PROTRANK: A BETTER MEASURE FOR PROTEIN ESSENTIALITY

Tolga Can

Department of Computer Engineering, Middle East Technical University
Inonu Bulvari 06531 Ankara Turkey
phone: + (90) 312 210 5537, fax: + (90) 312 210 5544, email: tcan@ceng.metu.edu.tr
web: http://www.ceng.metu.edu.tr/~tcan/

ABSTRACT

In the last few years, genome-scale protein networks of an increasing number of organisms have been available across databases. Many techniques have been developed to analyze these large-scale networks for inferring biological knowledge. In this paper, we propose a graph theoretic measure ProtRank, inspired from Pagerank originally introduced for web page ranking, for ranking proteins in a protein-protein interaction network. We analyze the correlation between ProtRank and protein essentiality. Our results show that ProtRank, which is a global topological measure, is a better indicator of protein essentiality compared to both the local measure of *degree* and another global topological measure of *betweenness centrality*.

1. INTRODUCTION

With the advances in high-throughput technologies in recent years, biological data is accumulating in databases at an ever increasing rate. Sequence, gene expression, functional annotation, subcellular localization, and molecular interaction data is available for almost all of the proteomes for many model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. High-throughput protein interaction assays such as yeast two-hybrid and co-immunoprecipitation provides very valuable information about the underlying molecular machinery that governs the cell. Furthermore, statistical learning techniques enable prediction of interactions using other types of genomic data. These predicted interactions provide confidence weighted and higher coverage interaction data. Experimental and predicted interactions for many organisms are available publicly across databases such as DIP [11], BioGRID [13], BIND [1], MIPS [9], and STRING [14].

Information about protein interactions for an organism is usually represented as a graph in which the nodes of the graph represent individual proteins and the edges represent physical interactions. Graph theoretic approaches are therefore applicable to analyze these networks. In the past years, many techniques have been proposed to mine protein interaction networks for molecular complex and pathway discovery [2,12], function prediction [7], and annotation [6]. Also, there have been many studies that topologically analyze these networks for understanding their degree distribution and connectivity [3], and for finding recurring network motifs [10]. Degree distribution analyses show that all of these networks are scale-free networks with power-law degree distributions [3]. Studies that analyze individual nodes of protein interaction networks for their topological properties, such as number of neighbors (*i.e.*, degree), clustering coefficient, and betweenness centrality, found correlation between these graph theoretic measures and biological properties such as protein essentiality [4,15]. These studies signify the use of these networks as a data source to predict biological properties of novel proteins.

A gene or protein of an organism is essential if it is crucial for the organism's viability. Prediction of essential genes and proteins of an organism is biologically a very appealing task, because such genes and proteins are potential drug targets. If the essential genes or the associated gene products of a pathogen or a diseased cell are known, the drug design process will be considerably more efficient, since the need to screen the complete genome will be eliminated. Previous studies have shown that proteins with higher degree and higher betweenness centrality in the protein interaction network tend to be essential proteins.

In this paper, we propose another graph theoretic measure, ProtRank, to rank the proteins of an organism for a given

protein-protein interaction network. ProtRank, is a global topological measure which indicates the visiting likelihood of a protein. ProtRank is essentially the same measure as the Google's PageRank measure which is used to identify the importance (or popularity) of web pages. ProtRank of a protein is computed by simulating a random walker on the protein-protein interaction network and measuring the amount of time the random walker spends on the corresponding node. We computed ProtRank for all the proteins of the *S. cerevisiae* protein interaction network and analyzed the correlation of ProtRank with protein essentiality. Comparison to degree and betweenness centrality measures shows that ProtRank is a better indicator of protein essentiality.

The rest of the paper is organized as follows. In Section 2, we describe the datasets and the computation details of the ProtRank measure. In Section 3, we present our experimental results and we conclude in Section 4.

2. MATERIALS AND METHODS

In this section, we first describe the datasets used in our experiments, and then give the details of the random walk method for computing ProtRank.

2.1 Materials

The *S. cerevisiae* interaction network is gathered from Yu *et al*. [15] Yu *et al.* assembled the interaction data from a number of different published high-throughput datasets and published databases. They also eliminated noise from the dataset by utilizing independent genomic features and using Bayesian integration. There are 23295 interactions for 4683 *S. cerevisae* proteins in the gathered network.

