2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization

The Topological Features of Nonessential-Nonhub Proteins in the Protein-Protein Interaction Network

Dong Yun-yuan[1], Yang Jun[2], Liu Qi-jun[3], and Wang Zheng-hua[1]
[1] College of Computer, National University of Defense Technology, Changsha, China
[2] No.91 Element of 92941 Troops of the Chinese People's Liberation Army, Huludao, China
[3] College of Science, National University of Defense Technology, Changsha, China
happydongyy@gmail.com, yjnudt@163.com, ivanliuqj@gmail.com, zhhwang188@sina.com

Abstract—**One of the important problems in system biology is to discover the relationship between topological properties and functional features of proteins in protein-protein interaction (PPI) networks. There are many essential-nonhub proteins and lots of nonessential-nonhub proteins in the PPI network. Both essential-nonhub proteins and nonessential-nonhub proteins are low-connectivity proteins, but they have different lethality. In order to explain why nonessential-nonhub proteins are not essential, we compare them with essential-nonhub proteins from topological view. The comparison results show that there are statistical differences between nonessential-nonhub proteins and essential-nonhub proteins in centrality measures and clustering coefficient.**

Keywords- protein-protein interaction network; nonessential-nonhub protein; essential-nonhub protein; centrality measure; clustering coefficient

I. INTRODUCTION

With the completion of human genome project and the finish of model organism sequencing, the field of proteomics stands on the threshold of significant advances. Crucial to furthering these investigations is a comprehensive understanding of the structures and functions of the proteins.

As we know, hub proteins are highly connected proteins. Essential proteins, which are detected by single gene knock-out[1], RNA interference[2] and conditional knockouts[3], are responsible for the viability of an organism.

Jeong *et al*. [4] observe that hub proteins are more likely to be essential than proteins selected by chance, which is the so called centrality-lethality rule. This rule has also been investigated in other species[5,6] and demonstrated by many other studies[7,8,9]. Also a number of researches [6,7,9,10] have confirmed that a protein's lethality is correlated with topological centrality in the PPI networks. However, in PPI networks, there are many low-connectivity essential proteins; we call them 'essential-nonhub proteins'. And there are also lots of low-connectivity nonessential proteins; we name them 'nonessential-nonhub proteins'. Both essential-nonhub proteins and nonessential-nonhub proteins are low-connectivity proteins, but why they have different lethality?

In this work, we explore the relationship between nonessential-nonhub proteins and essential-nonhub proteins to explain why nonessential-nonhub proteins are not lethal from topological point of view. The comparison results show that there are statistical differences between nonessential-nonhub proteins and essential-nonhub proteins in Betweenness Centrality (BC)[11], Eigenvector Centrality (EC)[12], Information Centrality (IC)[13], Subgraph Centrality (SC)[14], Bottle Neck (BN)[8,15], Edge Percolation Component (EPC)[16], Density of Maximum Neighborhood Component(DMNC)[17], L-index (LI)[18] and clustering coefficient.

II. MATERIALS AND METHODS

A. Experimental Data

1) Protein-protein interaction datasets

The PPI dataset of yeast, which is named DIP_core, is derived from the DIP database[19]. There are 2164 proteins and 4303 interactions in total after self-interactions and redundancy interactions are removed.

2) Essential proteins

Essential proteins of saccharomyces cerevisiaeare obtained from the Saccharomyces Genome Deletion Project[20]. It contains 1156 essential proteins. In dataset DIP_core, out of all the 2164 proteins, there are 405 essential-nonhub proteins and 1160 nonessential-nonhub proteins.

Note that, in this study, the nonhub proteins are those proteins whose degrees are less than five[21].

B. Method

The PPI network is regarded as an undirected graph *G*, in which proteins are represented as nodes and interactions are represented as edges. We assign *N* as the total number of nodes in the network and the adjacency matrix of the network is $A=a_{ij}$. If there is an edge between node *i* and node *j*, $a_{ij}=1$; otherwise $a_{ij}=0$. The commonly used centrality measures are defined as follows.

1) Centrality Measures

Betweenness Centrality (BC) is the fraction of shortest paths going through node *i*[11]. $BC(i)=\sum_s\sum_t \frac{\sigma_{st}(i)}{\sigma_{st}}, s \neq t \neq i$. Here σ_{st} is the number of shortest paths from node *s* to node *t*, and $\sigma_{st}(i)$ is the number of shortest paths passing through node *i* from node s to node *t*.

Closeness Centrality (CC) is the sum of graph-theoretic distances from all other proteins in the PPI network[22].

CC(i)=$\frac{1}{\sum_j dist(i,j)}$, here $dist(i,j)$ is the number of links in the shortest path from node i to node j.

