International Journal of u- and e- Service, Science and Technology
Vol.9, No. 8 (2016), pp.343-350
http://dx.doi.org/10.14257/ijunesst.2016.9.8.29

Identification of Essential Protein based on Functional Modules and Weighted Protein-Protein Interaction Networks

Yi Pan, Sai Hu and Bihai Zhao*

Department of Mathematics and Computer Science, Changsha University, Changsha, Hunan 410022, China
bihaizhao@163.com

Abstract

Identification of essential proteins plays a significant role in understanding minimal requirements for the cellular survival and development. Experimental methods for the identification of essential proteins are always costly, time-consuming, and laborious. High throughput technologies have resulted in a large number of protein-protein interaction data, which provided a stepping stone for predicting essential proteins using computational approaches. There have been a series of computational approaches proposed for predicting essential proteins based on network topologies. However, the network topology-based centrality measures are very sensitive to noise of network. In this paper, we propose a naive essential protein discovery method, named PMN, based on the integration of weighted interactome network and functional modules. The performance of PMN is validated based on the PPI network of Saccharomyces cerevisiae. Experimental results show that PMN significantly outperforms the classical centrality measures. The results also uncover relationship between the modularity and essentiality of proteins.

Keywords: *Interactome network; Essential protein; Functional modules*

1. Introduction

Proteins constitute the structure of cells and tissues of all essential ingredients that are the most important activities of life material base. However, different proteins on the importance of life activities are not the same. Usually those who have been excluded for protein complexes, resulting from loss of function and causing the organism can not survive are called essential proteins [1]. Essential protein is not only necessary for the organism to survive and reproduce, but also plays an important role in life activities. Therefore, identification of essential proteins from the system level helps to understand the internal organization of life activities and processes. Meanwhile, a large number of studies have shown that an essential protein (gene) is often the disease gene [2]. It can be seen that the identification of essential provides valuable information both for proteins biology, medicine and other related disciplines, especially with important applications in disease diagnosis and drug design.

In biology, the essential proteins and disease genes are mainly identified by biomedical experiments [3]. The biological methods of essential proteins identification are clear and effective. However, the cost of these methods is quite high, the efficiency is very low, and suitable species are limited. In recent years, with the rapid development of bioinformatics, prediction of essential proteins based on computer science and mathematical theory become the new direction of development. Especially the development of the yeast two-hybrid [4], tandem affinity purification [5], mass spectrometry [6] high-throughput proteomic techniques supply a large number of protein-protein interaction data. These large amounts of protein interaction data makes it possible to predict essential proteins through a computer method. Recently more attention has been

ISSN: 2005-4246 IJUNESST

paid to computational methods based on network topological characteristics. Many researchers have explored the correlations between network topological features and protein essentiality. Proteins in the network highly connecting with other proteins are more likely to be essential than those selected by chance. This is called the *centrality-lethality* rule [7].

Computational methods could be seen as useful preprocessing techniques which could help experimental methods to quickly find essential proteins. Many centrality measures have been proposed to capture the correlation between network topological properties and protein essentiality. Local network features based centrality measures include Degree Centrality (DC) [8], Betweenness Centrality (BC) [9], Closeness Centrality (CC) [10], Subgraph Centrality(SC) [11], Information Centrality(IC) [12] and Normalized a-Centrality (NC) [13]. Chua, etc. [14] proposed measurement methods combined with the existing central measures to identify essential proteins (including, edge clustering coefficient, NFC and ND). Del *etc.*, [15] analyzed 18 different reconstructed metabolic networks of 16 different center measures and found that among 16 centers measure, any combination two of them can improve the prediction performance, but the combination of three or over three did not improve the prediction.

