Identifying essential proteins via integration of protein interaction and gene expression data

Xiwei Tang[1,2], Jianxin Wang*[1], Yi Pan[1,3]
1. School of Information Science and Engineering, Central South University, Changsha, 410083, China
2. School of Information Science and Engineering, Hunan First Normal University, Changsha, 410205, China
3. Department of Computer Science, Georgia State University, Atlanta, GA 30302-4110, USA
*Email: *corresponding author, jxwang@mail.csu.edu.cn; tangxiwei2010@gmail.com; pan@cs.gsu.edu*

***Abstract*—Essential proteins are vital for an organism's viability under a variety of conditions. Computational prediction of essential proteins based on the global protein-protein interaction (PPI) network is severely restricted because of the insufficiency of the PPI data, but fortunately the gene expression profiles help to make up the deficiency.**

In this work, Pearson correlation coefficient (PCC) is used to bridge the gap between PPI and gene expression data. Based on PCC and Edge Clustering Coefficient (ECC), a new centrality measure, i.e., the weighted degree centrality (WDC), is developed to achieve the reliable prediction of essential proteins. WDC is employed to identify essential proteins in the yeast PPI network in order to estimate its performance. For comparison, other prediction technologies are also performed to identify essential proteins. Some evaluation methods are used to analyze the results from various prediction approaches.

The analyses prove that WDC outperforms other state-of-the-art ones. At the same time, the analyses also mean that it is an effective way to predict essential proteins by means of integrating different data sources.

***Keywords*-Protein-protein interaction network; gene expression profiles; Edge clustering coefficient; Pearson correlation coefficient;**

I. Introduction

Essential proteins are vital for the growth and development of an organism under a variety of conditions. The identification of essential proteins is important not only for the understanding of the minimal requirements for cellular life, but also for practical implications, e.g., in detecting bacterial drug and vaccine targets [1].

Some investigators have proposed a series of centrality measures such as degree centrality (DC) [2], betweenness centrality (BC) [3], closeness centrality (CC) [4], subgraph centrality (SC) [5], eigenvector centrality (EC) [6], and information centrality (IC) [7]. The measures are used to discover essential proteins based on network topological features. Analyses have shown that they are significantly better than pseudorandom selection in detecting essential proteins. Subsequently, Wang and his colleagues have proposed recently a new essential proteins discovery method based on edge clustering coefficient, named as neighborhood centrality (NC) [8]. Their experimental results show that NC far outweigh six measures of gene centrality in identifying essential proteins.

However, it is insufficient to use PPI data alone to identify essential proteins. Firstly protein interaction data generated by high-throughput technologies is currently flooded with false interactions [9]. Also protein interaction measurements descend from a certain range of experimental conditions, thus they succeed to identify only a small fraction of all possible protein-protein interactions. In addition PPI networks contain unstable interactions or interactions that take place at different time points, thus the resulting network does not represent the real one but an overlap of many different snapshots [10]. Considering the insufficiency of the PPI data, Li et al. have constructed a weighted protein interaction network to predict essential proteins [11]. But, after their experimental results are compared with those from NC [8], it is clear that the performance of the former is not better than that of the latter. So it is necessary to weight an interaction based on other more effective biological information. Recently, Li *et al.* propose a new method for predicting essential proteins based on the integration of PPI network and gene expression profiles [12].

In the present work, based on the integration of protein interaction network topology features and gene expression profiles for predicting essential proteins in $S.cerevisiae$, a new centrality measure will be introduced. That is, the weighted degree centrality (WDC) will be described, to achieve the reliable prediction of essential proteins from PPI and gene expression data. The prediction results from WDC will also be shown in the article. In addition, it will be proved that the integrated approach we present outperforms other state-of-the-art ones.

II. Method and Results

A. Method

The idea behind the use of PCC in our method is first that PCC is often used to determine the similarity between two sets of gene expression values. Furthermore, He et al. proposed that the majority of essential proteins are involved in one or more essential protein-protein interactions that are distributed uniformly at random along the network edges

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

[13]. There are studies like [14] that indicate that essentiality is a product of the protein complex rather than the individual protein. At the same time, studies have shown that genes showing similar pattern of expression tend to have similar function [15]. The studies mentioned above suggest that it may be reasonable to integrate the two data sources to predict essential proteins.