Eigenvector Centrality (EC) is defined as the ith component of the principal eigenvector of the adjacency matrix A[12]. The eigenvector equation is $\lambda e = Ae$, where λ is an eigenvalue and e is its corresponding eigenvector. EC(i)= $e_1(i)$, where e_1 corresponds to the largest eigenvalue of A.

Information Centrality (IC) is defined as: IC(i)= $\left[\frac{1}{N}\sum_j \frac{1}{I_{ij}}\right]^{-1}$, where I_{ij}= $(c_{ii}+c_{jj}-c_{ij})^{-1}$. Matrix $C=[D-A+J]$-1, where D is a diagonal matrix of the degree of each protein in G, and J is a matrix with all its elements equal to one[13]. For computational purposes, I_{ii} is defined as infinite, so $\frac{1}{I_{ii}} = 0$.

Subgraph Centrality (SC) is the total number of closed walks in which i takes part and gives more weight to closed walks of short lengths[14]. SC(i)= $\sum_{l=0}^{\infty} \frac{\mu_l(i)}{l!}$, where $\mu_l(i)$ denotes the number of closed walks of length l which starts and ends at node i.

Bottle Neck (BN) is defined as follows[8,15]. Node i is taking as the root of a tree T_i, in which all shortest paths are starting from i. The weight of a node j in the tree is the number of shortest paths starting from i passing through j. A node j is called a bottle-neck node in T_i if the weight of j is no less than $n/4$, where n is the number of nodes in T_i.

Edge Percolation Component (EPC) is defined as follows [16]. In a network G, assign a removing probability p to every edge. Let G' be a realization of the random edge removing from G. If nodes i and j are connected in G', set δ_{ij} be 1, otherwise set δ_{ij} be 0. The percolated connectivity of i and j, c_{ij}, is defined to be the average of δ_{ij} over realizations. EPC(i) is defined to be the sum of c_{ij} over nodes j.

Density of Maximum Neighborhood Component(DMNC) is defined as follows[17]. For a node i, N_i is the number of its neighbors and E_i is the number of edges in the maximum connected component in N_i. DMNC(i)= E_i/N_i^{ε} , $1 \le \varepsilon \le 2$. ε is set to be 1.7[17].

Lobby Index (LI) of a node i is the largest integer n such that i has at least n neighbors with a degree of at least n[18].

2) *Damage*

Damage is a quantitative criterion for the importance of a node in a network. It measures the consequences of the deletion of a node from the network [23,24]. Note that, we make a little change to the definition of damage. The PPI graph G has N nodes. After deletion of node i, there are N' nodes in graph G'. Damage is defines as $d=N-N'$.

3) *Clustering Coefficient*

Clustering coefficient measures tightness of a node with its direct neighbors: $C(i)=2k/n(n-1)$, where k is the number of direct connections among node i and its n neighbors.

4) *Evaluation Method*

Rank sum test is the most well-known non-parametric significance tests and it is also called non-parametric Mann-Whitney U test[25,26]. It is used for assessing whether two independent samples of observations have statistical difference. If the result of rank sum test is less than 0.05, the two samples have statistical difference.

III. RESULT AND DISCUSSION

A. *Differences in Centrality Measures*

At first, we compare the centrality of both essential-nonhub proteins and nonessential-nonhub proteins in nine centrality measures. The rank sum test results are shown in Table 1. Except for CC, essential-nonhub proteins and nonessential-nonhub proteins have statistical differences in other eight centrality measures.

TABLE I. RANK SUM TEST RESULTS OF NINE CENTRALITY MEASURES

Centrality Measure	***p*-value**
BC	1.7877×10-6
CC	0.0975
EC	5.9397×10-4
IC	8.9238×10-4
SC	5.8236×10-6
BN	0.0075
EPC	2.9396×10-4
DMNC	2.2617×10-6
LI	1.7647×10-8

Because the number of essential-nonhub proteins and nonessential-nonhub proteins differ greatly (there are 405 essential-nonhub proteins and 1160 nonessential-nonhub proteins in dataset DIP_core), we choose 405 nonessential-nonhub proteins randomly to demonstrate their differences. As shown in Figure 1 (a)~(h), there are distinct differences in the values of centrality between essential-nonhub proteins and nonessential-nonhub proteins. Essential-nonhub proteins have higher centrality values than nonessential-nonhub proteins, which mean essential-nonhub proteins are more important than nonessential-nonhub proteins in the PPI network. That is one reason that nonessential-nonhub proteins are not lethal.

B. *Differences in Damage*

Damage is the number of proteins that are disconnected from the network when a protein is deleted, and it accounts for the influence of a given protein in the PPI network. The rank sum test result of damage between essential-nonhub proteins and nonessential-nonhub proteins is 0.1875, which means there is no statistical difference between essential-nonhub proteins and nonessential-nonhub proteins and they take same effect in maintain the PPI network's robustness.