Though a great progress has been made on the computational methods for the identification of essential proteins based on network topologies, the identification of essential proteins based on topological property is still very challenging. One of the most important factors is that a significant proportion of PPI networks obtained from high-throughput biological experiments have been found to contain false positives and false negative. For false positives, a general approach is to evaluate the interactions by using different weighting methods. More recently, there is a new trend that improves the precision of essential protein discovery method by integration of network topology and other biological information. Hart *et al.*, [16] showed that the essentiality is a property of the protein complexes, and the experimental data shows that a large number of essential proteins tend to concentrate in certain protein complexes. Inspired by the researches and discoveries mentioned above, we propose a new method for predicting essential proteins in the yeast interactome network by integrating functional modules and weighted PPI networks. The performance of PMN was tested on the well studied species of *Saccharomyces cerevisiae*. Compared to other previous centrality measures: DC [8], BC [9], CC [10], SC [11], IC [12], NC [13], PMN achieves the highest precision for the identification of essential proteins. The experimental results show that the integration of network topology and functional modules increased the predictability of essential proteins in comparison with those centrality measures only based on network topological features.

2. Method

In this paper, a new essential proteins discovery method, PMN is proposed based on the integration of weighted PPI networks and functional modules. The basic ideas behind PMN are as follows: (1) A highly connected protein is more likely to be essential than a low connected one; (2) Essential proteins tend to form densely connected clusters; (3) Essential Proteins in the same cluster have a more chance to be co-expressed; (4) Essentiality is tied not to the protein or gene itself, but to the molecular module to which that protein belongs (5) Essential proteins

have a higher frequency to be in different functional modules than non-essential proteins. In PMN, a protein's essentiality is determined by the frequencies and weighted degrees of the protein in functional modules.

To describe PMN simply and clearly, we provide the following definitions and descriptions. A PPI network can be modeled as a simple graph $G = (V, E)$, in which a vertex in vertex set V represents a protein and an edge in edge set E represents an interaction between two distinct proteins. As we all known, PPI networks obtained from high-throughput biological experiments have been found to contain false positives. To reduce the negative impact of noise on the prediction of essential proteins, we construct a weighted PPI network using FS-Weight [17] to calculate the score of protein pairs.

Definition 1 FS-Weight Given a pair of proteins u and v, FS-Weight of edge (u,v) is defined as follows:

$$FS-Weight(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v| + \lambda_u} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v| + \lambda_v} \quad (1)$$

where N_u and N_v are sets consisting of all neighbors of u and v, respectively, λ_u and λ_v are used to penalize proteins with very few neighbors, and they are defined as follows:

$$\lambda_u = \max\left\{0, \frac{\sum_{x \in v}|N_x|}{|V|} - |N_u|\right\}$$
$$\lambda_v = \max\left\{0, \frac{\sum_{x \in v}|N_x|}{|V|} - |N_v|\right\} \quad (2)$$

Based on the definition, if the degree of a vertex u is below the average degree, then it is adjusted to the average degree.

For a protein pair of a weighted PPI network, the higher the weight is, the more likely the two proteins interact with each other. The intuition behind the weighting method is simple: if the weight of an interaction reflects its reliability, then the weighted interactions should better represent the actual interaction network than the initial binary ones.

Definition 2 Weighted degree (WD) Given a weighted PPI network $G= (V, E, W)$ and a vertex u, $u \in V$. $V = \{v_1, v_2, \ldots, v_n\}$, $E = \{e_1, e_2, \ldots, e_m\}$, $W = \{w(e_1), w(e_2), \ldots, w(e_m)\}$, $w(e_i)$ is the weight of an edge e_i. $WD(u, G)$ denotes the weighted degree of u within G and is defined as:

$$WD(u,G) = \sum_{i=1}^{n} w(u,v_i), (u,v_i) \in E \quad (3)$$

It has been proved that there exist a number of functional modules, which play a key role in carrying out biological functionality, and the essentiality tends to be a product of a functional module rather than an individual protein. Inspired by the researches and discoveries, we propose a new method to predict essential proteins, named PMN. The PMN calculates scores of all proteins by integrating weighted PPI netwok and benchmark functional modules. The score of a protein is used to judge whether the protein is essential. For a protein v, its score $S(v)$ is defined as