In order to weight each interaction in yeast PPI network, we will first compute the PCC between two interacting transcripts in PPI network based on their coding gene expression profiles. Second, the edge clustering coefficient (ECC) of the two proteins will also be calculated based on the number of their common neighbors in the PPI network. Based on the PCC and ECC, the weighing scoring model will be developed. After this, the degree centrality (DC) of proteins in the weighted PPI network will be calculated. Finally, all proteins in the weighted PPI network will be sorted in descending order according to their DC. Following, the method named the weighted degree centrality (WDC) will be described in detail.

1) Edge clustering coefficient: Radicchi et al. have proposed the edge-clustering coefficient (ECC) in analogy with the usual node-clustering coefficient [16]. ECC of a link between nodes i and j is defined as the ratio of the actual number of triangles $Z_{i,\,j}^{(3)}$ to which the link between i and j contributes and the number of possible triangles, determined by the minimum of the degrees k_i and k_j of the two nodes i and j:

$$ECC(i,\ j) = \frac{Z_{i,\,j}^{(3)}}{\min\left(k_i - 1, k_j - 1\right)} \tag{1}$$

2) Pearson correlation coefficient: The Pearson correlation coefficient (PCC) is a frequently used coefficient to express the degree of linear relationship between two sets of gene expression values. For two sequences of gene expressions such as $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n)$, the correlation coefficient is estimated by

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x_i})(y_i - \bar{y_i})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x_i})^2 \sum_{j=1}^{n}(y_j - \bar{y_j})^2}} \tag{2}$$

The value of the correlation coefficient is always between minus one and plus one.

3) weighted degree centrality: We will explain our methodology of adding weights to the interactions by exploiting the information of the gene expression profiles and the PPI network. The weight of an interaction is given by the metric described as:

$$W = ECC \times \lambda + PCC \times (1 - \lambda), \tag{3}$$

The *PCC* and *ECC* stand for the Pearson correlation coefficient and edge clustering coefficient of an interaction, respectively. The constant λ falls within the interval: [0, 1]. Following the principle that similar expression profiles are associated with function [15] and the definition of the edge clustering coefficient [16], the value of the above stated metric favors interactions whose corresponding genes have similar expression and thus enhances them in the overall network. We propose a new centrality measure, i.e., the weighted degree centrality (WDC), depending on the weighted yeast PPI network. WDC is calculated according to the weighted interactions between a protein and its direct neighbor. Specifically, the weighted degree centrality *WDC* of a node i is the sum of weighted values of the edges connecting node i and its neighbors. It can be expressed as

$$WDC(i) = \sum_{j}^{N_i} W_{i,\,j} \tag{4}$$

where N_i is the set of neighbors of node i and $W_{i,\,j}$ refers to the weighted value of edge between node i and its neighbor j. All proteins in the PPI network are ranked in descending order of WDC.

B. Results

1) Data sources: Protein interaction data: The yeast PPI data is downloaded from DIP (http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=7/), updated on Oct. 10, 2010.

Gene expression data: The yeast gene expression data comes from [17]. It is available in the form of a $9,335 \times 36$ matrix, includes expression profiles of 9,335 probes under 36 different time points, updated on Apr 14, 2011.

Integrated essential gene set: A list of essential proteins of $S.cerevisiae$ are downloaded from MIPS [18], SGD [19], DEG [20] and SGDP [21], which contain 1,285 essential proteins altogether.

2) ROC curves of various prediction methods: In our experiment, we use ROC AUC statistic as a classification measure so that the performance of different prediction technologies can be checked. Figure 1 shows the comparative results. As shown in Figure 1, the area under the curve (AUC) for WDC is 0.691 and NC's AUC is 0.689. For DC and PeC, we yield an AUC of 0.671 and 0.633, respectively. The resulting ROC of the four methods demonstrates that WDC and NC are more suitable for discrimination between essential proteins and other proteins in yeast compared with DC and PeC. The difference of WDC and PeC is how to weight the PPI network. When λ is 0.5, the weight of WDC is (ECC+PCC)/2. On the other side, the weight of PeC is ECC$\times$PCC. The weight method of ECC is used as a reference in the following discussion. A computed ECC value of a pair of interacting proteins may be bigger than their real ECC value because of the false interactions (false positives) relating the two proteins. On the contrary, the missed interactions (false negatives) may lead to that a

computed ECC value is smaller than the real ECC value. The error produced by the false positives and negatives can be improved in terms of the PCC between the two interacting proteins. When PCC is more than 0, it implies that two genes coding a pair of interacting proteins may be coexpressed. In this case, it is reasonable for WDC weight method to increase the weighted value between two interacting proteins. However, it is unreasonable for PeC weight method to decrease the weighted value of a pair of proteins. This may explain why the AUC of PeC is the smallest in Figure 1.

