Analysis of Protein-Protein Interaction Networks Using Random Walks

Tolga Can
Dept. of Computer Science
University of California
Santa Barbara, CA
tcan@cs.ucsb.edu

Orhan Çamoğlu
Dept. of Computer Science
University of California
Santa Barbara, CA
orhan@cs.ucsb.edu

Ambuj K. Singh
Dept. of Computer Science
University of California
Santa Barbara, CA
ambuj@cs.ucsb.edu

ABSTRACT

Genome wide protein networks have become reality in recent years due to high throughput methods for detecting protein interactions. Recent studies show that a networked representation of proteins provides a more accurate model of biological systems and processes compared to conventional pairwise analyses. Complementary to the availability of protein networks, various graph analysis techniques have been proposed to mine these networks for pathway discovery, function assignment, and prediction of complex membership. In this paper, we propose using random walks on graphs for the complex/pathway membership problem. We evaluate the proposed technique on three different probabilistic yeast networks using a benchmark dataset of 27 complexes from the MIPS complex catalog database and 10 pathways from the KEGG pathway database. Furthermore, we compare the proposed technique to two other existing techniques both in terms of accuracy and running time performance, thus addressing the scalability issue of such analysis techniques for the first time. Our experiments show that the random walk technique achieves similar or better accuracy with more than 1,000 times speed-up compared to the best competing technique.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*graph algorithms, network problems*; J.3 [**Life and Medical Sciences**]: Biology and Genetics

Keywords

protein networks; random walks on graphs; complex membership; pathway membership

1. INTRODUCTION

Recent developments in genome projects have shown that the complex biological functions of higher organisms are due to combinatorial interactions between their proteins. Therefore, in recent years much effort has gone into finding the complete set of interacting proteins in an organism [22]. Genome-scale protein networks have been realized with the help of high throughput methods, like yeast-two-hybrid (Y2H) [8, 19] and affinity purification with mass spectrometry (APMS) [6, 7]. However, as later studies show, the results from high throughput screens may contain significant number of false positive interactions [22]. Asthana *et al.* [1] assign probabilistic confidence values to experimentally derived interactions using the manually curated catalogs of known complexes in MIPS (Munich Information Center for Protein Sequences) [15] as a trusted reference set. In addition, information integration techniques that utilize indirect genomic evidence have provided both increased genome coverage by predicting new interactions and more accurate associations with multiple supporting evidence [4, 9, 12, 21].

Complementary to the availability of genome-scale protein networks, various graph analysis techniques have been proposed to mine these networks for pathway discovery [3, 17, 24], function assignment [11, 13, 18], and prediction of complex membership [1]. The intrinsic cluster structure of a protein network provides more accurate biological insights compared to local pairwise comparisons. Bader and Hogue [2] propose a clustering algorithm to detect densely connected regions in a protein interaction network for discovering new molecular complexes.

A biologically motivated problem is to predict new members of a partially known protein complex or pathway. In this problem, a particular *core* set of proteins is known, but the biologists are not confident that this core set is complete. The goal is to find a list of candidate proteins, preferably ranked by probability of membership in the partially known complex. As a solution to this problem, Asthana *et al.* [1] proposed a network reliability based technique to find close proximity proteins. They approximate the reliability between two nodes using Monte Carlo simulation, since the exact solution to the network reliability problem is NP-hard [20]. However, the proposed approximation technique is still computationally expensive as the number of samples for accurate reliability estimation of distant nodes can be very high. Therefore, this technique does not scale well for large protein-protein interaction networks. In this paper, as a computationally more efficient alternative, we propose using random walks on graphs for the complex membership problem.

The random walk technique exploits the global structure

BIOKDD '05, August 2005, Chicago, Illinois, USA
Copyright 2005 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

of a network by simulating the behavior of a random walker [14]. The random walker starts on an initial node, i.e., the query node, and moves to a neighboring node based on the probabilities of the connecting edges. The random walker may also choose to teleport to the start node with a certain probability, called the *restart probability.* The walking process is repeated at every time tick for a certain amount of time. At the end, the percentage of time spent on a node gives a notion of its proximity to the query node. Google search engine uses a similar technique to exploit the global hyperlink structure of the Web and produce better rankings of search results [5]. Weston *et al.* [23] use the random walk technique on a protein sequence similarity graph created using PSI-BLAST scores to provide better rankings for a given query protein sequence.