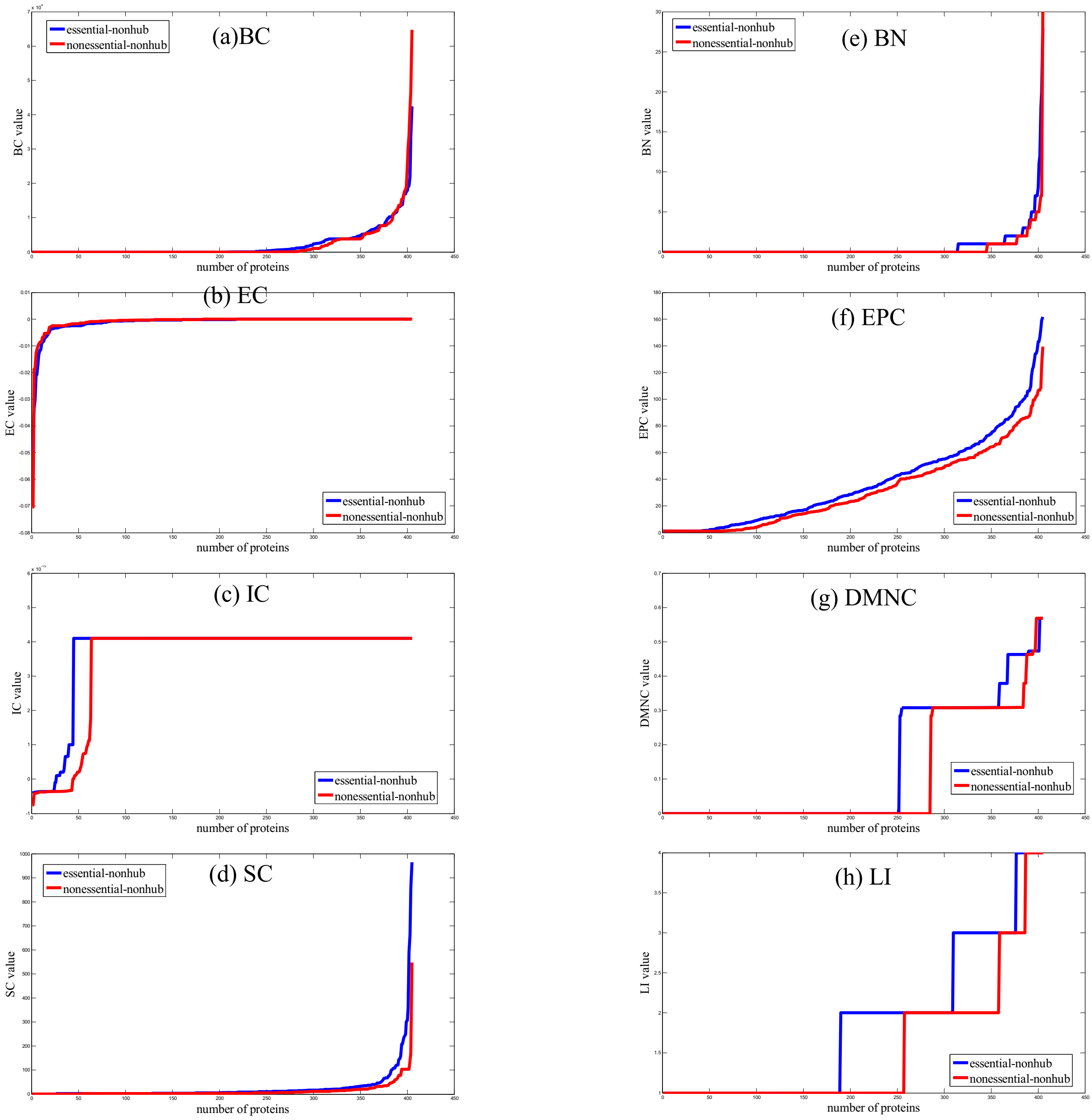

Figure 1. Distribution of centrality measures values in essential-nonhub proteins and nonessential-nonhub proteins

C. *Differences in Clustering Coefficient*

Clustering coefficient is the measurement of how close a node to its neighbors. We calculated clustering coefficient of both essential-nonhub proteins and nonessential-nonhub proteins. The rank sum test result of clustering coefficient between essential-nonhub proteins and nonessential-nonhub proteins is 6.8060×10^{-6}, which means essential-nonhub proteins and nonessential-nonhub proteins have statistical difference in clustering coefficient. They have different closeness with their neighbors.

As shown in Figure 2, we choose 405 nonessential-nonhub proteins randomly. Nonessential-nonhub proteins have higher clustering coefficient, and they are more close to their neighbors than essential-nonhub proteins.

