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

Chengwei Lei and Jianhua Ruan
Department of Computer Science
The University of Texas at San Antonio
San Antonio, TX 78249, USA
Email: {clei,jruan}@cs.utsa.edu

Abstract—**Recent advances in high-throughput technology have dramatically increased the quantity of available 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 protein-protein interaction networks in different biological processes. 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. Applying our algorithm to a yeast protein-protein interaction network, we found that the interactions in the reconstructed PPI network have more significant biological relevance than the original network, assessed by multiple types of information, including gene ontology, gene expression, essentiality, conservation between species, and known protein complexes. Comparison with several existing methods show that the network reconstructed by our method has the highest quality. Finally, using two independent graph clustering algorithms, we found that the reconstructed network has resulted in significantly improved prediction accuracy of protein complexes.**

Keywords-**Protein-protein interaction network; Protein complex; Link prediction; Clustering**

I. 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 [1–4]. 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 [5]. For example, global analysis of PPI networks have revealed important connections between topology and function [6–8]. PPI networks have also been utilized for predicting gene functions, functional pathways, or protein complexes, with both supervised and unsupervised methods [9–13]. 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 [14, 15].

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 [16]. 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 [11]. 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. Our method is partially inspired by the work of Kuchaiev *et al.*, 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 [17]. In computer science, many methods have been developed to predict missing links from networks ([18–20], and reviewed in [21]). 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 [9, 18]. 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) [19] and random walk with restart (RWR) [20]. ECT measures the number

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

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.

To evaluate the performance of our algorithm, we apply it to a yeast PPI network and examine the biological relevance of the predicted and removed PPIs, using multiple information sources. Results show that the predicted PPIs have much higher functional relevance than the removed ones. Comparison with several existing methods mentioned above show that the network reconstructed by our method has the highest overall quality. Furthermore, applying two independent graph clustering algorithms, we found that the reconstructed network has resulted in significantly improved prediction accuracy of protein complexes.

The remainder of the paper is organized as follows. In Section II we described our algorithm. We present the evaluation results in Section III and conclude in Section IV.

II. 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

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

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

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

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.

A. 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

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

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:

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

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.

B. 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 Θ_{ij} = Ψ_{ij} - H_j.

Then, we calculate the Pearson correlation coefficient between each pair of columns 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. [22]), 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. We are currently developing cutoff selection methods to further improve the quality of the reconstructed network, and particularly, reduce the false negative rate of PPI networks.

III. Results and discussion

For evaluation, we applied our algorithm to a yeast core PPI network obtained from [1], which covers 2708 genes with 7123 edges. By performing a random walk on this network and calculating similarities between every pair of nodes based on their topology equivalence, we 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.

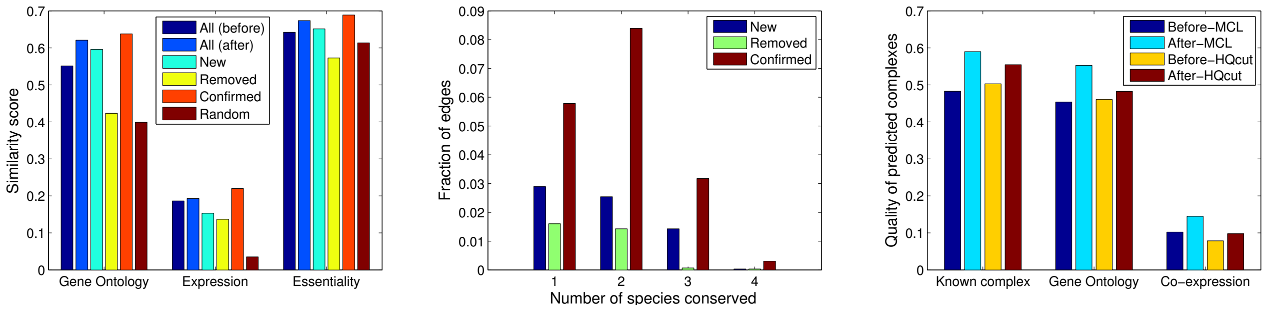
To facilitate discussion, we call the group of edges present in the original network "before" group, and that in the modified network "after" group. Furthermore, "new" edges designate the edges that are in "after" but not "before" group, "removed" edges are "before" but not "after". Finally, those present 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.

A. Reconstructed PPI network has better functional relevance

The most straightforward way to test the performance of the algorithm is to compare the different edge groups for the functional relevance between nodes connected by an edge. If our algorithm indeed reduces noise in the PPI network, we should find the new edges functionally more relevant than the removed ones.

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 [23, 24]. 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 [25]. 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.

Next, we used essential genes to compare different edge groups. The list of essential genes in yeast is retrieved from the Saccharomyces Genome Database [26]. As two



(a) Functional relevance of predicted PPIs

(b) Conservation of predicted PPIs

(c) Complex prediction accuracy

Figure 1. Evaluation results

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. 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/) [27]. As shown in Fig. 1(b), 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 suggests that the predicted edges are likely bona fide physical interactions rather than functional links.

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.