The solution to the problem of finding final rankings of a random walk process can be formulated as an iterative matrix multiplication that provably converges [23]. In addition to providing a computationally much efficient alternative, the matrix formulation also allows for the random walker to start from a *set of nodes* instead of a single node. Therefore, by using the proteins of a partially known complex as the start set, the random walk technique ranks the remaining proteins in the network with respect to their proximity to the query complex. This makes the random walk technique a suitable solution for complex membership problem.

We evaluate the random walk technique on three probabilistic yeast networks using a benchmark dataset of 27 complexes from the MIPS complex catalog database [15] and 10 pathways from the KEGG [10] pathway database. Our experiments show that the ranking results provided by the random walk technique is as accurate as the network reliability technique [1] with more than 1,000 times speed-up.

The rest of the paper is organized as follows. In Section 2, we give technical details of the random walk method for the complex membership problem. In Section 3, we evaluate the proposed technique on three probabilistic yeast networks and present comparative analysis results. We conclude in Section 4.

2. METHODS

In this section, we describe the complex membership problem and present the random walk algorithm as a solution to this problem. We also discuss the competing techniques that are used in the comparative analysis.

Complex membership problem: Given a set of core proteins in a protein complex, the complex membership problem is defined as the problem of finding a set of candidate proteins, ranked according to the probability that each connects to the core complex. A good solution to this problem provides better targets for *in vivo* screening of candidate members of a protein complex. The same solution can be used for predicting candidate members of a partially known pathway if the underlying network captures functional associations as well as protein-protein interactions.

2.1 Random walks on graphs

Let $G = (V, E)$ be the graph representing a protein-protein interaction network, where V is the set of nodes (proteins), and E is the set of weighted undirected edges, where the weight shows the probability of interaction (or functional association) between protein pairs. We define the proximity of a node v to a start node s, $p_s(v)$, as follows:

Definition 2.1. *$p_s(v)$ is the steady state probability that a random walk starting at node s will end at node v.*

```
Input: the similarity network $G = (V, E)$;
       a start node $s$;
       restart probability $c$;
Output: the proximity vector $\vec{p_s}(V)$;

Let $\vec{r_s}(V)$ be the restart vector with 0 for all its entries
          except a 1 for the entry denoted by node $s$;
Let $\mathbf{A}$ be the column normalized adjacency matrix
          defined by E;
Initialize $\vec{p_s}(V) := \vec{r_s}(V)$;
while ($\vec{p_s}(V)$ has not converged)
     $\vec{p_s}(V) := (1 - c)\mathbf{A}\vec{p_s}(V) + c\vec{r_s}(V)$;
```

Figure 1: The iterative algorithm to compute the proximity of all the nodes in the graph to a given start node s.

Random walk method simulates a random walker that starts on a source node, s (or a set of source nodes simultaneously). At every time tick, the walker chooses randomly among the available edges (based on edge weights), or goes back to node s with probability c. The restart probability c enforces a restriction on how far we want the random walker to get away from the start node s. In other words, if c is close to 1, the affinity vector reflects the local structure around s, and as c gets close to 0, a more global view is observed.

The probability $p_s(v)^{(t)}$, describes the probability of finding the random walker at node v at time t. The steady state probability $p_s(v)$ gives a measure of proximity to node s, and can be computed efficiently using iterative matrix operations. Figure 1 shows the iterative algorithm, which provably converges [23]. The number of iterations to converge is closely related to the restart probability c. As c gets smaller the diameter of the observed neighborhood increases, thus the number of iterations to converge gets larger. The convergence check requires the L_1-norm between consecutive $\vec{p_s}(V)$s to be less than a small threshold, e.g., 10^{-12}. In our experiments, for $c = 0.30$ the average number of iterations to converge is around 55. We give the running time performance of the random walk method for different c values in Section 3.