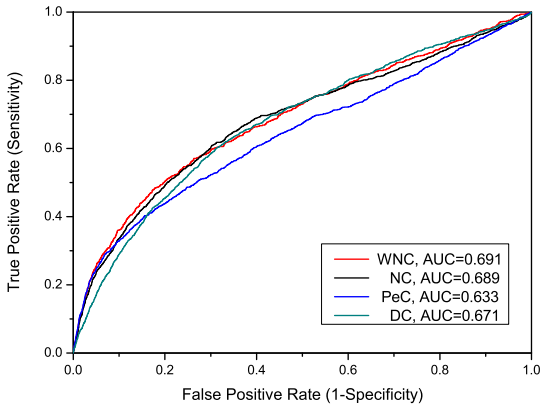

Figure 1. **ROC curves of the prediction performances. Figure 1 gives the ROC curve of four prediction approaches: WDC, NC, PeC and DC.**

3) Assessment of the proportion of essential proteins in proteins sorted in terms of various prediction methods: We select the top 1%, top 5%, etc. of the 5,093 proteins ranked in descending order according to their value of WDC, and determine how many of these are essential in the yeast PPI network [22]. Table 1 shows the number of essential proteins identified from various prediction technologies up to 25% of proteins in the PPI network. For instance, as shown in the first column of Table 1, when the top 1% of ranked proteins are selected, WDC and other methods (DC, BC, CC, EC, SC, IC, NC and PeC) identify 36, 22, 24, 24, 24, 24, 24, 32 and 41 essential proteins, respectively. As mentioned above, 5,093 proteins in the yeast PPI network include 1,167 essential proteins. The top 25% of proteins, i.e., 1,274 ranked proteins thus are already more than the total number of essential proteins in the PPI network. From Table 1, it can be found that the results from WDC are better than those from NC -not to mention DC, BC, CC, SC, EC and IC. Specifically, compared with other methods excepting PeC, WDC predicts consistently more essential proteins in each top percentage. PeC finds the most essential proteins in top 1% or 5%. If some biologists prefer to pick proteins from the highest 1% or 5%, PeC is a good choice. However, the number of essential proteins discovered by PeC is less than that of essential proteins predicted by WDC in other intervals. It means that if we want to select some proteins to test their lethality, so as to select some of them as targets for new drugs, it will be more appropriate to adopt WDC than to use other methods. Incidentally, it hints that a certain gene may be suppressed in order to allow another one to be expressed with PCC$<$0. In the interval of PCC, it is reasonable for WDC ($\lambda = 0.5$) and PeC to decrease the weighted value of every pair of proteins. At the same time, it can be observed that the decreasing rate of PeC weight values becomes faster than that of WDC weight values, which may be the reason that PeC identifies the most essential proteins in top 1% or 5% of ranked proteins.

Table I
NUMBER OF ESSENTIAL PROTEINS IN THE RANKED PROTEINS.

	1%	5%	10%	15%	20%	25%
DC	22	101	207	320	413	502
BC	24	95	182	271	361	433
CC	24	104	193	284	364	448
SC	24	96	195	279	377	467
EC	24	96	195	279	377	467
IC	24	102	210	316	406	504
NC	32	159	282	373	465	545
PeC	**41**	**171**	293	386	467	536
WDC	**36**	**165**	**307**	**402**	**489**	**566**

4) Effect of the parameter λ: WDC employs the parameter λ to integrate PCC and ECC. We now investigate how the variation of λ affects the performance of our WDC method. Table 2 shows the number of essential proteins predicted by WDC in each top percentage under different values of λ. For instance, WDC detects 36, 164, 293, 398, 491 and 552 essential proteins under different percentages when $\lambda = 0.3$. From Table 2, it can be found that the performance of WDC is better when $\lambda \in [0.3,\ 0.5]$. Its performance in particular is optimal when $\lambda = 0.5$. So the parameter λ are set 0.5 in our experiment.

Table II
NUMBER OF ESSENTIAL PROTEINS PREDICTED BY WDC IN EACH TOP % UNDER DIFFERENT VALUES OF λ.