B. 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 [28], 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 [29, 30]. HQcut is a community discovery algorithm developed by one of the co-authors, based on the optimization of a so-called modularity function [31]. For MCL, we set the inflation parameter to 1.8 as suggested by others [29]. HQcut does not require any user-tuned parameters. To measure the accuracy of the prediction, we used the Fowlkes-Mallows index for comparing clustering [32, 33]. 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. 1(c), 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 III-A) between every pair of nodes that are 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. 1(c)).

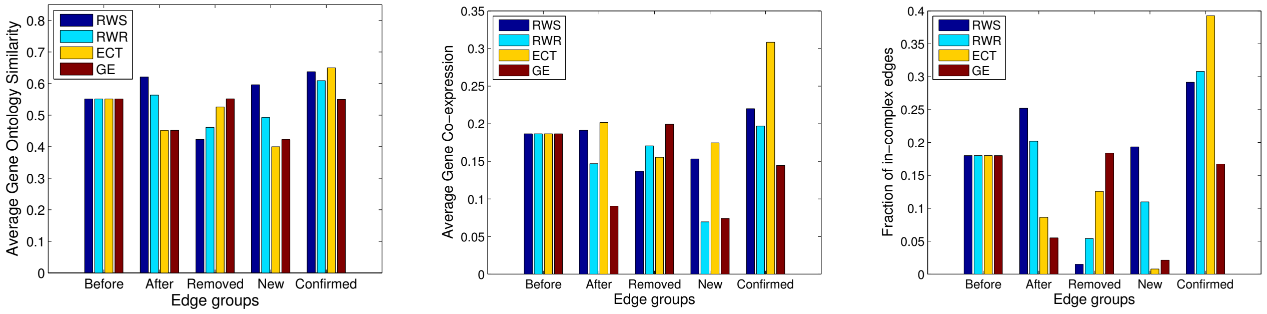
C. Comparison with existing methods

We compared our algorithm with three existing methods, including Euclidean commute time (ECT) [19], random walk with restart (RWR) [20], and a geometric embedding method (GE) [17]. 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 III-A. As shown in Fig. 2, 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 and highest fraction of in-complex edges. Analysis of the other three edge groups (removed, new, and confirmed) also shows

Table I

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

Figure 2. Comparison with other algorithms. Note that ECT predicted a much smaller number of confirmed edges than the other approaches (Table 1), which may partially explain that the edges confirmed by ECT tend to have very high functional relevance scores (see text).

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. 2 (b,c)). 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.

IV. CONCLUSIONS

In this paper we have presented a novel 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 may belong to the same protein complex. Overall, the reconstructed yeast PPI network has much higher biological relevance than the original network, 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. Finally, the PPI network reconstructed by our algorithm has better quality than those reconstructed by several existing algorithms.

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 [34]. With our evaluation results presented here, we believe 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 [10, 13]. It should be straightforward to utilize our method in these approaches, by replacing the PPI network with our modified PPI network.

ACKNOWLEDGMENT

This research was supported partially by NIH grants SC3GM086305, U54CA113001 and R01CA152063 and a grant from the San Antonio Life Science Institutes.

REFERENCES

[1] N.J. Krogan, G. Cagney, H. Yu, et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440:637–643, 2006.

[2] A.C. Gavin, P. Aloy, P. Grandi, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–6, 2006.

[3] H. Yu, P. Braun, M.A. Yildirim, et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 322(5898):1158684–110, 2008.

[4] K. Tarassov, V. Messier, C.R. Landry, et al. An in vivo map of the yeast protein interactome. *Science*, 320(5882):1465–1470, 2008.

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

[6] H. Jeong, S.P. Mason, A.L. Barabasi, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–2, 2001.

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

[8] J.D. Han, N. Bertin, T. Hao, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

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

[10] S Asthana, OD King, FD Gibbons, and FP Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Res*, 14:1170–1175, 2004.

[11] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3:88, 2007.

[12] K. Lee, H.Y. Chuang, A. Beyer, et al. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res*, 36:e136, 2008.

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

[14] H Y Chuang, E Lee, Y T Liu, D Lee, and T Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

[15] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res*, 18:644–52, 2008.

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

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

[18] J. Ruan and W. Zhang. Identification and evaluation of weak community structures in networks. In *Proc. National Conf. on AI (AAAI-06)*, pages 470–5, Boston, MA, 2006.

[19] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. 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, 2007.

[20] H. Tong, C. Faloutsos, and J.-Y. Pan. 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, 2006. IEEE Computer Society.

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

[22] J. Ruan. 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, 2009. IEEE.

[23] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, and C.F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23:1274–81, 2007.

[24] G. Yu, F. Li, Y. Qin, et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26:976–8, 2010.

[25] AP Gasch, PT Spellman, CM Kao, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241–57, 2000.

[26] S.S. Dwight, R. Balakrishnan, K.R. Christie, et al. Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform*, 5:9–22, 2004.

[27] AM Wiles, M Doderer, J Ruan, et al. Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology*, 4:36, 2010.

[28] HW Mewes, D Frishman, KF Mayer, et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, 34:D169–172, 2006.

[29] S. Brohee and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.

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

[31] J Ruan and W Zhang. Identifying network community structures with a high resolution. *Physical Review E*, 77:016104, 2008.

[32] E.B. Fowlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. *J. Amer. Statist. Assoc.*, 78:553–569, 1983.

[33] M. Meila. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98:873–95, 2007.

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