The details of the random walk method can be found in [14]. The main advantage of the random walk method is that it is very fast and therefore applicable to large protein networks. Another advantage is that, the method can be used to compute the proximity of a node to a set of source nodes (not just a single source node). This property is especially beneficial when a core set of members of a pathway or complex is known and the network is queried for candidate members.

2.2 Other techniques for the complex membership problem

Network reliability using Monte Carlo simulation: The solution to the two-terminal network reliability problem can be used to predict functional associations between proteins. In the reliability problem, we have a graph of connections between nodes in which each connection is weighted by the probability that the corresponding wire (edge) is functioning at a given time. The probability that some path of

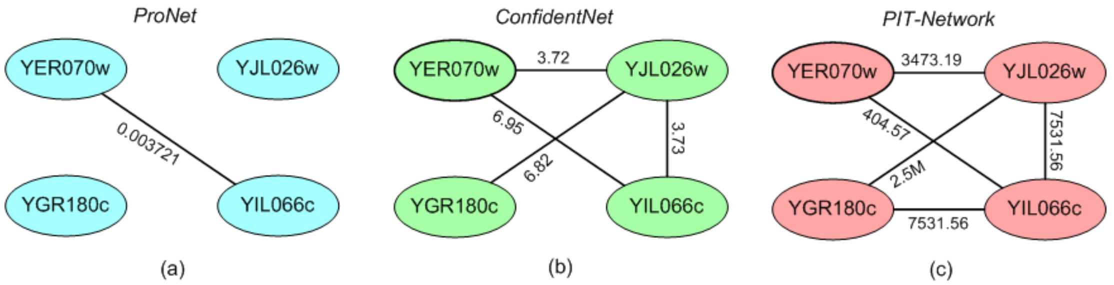

Figure 2: Associations between four members of the Ribonucleoside-diphosphate reductase complex in (a) ProNet, (b) ConfidentNet, and (c) PIT-Network. The edge weights for ProNet are probabilities with prior probability of interaction 0.007. The edge weights for ConfidentNet and PIT-Network are products of likelihoods of individual data sources. The likelihoods of these networks are not directly comparable since they are built using different number of data sources. For the network reliability technique, these likelihoods are normalized to range [0,1].

functioning wires connects the two terminals at a given time gives a measure of proximity between these terminals. The same idea can be extended to discover neighboring proteins in a protein network. The exact solution to the network reliability problem is NP-hard [20]. Monte Carlo simulation [1] is one of the approximation methods proposed for this problem. In this method, a sample of N binary networks from the probabilistic network is created according to a Bernoulli trial on each edge based on its probability. Then, breadth-first search is used to determine the existence of a path between a node in the network and the *core* complex/pathway. For each protein p in the network, the fraction F_i of sampled networks in which there exists a path between i and the core complex/pathway is counted. This process provides a ranking of all the proteins in the network. Unlike the random walk technique, this method does not normalize the incoming edges of a node when computing the *connectivity* of a protein to the core complex/pathway. The two parameters that affect the accuracy of the results and the computational efficiency of the technique are the choice of N (the number of samples) and the maximum depth for breadth-first search. In Section 3, we give accuracy and running time performance results for different values of N.

Markov random field: Markov random field method is based on belief propagation and is used to analyze protein networks by Letovsky and Kasif [13]. The method is originally proposed for function prediction but can be used to predict new members of a partially known complex or pathway. At every iteration, each node receives information about its neighbors' labels and their beliefs on the label. Each node then updates its own belief based on the distribution of its neighbors' beliefs. The updated belief is the probability of having k of M neighbors having the label. Since the belief propagation is an iterative process, nodes may mutually enhance their beliefs in the case of cycles in the network. To avoid such traps, Letovsky and Kasif propose resetting the beliefs every two iterations. The resetting is accomplished by labeling only the nodes with probability higher than some threshold (e.g., 0.8). The Markov random field method is very fast, and the underlying idea of belief propagation is very intuitive. However, there are a number of disadvantages for practical use of this method for the complex membership problem: 1) there are too many parameters to adjust, 2) no formal proof of belief bounds exist, 3) the method needs a large negative label set to suppress propagation of belief to all of the network, and 4) the result provided by the Markov random field is not a ranking but a set of nodes that are predicted to be candidate members of the core complex.