the sum of weighted degree (*WD*) of *v* in all benchmark functional modules. Not all the proteins are contained in benchmark functional modules. So, if a vertex *v* is not contained in any functional module, *v* is identified as a non-essential protein and *S*(*v*) is assigned to zero. Let $FM = \{fm_1, fm_2, \ldots, fm_m\}$ is a set of benchmark functional modules predicted by experimental methods. Generally, *S* (*v*) can be calculated by the follow formula:

$$S(v) = \begin{cases} \sum_{i=1}^{m} WD(v, fm_i) & , \quad \exists fm_i, v \in fm_i \\ 0 & , \quad \forall fm_i, v \notin fm_i \end{cases} (i \in [1, m]) \tag{4}$$

3. Results and Discussion

3.1. Experimental Data

To evaluate the performance of the proposed method, the PPI network of *Saccharomyces cerevisiae* was used, as it has been well characterized by knockout experiments and widely used in the evaluation of methods for essential proteins discovery. The test data used in this paper are as following:

The PPI data of *Saccharomyces cerevisiae* was downloaded from DIP database [18]. There are 24,743 interactions among 5093 proteins in total after the self-interactions and the repeated interactions were filtered. The PPI network consists of 21 components. The largest component consists of 5052 proteins.

Essential proteins of *Saccharomyces cerevisiae* were collected from several databases, such as MIPS [19], SGD [20], DEG [21] and SGDP [22]. Out of all the 5093 proteins in the PPI network, 1167 proteins are essential among which 1165 proteins are in the largest component of the PPI network.

A benchmark functional modules set is adopted from CYC2008 [23], which consist of 408 functional modules. There are 1627 distinct proteins in CYC2008.

3.2. Compare PMN with other Methods

To validate the performance of the proposed new method PMN, we carry out a comparison between it and six other previously proposed centrality measures: Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Subgraph Centrality(SC), Information Centrality (IC) and Normalized a-Centrality (NC). Proteins are ranked according to their values calculated by each method. A certain number of top proteins are selected as candidates for essential proteins. Then we determine how many of them are true essential proteins. The number of essential proteins detected by PMN and six other centrality measures (DC, BC, CC, SC, IC and NC) from the yeast protein-protein interaction network is shown in Figure 1.

From Figure 1 we can see that PMN performs significantly better than all the six previous aforementioned methods for predicting essential proteins from the yeast protein interaction network. Especially, the improvement of PMN over the classic centrality measures (DC, BC, IC, SC) is more than 50%.

A more general comparison between the proposed new method PMN and the six previously proposed methods (DC, BC, CC, SC, IC and NC) is tested by using a jackknife methodology [24]. The comparison results are shown in Figure 2. In Figure 2, the X-axis from left to right represents the proteins in PPI networks

ranked in the descending order according to their ranking scores computed by corresponding methods, while the Y-axis is the cumulative count of essential proteins with respect to ranked proteins moving left to right. The areas under the curve for PMN and the six other methods are used to compare their prediction performance. As shown in Figure 2, it is clear that the sorted curve of PMN appears to be much better than that of the six previously proposed centrality measures: DC, BC, CC, SC, IC and NC. The comparison results indicate that the integration of weighted protein-protein interaction and functional modules can help improve the predicted precision of identifying essential proteins.