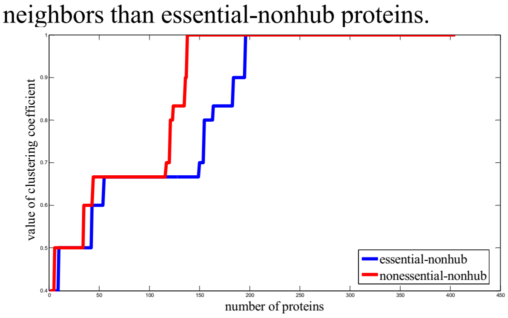

Figure 2. Distribution of clustering coefficient in essential-nonhub proteins and nonessential-nonhub proteins

IV. Conclusion

In order to explain why nonessential-nonhub proteins are not essential in the PPI network, we explore the relationship between nonessential-nonhub proteins and essential-nonhub proteins from topological point of view. We compare them in nine centrality measures, damage and clustering coefficient. The results show that there are statistical differences between nonessential-nonhub proteins and essential-nonhub proteins in centrality measures and clustering coefficient. Nonessential-nonhub proteins have lower centrality values and higher clustering coefficient than essential-nonhub proteins.

References

[1] Giaever G, Chu AM, Ni L, Connelly C, Riles L, e. al, Functional profiling of the Saccharomyces cerevisiae genome, Nature 418 (2002) 387-391.

[2] L.M. Cullen, G.M. Arndt, Genome-wide screening for gene function using RNAi in mammalian cells, Immunology and Cell Biology 83 (2005) 217-223.

[3] T. Roemer, B. Jiang, J. Davison, et al., Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery, Molecular Microbiology 50 (2003) 167-181.

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

[5] Yu H, Greenbaum D, Xin Lu H, Zhu X, G. M, Genomic analysis of essentiality within protein networks, Trends Genet 20 (2004) 227-231.

[6] Hahn MW, K. AD, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Mol Biol Evol 22 (2005) 803-806.

[7] Batada NN, Hurst LD, T. M, Evolutionary and physiological importance of hub proteins, PLoS Comput Biol 2 (2006) e88.

[8] Yu H, Kim PM, Sprecher E, Trifonov V, G. M, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, PLoS Comput Biol 3 (2007) e59.

[9] R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, A. Raval, Identifying Hubs in Protein Interaction Networks, Plos One 4 (2009) 1-10.

[10] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, Proteomics 6 (2006) 35-40.

[11] L.C. Freeman, A set of measures of centrality based upon betweenness, Sociometry 40 (1977) 35-41.

[12] P.F. Bonacich, Power and centrality: A family of measures, American Journal of Sociology 92 (1987) 1170-1182.

[13] K. Stevenson, M. Zelen, Rethinking centrality: Methods and examples, Social Networks 11 (1989) 1-37.

[14] E. Estrada, J.A. Rodríguez-Velázquez, Subgraph Centrality in Complex Networks, Phys Rev E Stat Nonlin Soft Matter Phys. 71 (2005) 056103.

[15] N. Przulj, D.A. Wigle, I. Jurisica, Functional topology in a network of protein interactions, Bioinformatics 20 (2004) 340-348.

[16] C.-S. Chin, P.S. Manoj, Global snapshot of a protein interaction network——a percolation based approach, Bioinformatics 19 (2003) 2413-2419.

[17] Chung-Yen Lin, Chia-Hao Chin, Hsin-Hung Wu, Shu-Hwa Chen, Chin-Wen Ho, Ming-Tat Ko, Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology, Nucleic Acids Research 36 (2008) W438–W443.

[18] Korna A., Schubert A., Telcs A., Lobby index in networks, Physica A 388 (2009) 2221-2226.

[19] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, E. D, DIP: the database of interacting proteins., Nucleic Acids Research 28 (2000) 289-291.

[20] Saccharomyces Genome Deletion Project. , 2011.

[21] Kar Leong Tew, X.-L. Li, Functional centrality detecting lethality of proteins in protein interaction networks, Genome Inform 19 (2007) 166-177.

[22] S. Wuchty, P. Stadler, Centers of complex networks, Journal of Theoretical Biology 223 (2003) 45-53.

[23] Jean Schmith, Ney Lemke, Jose´ C.M. Mombach, Patrı´cia Benelli, Cla´ udia K. Barcellos, G.B. Bedin, Damage, connectivity and essentiality in protein-protein interaction networks, Physica A 349 (2005) 675-684.

[24] Ney Lemke, Fabiana Herédia, Cláudia K. Barcellos, Adriana N. dos Reis, J.C.M. Mombach, Essentiality and damage in metabolic networks, Bioinformatics 20 (2004) 115-119.

[25] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (1945) 80-83.

[26] H. B. Mann, D.R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, Annals of Mathematical Statistics 18 (1947) 50-60.