Diffusion kernels: Diffusion kernels provide a global similarity metric for the nodes of a graph. The computation of a diffusion kernel is based on the Gaussian radial basis function kernel [16, 18]. The advantages of the diffusion kernels are: 1) they are suitable for integration of multiple data sources and 2) existing kernel methods, e.g., support-vector machines, can be used for classification. The main disadvantage is that it is a measure between two nodes; therefore, a decision as to which metric should be used to compute similarity of a set of nodes to a single node (e.g., max, average, sum, etc.) is needed. The other disadvantages are: 1) computation of the diffusion kernel is expensive, 2) the only parameter β is not as intuitive as the restart probability in random walks, and 3) the effect of the edge weights on the resulting kernel is unclear. Our efforts to use diffusion kernels for the complex membership problem with default parameters were not successful as the accuracy of the results were very low compared to those of random walk, network reliability, and Markov random field techniques. Kernel methods work best with the optimum parameters whose discovery can be tedious. Therefore, we do not compare the proposed random walk method to the diffusion kernel technique.

In the next section, we evaluate the random walk technique on three probabilistic yeast networks and provide comparative results for the complex membership problem.

3. RESULTS

Many biological studies for identification of functional interactions between proteins have targeted the model organism yeast due to its small genome, extensive genetic information, and well-known biochemistry. Therefore, due to the availability of extensive experimental data, most of the computational studies on construction of protein networks have been on the yeast genome. Below, we describe the probabilistic yeast networks used in our experiments. The first network, ProNet (Asthana *et al.* [1]), is a proba-

Table 1: KEGG pathways used in the experiments.

KEGG pathway id:	Number of pathway members:	Pathway description:
sce00030	27	*Pentose phosphate pathway*
sce00193	30	*ATP synthesis*
sce00510	30	*N-Glycan biosynthesis*
sce00513	15	*High-mannose type N-glycan biosynthesis*
sce00600	18	*Glycosphingolipid metabolism*
sce03020	29	*RNA polymerase*
sce03022	23	*Basal transcription factors*
sce03030	21	*DNA polymerase*
sce03050	32	*Proteasome*
sce03060	10	*Protein export*

bilistic network derived from the results of four large scale experimental interaction detection techniques [6, 7, 8, 19]. ProNet contains 3,112 yeast proteins and 12,594 undirected probabilistic interactions, i.e., edges. The second network, ConfidentNet (Lee *et al.* [12]), is a probabilistic functional network of yeast genes. The associations between proteins are predicted using a Bayesian approach by combining five different information sources: mRNA coexpression, gene-fusions, phylogenetic profiles, co-citation, and protein interaction experiments. ConfidentNet contains 4,681 yeast proteins and 34,000 undirected probabilistic associations. The third network, PIT-Network (probabilistic interactome-total) (Jansen *et al.* [9]), is a combination of predicted and experimental interaction networks using a naive Bayesian approach. The predicted network is constructed using mRNA expression, GO processes, MIPS function, and essentiality data. The experimental network is constructed with the same data sources used in ProNet, but by using a fully connected Bayesian network. PIT-Network contains 2,879 yeast proteins and 24,820 interactions. To illustrate the differences between the three networks, Figure 2 shows associations between the members of a Ribonucleoside-diphosphate reductase complex in ProNet, ConfidentNet, and PIT-Network respectively.

In order to evaluate the performance of the random walk technique for the complex membership problem, we used the 27 MIPS [15] complexes examined by Asthana *et al.* [1] and 10 selected pathways from the KEGG pathway database [10]. Table 1 shows the KEGG pathways used in our experiments. We used the leave-one-out benchmark to assess the accuracy of the analysis techniques. In this benchmark, for each of the complexes and pathways examined, one member protein is left out in turn and the remaining set of member proteins is used as the core complex or the partially known pathway in a membership query. The rank of the left out protein as given by the query results provides a measure of accuracy. A successful analysis method should report the left out protein in top ranks. Therefore, in the accuracy result graphs given below, the fraction of leave-one-out queries in which the left-out protein was found above a threshold rank k is assessed.