The essential genes of the *S. cerevisiae* is obtained from Winzeler *et al*.'s work [17], which is a result of the Saccharomyces Genome Deletion Project. Winzeler *et al.* identify 356 *S. cerevisiae* genes as essential.

The betweenness centrality values for the *S. cerevisiae* interaction network used in this study is downloaded as precomputed values from the supplementary material available as part of Yu *et al*.'s work[15][1].

2.2 Methods

In this section, we describe the algorithm to compute ProtRank measure for every protein of a protein-protein interaction network. The computation is based on random walks on graphs.

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be the graph representing a protein-protein interaction network, where $\mathbf{V}$ is the set of nodes (proteins), and $\mathbf{E}$ is the set of unweighted undirected edges. We define the ProtRank of a node $\boldsymbol{v}$, $\mathbf{P}(\boldsymbol{v})$, as follows:

Definition: $\mathbf{P}(\boldsymbol{v})$ is the steady state probability that a random walk will end at node $\boldsymbol{v}$.

Random walk method simulates a random walker that starts on a random node. At every time tick, the walker chooses randomly among the available edges (i.e., the neighbors of the current node), or goes to another random node with probability $\boldsymbol{c}$. The restart probability $\boldsymbol{c}$ prevents the random walker to get stuck in an isolated island of the graph. In our experiments, we have set $\boldsymbol{c}$ to 0.1. We experimented with other values of $\boldsymbol{c}$; however, we found out that the ProtRank of a protein is not very sensitive to the choice of $\boldsymbol{c}$. Therefore, in this paper, we do not report our results for different values of $\boldsymbol{c}$.

The probability $\mathbf{P}(\boldsymbol{v})^{\boldsymbol{t}}$ describes the probability of finding the random walker at node $\boldsymbol{v}$ at time $\boldsymbol{t}$. The steady state probability $\mathbf{P}(\boldsymbol{v})$ gives a measure of how frequent node $\boldsymbol{v}$ is visited and hence its relative popularity among the other nodes. The steady state probability $\mathbf{P}(\boldsymbol{v})$ can be computed efficiently using iterative matrix operations. Below, we give the iterative algorithm to compute ProtRank.

Algorithm: ProtRank
Input: The protein-protein interaction network, $\mathbf{G}=(\mathbf{V},\mathbf{E})$
Restart probability, $\boldsymbol{c}$
Output: The ProtRank vector $\mathbf{P}(\mathbf{V})$ for all the proteins

(1) Let $\mathbf{r}(\mathbf{V})$ be the restart vector with $1/|\mathbf{V}|$ for all its entries
(2) Let $\mathbf{A}$ be the column normalized transpose of the adjacency matrix defined by $\mathbf{G}$
(3) Initialize $\mathbf{P}(\mathbf{V}) := \mathbf{r}(\mathbf{V})$
(4) while $\mathbf{P}(\mathbf{V})$ has not converged
(5) $\quad \mathbf{P}(\mathbf{V}) := (1-\boldsymbol{c})\,\mathbf{A}\,\mathbf{P}(\mathbf{V}) + \boldsymbol{c}\;\mathbf{r}(\mathbf{V})$

In the algorithm given above, $\mathbf{r}(\mathbf{V})$ and $\mathbf{P}(\mathbf{V})$ are one dimensional column vectors and $\mathbf{A}$ is a square matrix, which is the adjacency matrix representation of graph $\mathbf{G}$. The dimensions of $\mathbf{A}$ are $|\mathbf{V}|\times|\mathbf{V}|$. In the adjacency matrix representation, each row and column of the adjacency matrix corresponds to a node of the graph. For a row that represents a node, the edges (i.e. interactions) of that node are represented by entries of 1s at the corresponding columns.

The ProtRank algorithm given above provably converges [16]. The number of iterations to converge is closely related to the restart probability $\boldsymbol{c}$. The convergence check requires the $\mathbf{L}_1$-norm between consecutive $\mathbf{P}(\mathbf{V})$s to be less than a small threshold, e.g.,10^{-10}. In our experiments, for $\boldsymbol{c}$=0.1 the average number of iterations to converge is around 160. The details of the random walk method can be found in [8]. The main advantage of the random walk method is that it is very fast and therefore applicable to large protein networks.

[1] http://www.gersteinlab.org/proj/bottleneck/

3. RESULTS

We computed the ProtRank measure for the 4683 proteins of the *S. cerevisiae* network. The network does not contain all the proteins of the *S. cerevisiae* proteome. As a result, 306 of the 356 essential proteins are present in the network. In order to measure the correlation of ProtRank, degree, and betweenness centrality measures with protein essentiality we use the following approach.