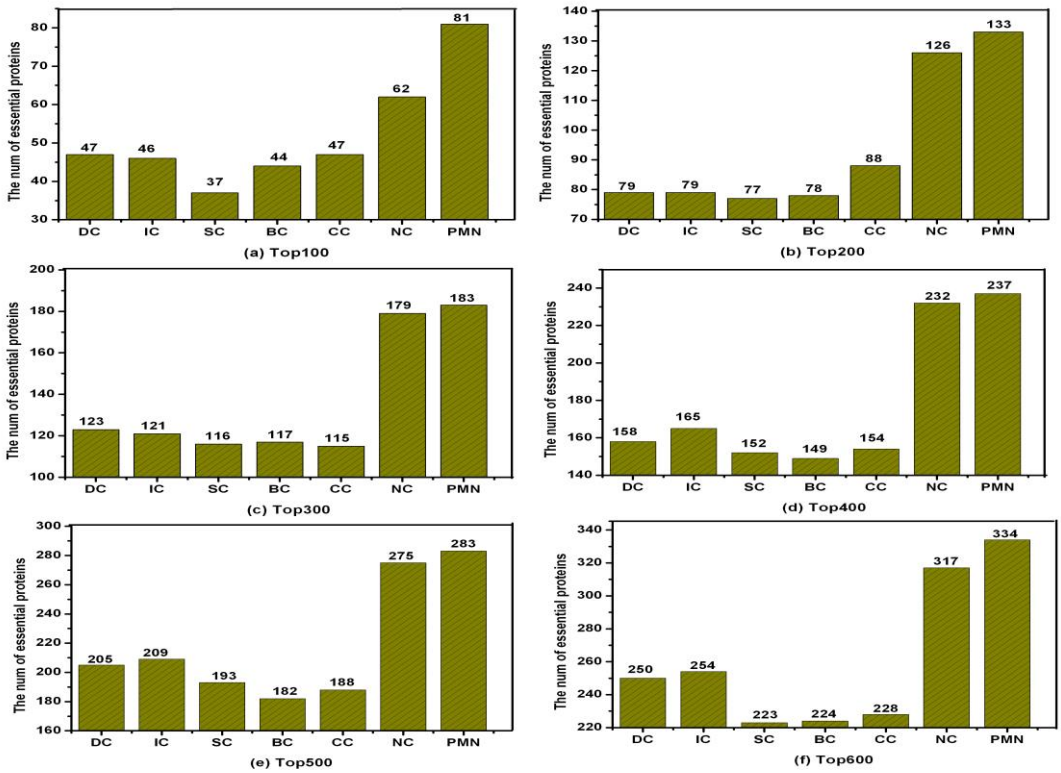


Figure 1. Comparison of the Number of Essential Proteins Predicted by PMN and Six other Methods

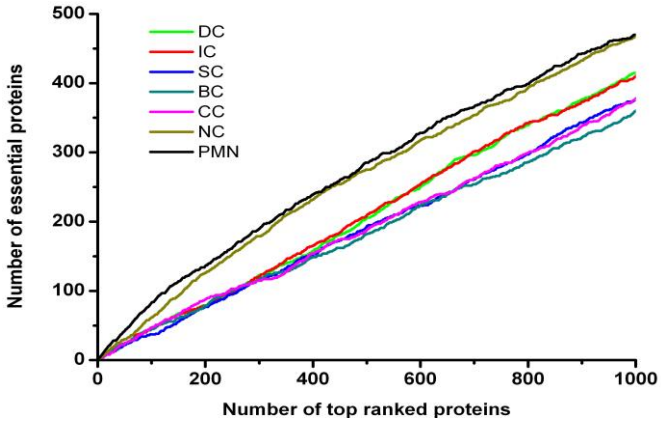


Figure 2. Jackknife Curves of PMN and Six other Methods

3.3. Analysis of the Differences between PMN and other Methods

To further analyze why and how PMN performs well on the identification of essential proteins, we study the relationship and difference between it and six other methods by predicting a small fraction of proteins. For each method, the top 100 proteins are selected.

Firstly, we compare PMN with DC, BC, CC, SC, IC and NC by investigating how many proteins are both predicted by PMN and by anyone of the other methods. The number of overlaps between PMN and one of the other methods is shown in Table 1. In Table 1, $|PMN \cap Mi|$ denotes the number of common proteins detected by PMN and by a method Mi. $\{M_i - PMN\}$ (or $\{PMN - M_i\}$) represents the set of proteins detected by M_i (or PMN), but not by PMN (or M_i). $|M_i - PMN|$ is the number of proteins in set $\{M_i - PMN\}$.

From Table 1, we can see that the common proteins identified by PMN and DC, IC, SC, BC and CC are all less than 10%, and that common proteins both predicted by PMN and NC is less 30%. Such a small overlap between the predicted proteins of PMN and DC, IC, SC, BC, CC and NC shows that PMN is a special centrality measure which is much different from others.