Figure 3 and Figure 4 show the comparison results for MIPS complex queries and KEGG pathway queries on ProNet respectively. The result of the Markov random field (MRF) method is depicted as a constant height bar, because MRF method does not return a ranked list, but a set of genes predicted to be members of the complex or the pathway. The size of the set returned by the MRF method

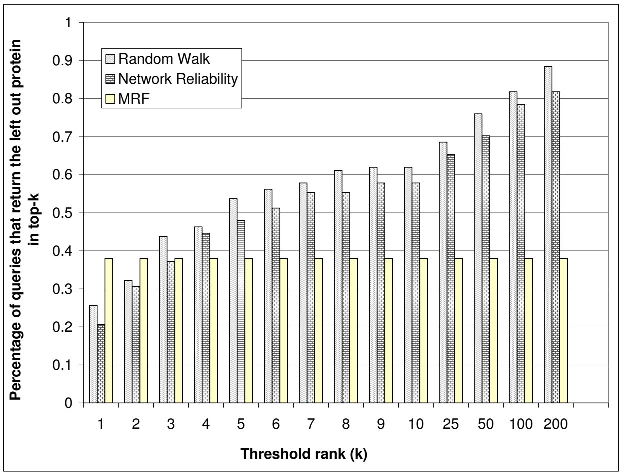

Figure 3: Comparison of analysis methods for protein complex queries on ProNet. The x-axis shows the rank threshold for the left out protein and the y-axis shows the percentage of complex queries (for a total of 121 left-out complex proteins) that the left out protein is found at (or below) the specified rank threshold.

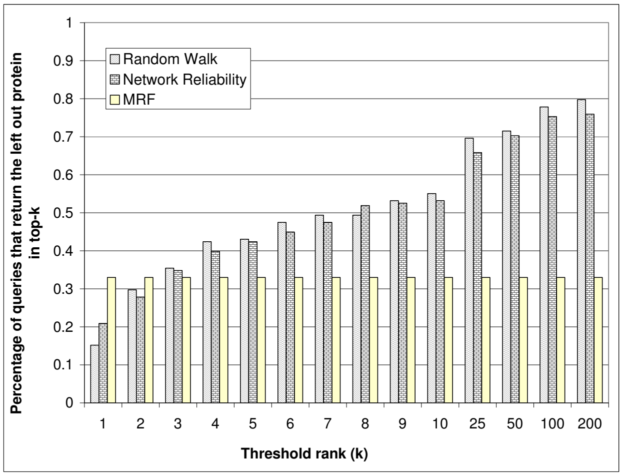

Figure 4: Comparison of analysis methods for KEGG pathway queries on ProNet.

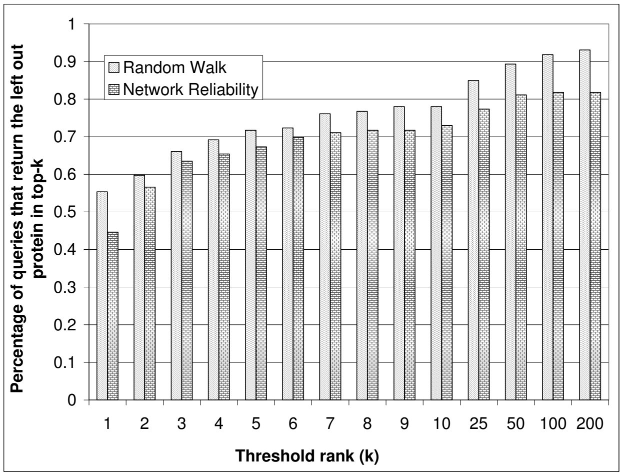

Figure 5: Comparison of random walk and network reliability techniques for MIPS complex queries on ConfidentNet.

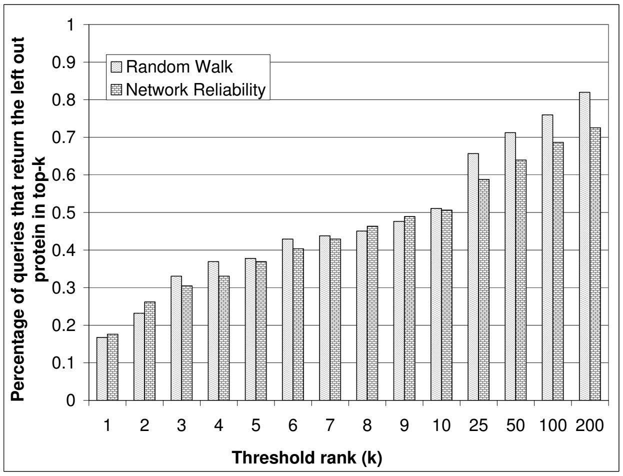

Figure 6: Comparison of random walk and network reliability techniques for KEGG pathway queries on ConfidentNet.

