.

Ruan, J. and Zhang, W. (2006). Identification and evaluation of weak community structures in networks. In *Proc. National Conf. on AI, (AAAI-06)*, pages 470–475, Boston, MA.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, **3**, 88.

Stark, C., Breitkreutz, B.-J. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, **39**(Database issue), D698–D704.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna Molina, M. M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, **320**(5882), 1465–1470.

Tong, H., Faloutsos, C., and Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 613–622, Washington, DC, USA. IEEE Computer Society.

Ulitsky, I. and Shamir, R. (2009). Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, **25**, 1158–64.

Vlasblom, J. and Wodak, S. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, **10**, 99.

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

Wang, J., Du, Z., Payattakool, R., Yu, P., and Chen, C. (2007b). A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**, 1274–81.

Wang, J., Li, M., Deng, Y., and Pan, Y. (2010). Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, **11 Suppl 3**, S10.

Wiles, A., Doderer, M., (joint-first author), J. R., Gu, T., Ravi, D., Blackman, B., and Bishop, A. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology*, **4**, 36.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, **26**, 976–8.

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

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., Barabasi, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, **322**(5898), 1158684–110.