The fourth column in Table 1 refers to the number of non-essential proteins among different proteins identified by Mi but not by PMN. According to the further investigation about these non-essential proteins predicted by other methods, we have found that more than three-quarter of these non-essential proteins detected by other methods (DC, IC, SC, BC, CC and NC) have very low scores of PMN (less than 0.2).

Table 1. Overlap and Different Proteins Predicted by PMN and other Methods Ranked in top 100 Proteins

M_i	$\|PMN \cap M_i\|$	$\|M_i - PMN\|$	Non-essential proteins in $\{M_i - PMN\}$	Percentage of non-essential proteins in $\{M_i - PMN\}$ with low PMN
DC	3	97	51	78.43%
IC	3	97	54	83.33%
SC	2	98	63	82.54%
BC	3	97	56	80.36%
CC	6	94	53	79.25%
NC	27	73	33	75.76%

Secondly, we evaluate the different proteins identified by PMN and those by other methods.

Table 2. Comparison of the Percentage of Essential Proteins out of all the Different Proteins between PMN and other Methods

M_i	$\|M_i - PMN\|$	Percentage of essential proteins in PMN	Percentage of essential proteins in M_i
DC	97	82.47%	47.42%
IC	97	82.47%	44.33%
SC	98	82.65%	35.71%
BC	97	82.47%	42.27%
CC	94	81.91%	43.62%
NC	73	83.56%	54.79%

Table 2 shows how many essential proteins are predicted out of all the different proteins identify by PMN and those identified by DC, BC, CC, SC, IC and NC. As expected, the results shown in Table 2 illustrates that the percentage of essential proteins

identified by PMN is consistently higher than that explored by six other methods for the different proteins between them. Take SC as an example, out of all the top 100 proteins, there are 98 different proteins detected by PMN. About 82.65% of these proteins are essential, while there are only 35.71% of different proteins detected by SC but not by PMN are essential proteins.

4. Conclusions

Essential proteins play a key role in the life activities of cells. In this work we propose a new method, named PMN, for predicting essential proteins based on functional modules and weighted PPI networks. PMN is applied to the PPI network of *Saccharomyces cerevisiae*. The experimental results show that the predicted precision of PMN is clearly higher than those of the six other topology-based centrality measures: DC, IC, SC, BC, CC and NC. Although PMN performs well on the discovery of essential proteins, there should be still a space to improve the prediction precision. besides the functional modules data, some other protein related data, such as gene expression data, should be also integrated into PPI networks for identifying essential proteins. The integration of multiple protein related data may contribute a good deal to the identification of essential proteins with further research efforts.

Acknowledgments

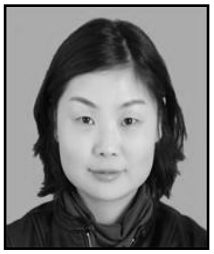
This work is supported National Natural Science Foundation of China No.11501054, Natural Science Foundation of Hunan Province No. 13JJ4106, No.14JJ3138, Foundation of Education Bureau of Hunan Province (Grant14A016),Science and Technology Plan Project of Hunan Province No. 2010FJ3044, No. 2015GK3072.