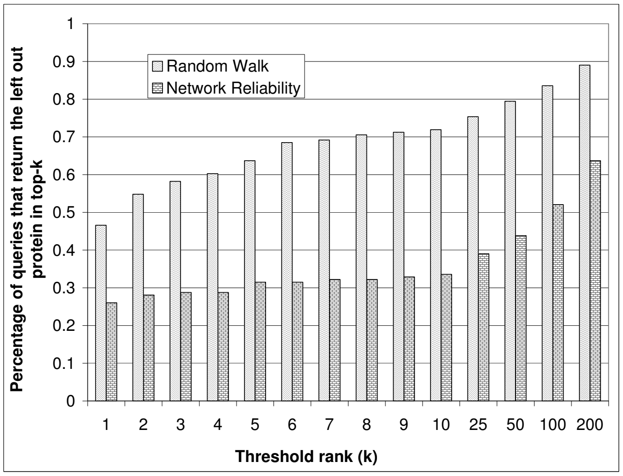

Figure 7: Comparison of random walk and network reliability techniques for MIPS complex queries on PIT-Network.

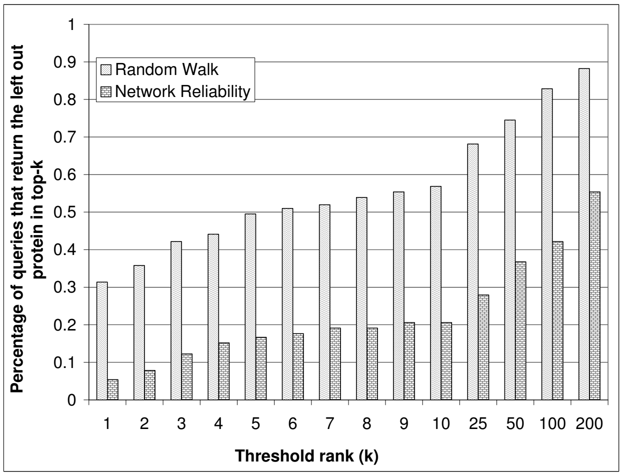

Figure 8: Comparison of random walk and network reliability techniques for KEGG pathway queries on PIT-Network.

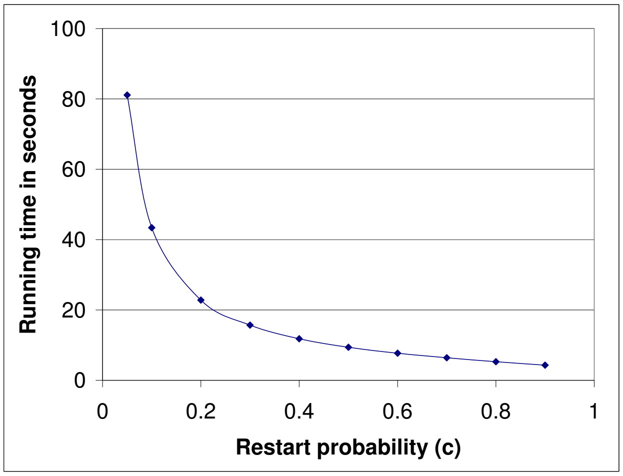

Figure 9: Running time performance of the random walk technique for varying restart probability. The queries are performed on ProNet and the time on y-axis shows the total time to complete all 121 MIPS complex leave-one out queries.

is approximately 300 for the protein networks we consider in this paper. The accuracy ratio indicates the percentage of left out proteins that are correctly predicted to be a member of the core complex/pathway. The results show that the random walk technique has similar or better accuracy compared to the network reliability technique for both complex and pathway queries. In these tests, restart probability of 0.50 was used for the random walk method and sampling size of 10,000 was used for the network reliability by Monte Carlo sampling technique. The slight decrease in the accuracy values for pathway queries is because ProNet captures only direct interactions but not functional associations.