For each measure we have a ranked list of proteins. We analyze the top 400 proteins in the ranked lists and compare these top-400 proteins to the essential proteins one by one from top to bottom. We chose 400, because it is a number close to the number of existing essential proteins in *S. cerevisiae* and it includes about 10% of all proteins in the protein-protein interaction network which is a reasonable percentage to screen as drug targets. At each step we compute the percentage of essential proteins among the top-*k* proteins in the ranked list. A measure which correlates perfectly with essentiality should detect all the essential proteins among the top-400 proteins. Unfortunately, the inherent noise in protein-protein interaction networks, and other factors which effect protein essentiality, we can identify only a small percentage of essential proteins in top-400. Figure 1 shows the result of this analysis. The best identification percentage at top-400 is attained by the ProtRank measure with 19% of the essential proteins identified. Betweenness measure [15] can identify about 12% of the essential proteins and the degree measure [4] can identify about 17% of the essential proteins. It can also be seen from this figure that there is a correlation between the degree measure and the ProtRank measure. However, overall, ProtRank is a better indicator of protein essentiality.

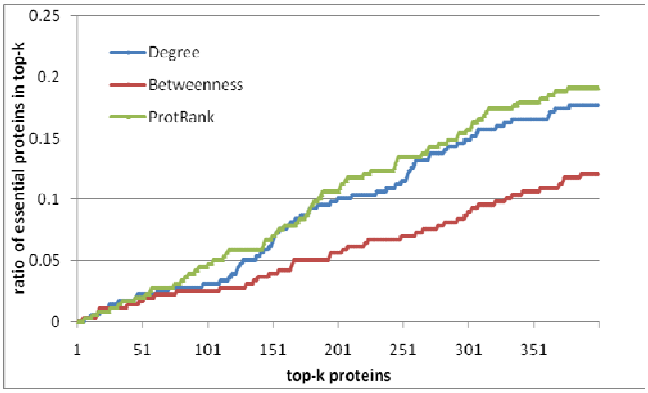

Figure 1. The ratio of essential proteins identified among the top-k proteins

4. CONCLUSIONS

In this paper, we proposed a graph theoretic measure, ProtRank, to rank the proteins of an organism using a protein-protein interaction network. ProtRank, which is inspired from Google's PageRank, is a global topological measure, which indicates the relative visiting frequency of a protein. We computed ProtRank for all the proteins of the *S. cerevisiae* protein interaction network and analyzed the correlation of ProtRank with protein essentiality. Our results showed that ProtRank is a better measure of protein essentiality compared to the previously studied measures of degree and betweenness centrality.

5. ACKNOWLEDGEMENTS

This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Career Program Grant #106E128.

REFERENCES

[1] C. Alfarano *et al.*, "The Biomolecular Interaction Network Database and related tools 2005 update," *Nucleic Acids Research*, 33(S1):D418-D424, 2005.

[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] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks, " *Science*, 286:509-512, 1999.

[4] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, 411:41-42, May 2001.

[5] H. Jeong, Z. N. Oltvai, and A.-L. Barabási, "Prediction of Protein Essentiality Based on Genomic Data," *ComPlexUs*, 1:19-28, 2003.

[6] M. Kirac, G. Ozsoyoglu, and J. Yang, "Annotating proteins by mining protein interaction networks," *Bioinformatics*, 22(14):e260-e270, 2006.

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

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

[9] H. W. Mewes HW, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research* 30(1):31-4, January 2002.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, et. al., "Network motifs: Simple building blocks of complex networks," *Science*, 298:824-827, 2002.

[11] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Research* 32(Database issue):D449-51, 2004.

[12] J. Scott, T. Ideker, R. M. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *Proceedings of the Research in Computational Molecular Biology (RECOMB2005)*, pp. 1-13, 2005.

[13] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, 34:D535-9, January 2006.

[14] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, and P. Bork, "STRING 7 — recent developments in the integration and prediction of protein interactions," *Nucleic Acids Research*, 35:D358-D362, January 2007.

[15] 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 Computational Biology*, 3(4):e59, April 2007.

[16] 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.

[17] E. Winzeler, E., et al., "Functional Characterization of the Saccharomyces cerevisiae Genome by Gene Deletion and Parallel Analysis," *Science*, 285:901-906, 1999.