λ	1%	5%	10%	15%	20%	25%
0.0	36	168	282	375	443	505
0.1	36	167	286	387	455	528
0.2	36	168	291	388	477	542
0.3	36	164	293	398	491	552
0.4	36	164	303	405	488	561
0.5	**36**	**165**	**307**	**402**	**489**	**566**
0.6	33	166	304	397	487	561
0.7	32	164	297	392	487	558
0.8	34	163	290	391	479	558
0.9	31	164	287	380	478	546
1.0	32	159	282	373	465	545

III. Conclusion

In this research, the gene expression profiles are successfully incorporated into the PPI network. Based on PCC and ECC, a new essential protein prediction method named WDC is developed. The experimental results indicate that WDC is constantly superior to other prediction technologies with regard to the proportion of selected proteins that are essential proteins. Our research suggests that it is important to predict essential proteins by integrating multiple data sources. Therefore, future work should focus on how to merge different data sources and developing new centrality measures for increased power to discriminate between essential and non-essential proteins.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grant Nos.61073036, 61003124, the Ph.D. Programs Foundation of Ministry of Education of China No.20090162120073, the Freedom Explore Program of Central South University No.201012200124, High-tech Program of China Hunan Provincial Science and Technology Department Nos.2010GK3049, 2011GK3138, Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province No.2010212, Hunan Provincial Department of education Science Foundation under Grant No.11C0281, the U.S. National Science Foundation under Grants CCF-0514750, CCF-0646102 and CNS-0831634.

References

[1] N. Judson and J.J. Mekalanos, "TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes," *Nat Biotechnol*, Vol. 18, No. 7, pp. 740-5, Jul. 2000.

[2] R.R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, "Identifying Hubs in Protein Interaction Networks," *PLoS ONE*, Vol. 4, No. 4, pp. e5344, 2009.

[3] L.C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, Vol. 40, No. 1, pp. 35-41, 1977.

[4] S. Wuchty,and P.F. Stadler, "Centers of Complex Networks," *J. Theoretical Biology*, Vol. 223, No. 1, pp. 45-53, 2003.

[5] E. Estrada and J.A. Rodríguez-Velázquez, "Subgraph Centrality in Complex Networks," *Physical Rev. E*, Vol.71, No. 5, pp. 056103, 2005.

[6] P.Bonacich, "Power and Centrality: A Family of Measures," *Am. J. Sociology*, Vol. 92, No. 5, pp. 1170-1182, 1987.

[7] K. Stevenson and M. Zelen, "Rethinking Centrality: Methods and Examples," *Social Networks*, Vol. 11, No. 1, pp. 1-37, 1989.

[8] J. Wang, Li M., H. Wang and Y. Pan, "A New Method for Identifying Essential Proteins Based on Edge Clustering Coefficient," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 1, 2012.

[9] E. Sprinzak, S. Sattath and H. Margalit, "How Reliable are Experimental Protein-Protein Interaction Data?" *Journal of Molecular Biology*, Vol. 327, pp. 919-923, 2003.

[10] J. Chen and B. Yuan, "Detecting Functional Modules in the Yeast Protein-Protein Interaction Network," *Bioinformatics*, Vol. 22, pp. 2283-2290, 2006.

[11] M. Li, J. Wang, H. Wang amd Y. Pan, "Essential Proteins Discovery from Weighted Protein Interaction Networks," *Proceedings of the 7th international conference on Bioinformatics research and applications*, Vol. 6053, pp. 89-100, 2010.

[12] M. Li,H. Zhang, J. Wang and Y. Pan, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data," *BMC Systems Biology*, Vol. 6, pp. 15, 2012.

[13] X. He and J. Zhang, "Why Do Hubs Tend to Be Essential in Protein Networks?" *PLoS Genet*, Vol. 2, pp. e88, 2006.

[14] G.T. Hart, I. Lee, and E.M. Marcotte, "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality," *BMC Bioinformatics*, Vol. 8, No. 236, 2007.

[15] C.J. Wolfe, I.S. Kohane and A.J. Butte, "Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks," *BMC Bioinformatics*, Vol. 6, No. 79, 2005.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, "Defining and identifying communities in networks," *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 101, No. 9, pp. 2658-2663, 2004.

[17] B.P. Tu, A. Kudlicki, M. Rowicka and S.L. McKnight, "Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes," *Science*, Vol. 310, pp. 1152-1158, 2005.

[18] H.W. Mewes, et al., "MIPS: analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Res,*, Vol. 34, No. Database issue, pp. 169-172, 2006.

[19] J.M. Cherry, et al., "SGD: Saccharomyces Genome Database," *Nucleic Acids Res.*, Vol. 26, No. 1, pp. 73-79, 1998.

[20] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Res.*, Vol. 37, No. Database issue, pp. 455-458, 2009.

[21] http://www-sequence.stanford.edu/group/yeast_deletion_project, 2011.

[22] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *PROTEOMICS*, Vol. 6, No. 1, pp. 35-40, 2006.