It is clear that the accuracy of any analysis method depends also on the quality of the probabilistic network. Therefore, we performed the same benchmark tests for random walk and network reliability on ConfidentNet and PIT-Network (Figures 5 to 8). These results indicate that, regardless of the network used, random walk technique achieves similar results similar to those of the network reliability technique for the complex/pathway membership problem. One interesting observation is that the network reliability technique performs significantly worse than the random walk technique on the PIT-Network. A possible reason for this finding may be the breadth-first search threshold of 4 that is specially tuned for ProNet. The network reliability technique will perform poorly for graphs on which complex/members are placed farther apart.

Next, we analyze the effect of the restart probability for the random walk method and sample size for the Monte Carlo sampling technique (network reliability) on ProNet for MIPS complex queries. Running time behaviors of these methods on other networks are similar. Also, the running time of Markov random field method is close to that of the random walk method.

Figure 9 and Figure 10 show the running time performances of the random walk method and network reliability by Monte Carlo sampling method respectively. In order to compare the timing results effectively, one needs to find the optimum parameters that gives best accuracy results. Figure 11 and Figure 12 present accuracy results with respect to varying restart probability and sample size (Figure 12 is depicted as a bar graph in order show variable scale values of sample sizes more clearly). Figure 11 shows that the accuracy of the random walk technique is not sensitive to the value of restart probability. The random walk method attains the best accuracy of 54% for restart probability 0.5. On the other hand, the Monte Carlo sampling technique has the best accuracy of 51% for sample sizes 5,000 and 10,000. The running time at sample size of 5,000 is approximately 6 hours for the Monte Carlo sampling technique, whereas random walk technique achieves a better accuracy in only 9.4 seconds. This gives a speed-up of more than 2,000. Even with small sampling sizes, such as 100, where network reliability has acceptable accuracy, random walk is much faster than the Monte Carlo sampling technique, i.e. 9.4 seconds versus 437.81 seconds.

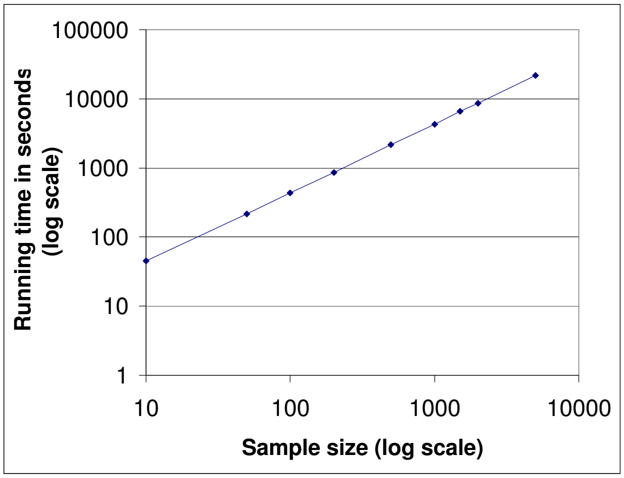

Figure 10: Running time performance of the Monte Carlo sampling approximation to the network reliability problem for varying sample size. Both axes are shown in log scale for better illustration of wide range of values. The queries are performed on ProNet and the time on y-axis shows the total time to complete all 121 MIPS complex leave-one out queries.

4. CONCLUSIONS

In this paper, we proposed using random walks on protein-protein interaction networks for the complex membership problem. We assessed the accuracy of the random walk technique on three different probabilistic yeast networks using a benchmark dataset of 27 complexes from the MIPS complex catalog database and 10 pathways from the KEGG pathway database. We showed that the random walk method is suitable for predicting candidate members of a core complex or partially known pathway. The most prominent property of the random walk technique is its computational efficiency. Our experiments showed that the random walk technique achieves similar or better accuracy with more than 1,000 times speed-up compared to the best competing technique.