References

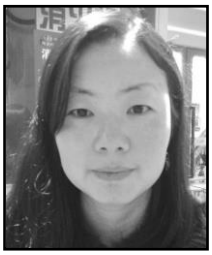
[1] X. Zhang, J. Xu and W. Xiao, "A new method for the discovery of essential proteins", PloS one, vol. 8, no. 3, **(2013)**, pp. e58763.
[2] P. Yang, X. Li and H. N. Chua, "Ensemble Positive Unlabeled Learning for Disease Gene Identification", PloS one, vol. 9, no. 5, **(2014)**,pp. e97079.
[3] P. V. Maillard, C. Ciaudo and A. Marchais, "Antiviral RNA interference in mammalian cells", Science, vol. 342, no. 6155, **(2013)**, pp. 235-238.
[4] T. Ito, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", PNAS, vol. 98, no. 8, **(2001)**, pp. 4569-4574.
[5] G. Rigaut, "A generic protein purification method for protein complex characterization and proteome exploration", Nature Biotechnology, vol. 17, **(1999)**, pp. 1030-1032.
[6] Y. Ho, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry", Nature, vol. 405, **(2002)**, pp. 180-183.
[7] C. J. Ryan, N. J. Krogan and P. Cunningham, "All or nothing: protein complexes flip essentiality between distantly related eukaryotes", Genome biology and evolution, vol. 5, no. 6, pp. 1049-1059.
[8] H. Jeong and S. P. Mason, "Lethality and centrality in protein networks", Nature, vol. 411, no. 6833, **(2001)**, pp. 41-42.
[9] M. Joy, "High-betweenness proteins in the yeast protein interaction network", Journal of Biomed Biotechnol, vol. 2, **(2005)**, pp. 96-103.
[10] S. Wuchty and P. F. Stadler, "Centers of complex networks", Journal of Theor Biol, vol. 223, **(2003)**, pp. 45-53.
[11] E. Estrada and J. A. Rodríuez-Veláquez, "Subgraph centrality in complex networks", Physical Review E, vol. 71, no. 5, **(2005)**, pp. 1-9.
[12] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples", Social Networks, vol. 11, **(1989)**, pp. 1-37.
[13] R. Gosh and K. Lerman, "A Parameterized Centrality Metrics for Network Analysis", Phys Rev, **(2011)**, E 83:066118.
[14] H. N. Chua, K. L. Tew and X. L. Li, "A unified scoring scheme for detecting essential proteins in protein interaction networks", ICTAI'08. 20th IEEE International Conference on. IEEE, vol. 2, **(2008)**, pp. 66-73.
[15] G. Del Rio, K. Dirk and C. Gerardo, "How to identify essential genes from molecular networks?", BMC Systems Biology, vol. 3, no. 1, **(2009)**, pp. 102.

[16] 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. 1, **(2007)**, pp. 236.
[17] H. N. Chua, W. K. Sung and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions", Bioinformatics, vol. 22, no. 13, **(2006)**, pp. 1623-1630.
[18] B. Zhao, J. Wang and M. Li, "Detecting Protein Complexes Based on Uncertain Graph Model", IEEE/ACM Transactions on Computational Biology and Bioinformatics, TCBB.2013.2297915, **(2014)**.
[19] H. W. Mewes, D. Frishman and K. F. X. Mayer, "MIPS: analysis and annotation of proteins from whole genomes in 2005", Nucleic acids research, vol. 34, suppl. 1, **(2006)**, pp. D169-D172.
[20] J. M. Cherry, "SGD: Saccharomyces Genome Database", Nucleic acids research, vol. 26, no. 9, **(1998)**.
[21] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes", Nucleic acids research, vol. 37, **(2009)**, pp. D455-D458.
[22] Saccharomyces Genome Deletion Project, http://www-sequence.stanford.edu/group/.
[23] S. Pu, J. Wong and B. Turner, "Up-to-date catalogues of yeast protein complexes", Nucleic acids research, vol. 37, no. 3, **(2009)**, pp. 825-831.
[24] M. Li, H. Zhang and J. Wang, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data", BMC systems biology, vol. 6, no. 1, **(2012)**, pp. 15.

Authors



Pan Yi, she received her M.S. degree (2000) and Ph.D degree (2005) in Computer Software and Theory from Huazhong University. Now she is an associate professor in Department of Mathematics and Computer Science of Changsha University. Her current main research interests include data mining, bioinformatics.



Hu Sai, she received her M.S. degree (2003) from Hunan University. Now she is an associate professor in Department of Mathematics and Computer Science of Changsha University. Her current main research interests include mathematical statistics.

Zhao Bi-hai, he received his Ph.D degree (2014) from Center South University. Now he is an associate professor in Department of Mathematics and Computer Science of Changsha University. His current main research interests covers bioinformatics and data mining. (Corresponding author of the paper).