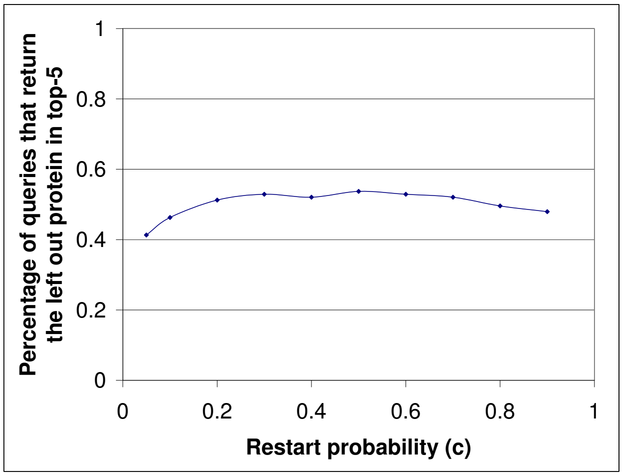

Figure 11: Accuracy of the random walk technique for varying restart probability for top-5 queries. The queries are performed on ProNet and using MIPS complexes.

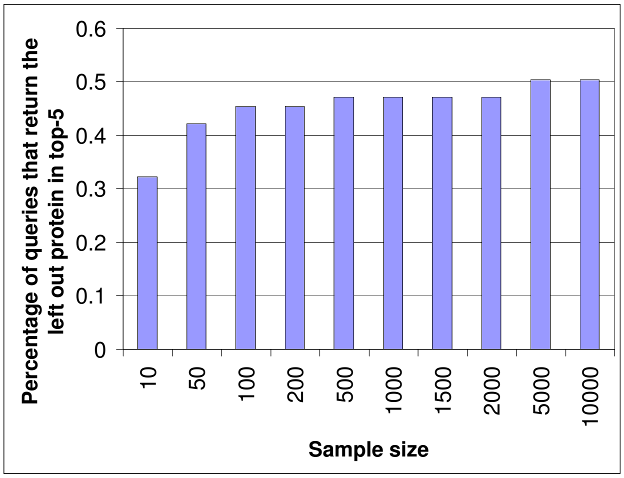

Figure 12: Accuracy of the Monte Carlo sampling technique for varying sample size for top-5 queries. The queries are performed on ProNet and using MIPS complexes.

Therefore, it is a promising method that can scale well for large, genome-scale protein networks.

5. ACKNOWLEDGEMENTS

This research was supported in part by funding from the National Science Foundation: ITR-0331697, DBI-0213903, and EIA-0080134.

6. REFERENCES

[1] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth. Predicting protein complex membership using probabilistic network reliability. *Genome Research*, 14:1170–1175, May 2004.

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

[3] J. S. Bader. Greedily building protein networks with confidence. *Bioinformatics*, 19(15):1869–1874, 2003.

[4] P. M. Bowers, M. Pellegrini, M. J. Thompson, J. Fierro, T. O. Yeates, and D. Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, 5(5):R35, 2004.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[6] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, and C. M. Cruciat. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[7] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, and K. Boutilier. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[8] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.*, 98:4569–4574, 2001.

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

[10] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30:42–46, 2002.

[11] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of PSB*, 2004.

[12] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306:1555–1558, November 2004.

[13] S. Letovsky and S. Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19:i197–i204, 2003.

[14] L. Lovasz. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2:353–398, 1996.

[15] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32:D41–D44, 2004.

[16] B. Schoelkopf, K. Tsuda, and J.-P. Vert, editors. *Kernel methods in computational biology.* MIT Press, 2004.

[17] J. Scott, T. Ideker, R. M. Karp, and R. Sharan.

Efficient algorithms for detecting signaling pathways in protein interaction networks. In *Proceedings of RECOMB*, 2005.

[18] K. Tsuda and W. S. Noble. Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, 20(S1):i326–i333, 2004.

[19] P. Uetz, G. Cagney, T. A. Mansfield, R. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, and P. Pochart. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[20] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8:410–421, 1979.

[21] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33:D433–D437, 2005.

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

[23] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble. Protein ranking: From local to global structure in the protein similarity network. *Proc. Nat. Acad. Sci.*, 101(17):6569–6563, 2004.

[24] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(S1):i363–i370, 2004.