Identification of Essential Proteins by Using Complexes and Interaction Network*

Min Li[1], Yu Lu[1], Zhibei Niu[1], Fang-Xiang Wu[3], and Yi Pan[1,2]

[1] School of Information Science and Engineering,
Central South University, Changsha 410083, P.R. China
[2] Department of Computer Science,
Georgia State University, Atlanta, GA 30302-4110, USA
[3] Department of Mechanical Engineering and Division of Biomedical Engineering
University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada
limin@mail.csu.edu.cn

Abstract. Essential proteins are indispensable in maintaining the cellular life. Identification of essential proteins can provide basis for drug target design, disease treatment as well as synthetic biology minimal genome. However, it is still time-consuming and expensive to identify essential protein based on experimental approaches. With the development of high-throughput experimental techniques in the post-genome era, a large number of PPI data and gene expression data can be obtained, which provide an unprecedented opportunity to study essential proteins at the network level. So far, many network topological methods have been proposed to identify the essential proteins. In this paper, we propose a new method, United complex Centrality(UC), to identify essential proteins by integrating protein complexes information and topological features of PPI network. By analysis of the relationship between protein complexes and essential proteins, we find that proteins appeared in multiple complexes are more inclined to be essential compared to these only appeared in a single complex. The experiment results show that protein complex information can help identify the essential proteins more accurate. Our method UC is obviously better than traditional centrality methods(DC, IC, EC, SC, BC, CC, NC) for identifying essential proteins. In addition, even compared with Harmonic Centricity which also used protein complexes information, it still has a great advantage.

Keywords: essential proteins, PPI network, protein complexes, traditional centrality methods.

1 Introduction

Essential proteins are the proteins that play an irreplaceable role in the cellular life, they are closely related to survival and reproduction of the organism[1-3].

* This work is supported in part by the National Natural Science Foundation of China under Grant No.61370024, No.61232001, and No.61379108, the Program for New Century Excellent Talents in University (NCET-12-0547).

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 255–265, 2014.

Removal of any one of essential proteins leads to a fatal defect on an organism[4,5].Compared with non-essential proteins, essential proteins tend to be more conservative in biological evolution[6-8], so that they are not easy to change. Meanwhile, it is proved that essential proteins are often disease genes in the human body cell[9,10]. Thus, the identification of essential proteins in lower organisms is of great importance to finding disease genes and drug targets. At the same time, it has important applications in disease diagnosis and drug design, etc.

Recent years, many researchers concerned more about the connections between topological properties in PPI networks and essential proteins[11-18]. Highly connected center nodes(hubs) in PPI networks are often the essential proteins, that is so-called centrality-lethality rule. After that, researchers successively proposed a series of centrality methods. For examples:Degree Centrality (DC)[19-21], Betweenness Centrality (BC)[22,23], Closeness Centrality (CC)[24], Subgraph Centrality (SC)[25], Eigenvector Centrality(EC)[26], Information Centrality (IC)[27], Neighborhood Centrality (NC)[28], Local Average Connectivity-based Method (LAC)[29]. These centrality methods are used to predict essential proteins based on topological features of PPI networks. Experimental analyses had shown that these centrality methods are much more effective than choosing essential proteins randomly. Specially, NC and LAC are better than six traditional centrality methods (DC, BC, CC, SC, EC, IC).

However, merely using PPI to predict essential protein is far less, the PPI data produced by current high throughput technology exists some false interactions(false positive)[33], these will effect the accuracy of essential protein predicting. Thus, researches hope to integrate multi-information to help identify the essential proteins more accurate[30-32,45]. Meanwhile, a large number of studies have shown that there exists close relationship between protein complex and essential proteins. Hart, etc.[34] pointed out that the essentiality is one of properties of protein complex. The essentiality of proteins are not merely decided by a single protein, often by functional protein complex. Zotenko, etc[14] proposed the concept of Essential Complex Biological Modules, a group of highly connected function modules which have the same or similar biological functions. They also consider that the reason why hub nodes are inclined to be essential proteins is there exists a lot in essential complex modules, so they joined in more biological functions.On this basis, Ren, etal.[45] introduced the information of protein complex and proposed a new centrality measuring method of proteins named Harmonic Centricity(HC).

In this paper, we propose a new method to identify essential proteins based on information of protein complexes and topology features of PPI networks. We first verify that proteins in complexes are more inclined to be essential proteins and sort the proteins of PPI network into two parts: proteins in complexes and proteins not in complexes. Secondly, after the classification, we analyze the proteins in complexes and find that proteins appeared in multiple complexes are more likely to be essential. So we classify the proteins of complexes into two parts: proteins appeared in a single complex and proteins appeared in multiple complexes. Finally, we combine the classified proteins and the topology characteristics of PPI network to identify the essential proteins. The results show that this method can improve the prediction precision of essential proteins compared with other centrality measures. Even compared to Harmonic Centricity which also used protein complexes information, it still has a great advantage.

2 Materials and Methods

Materials

PPI Network: PPI network data of *S.cerevisiae* are downloaded from DIP dataset[37], we call it as YDIP. It contains 4950 proteins and 21788 interactions.

Protein Complex: Protein complex datasets are obtained from MIPS_216[36], MIPS_408[36], krogan[38], Gavin[39], Ho[40], the map of integrated dataset in YDIP is named YC_union. YC_union contains 1208 complexes and 2572 proteins. We only wiped out the same complexes when integrate YC_union.

Essential Proteins: The essential proteins of *S.cerevisiae* are obtained from the following databases:MIPS(Mammalian Protein-Protein Interaction Database) [36], SGD(Saccharomyces Genome Database)[41], DEG(Database of Essential Genes)[42] and SGDP(Saccharomyces Genome Deletion Project)[43]. A protein in the yeast PPI network is considered as an essential protein if it can be matched at least in one yeast essential database. Out of all the 4950 proteins in YDIP, 1151 proteins are essential, 3799 are non-essential if it can't be marked in our essential datasets.

Methods

Protein complex is the basis of executing biological procedure, which generates all kinds of molecular mechanisms to executing a large number of biological functions. Complexes consist of specific proteins which usually have constant structures and functions[35]. In order to investigate the connections between protein complexes and essential proteins, we download the PPI network and complex information of *S.cerevisiae*, and list the number of proteins and essential proteins in yeast PPI network and several complex datasets in Table 1. Among these, PPI network is named YDIP. Different from usually that only concern the relationship between the single known complex and essential proteins, we integrate five known complex datasets (MIPS_216, MIPS_408, krogan, Gavin, Ho) named YC_union. As shown in Table 1, only 23.3% of the 4950 proteins in YDIP are essential proteins. While in the single protein complex datasets, the proportion of essential proteins are 42.6%, 38.4%, 66.00%, 43.40%, 34.80%, respectively. And out of the 2572 proteins in YC_union, 35.2% proteins are essential proteins.

In order to investigate that whether there exists the relationship between protein complexes and essential proteins or not, we divide all the proteins in YDIP into two part: proteins in YC_union and proteins not in YC_union. We also count the number and the proportion of their essential proteins. The results are

Table 1. Numbers of proteins and essential proteins in PPI network and protein complexes

	dataset Name	Number of proteins	Number of essential proteins	Ratio of essential proteins
PPI network	YDIP	4950	1151	23.30%
Protein complex	MIPS_216	1184	504	42.60%
	MIPS_408	1835	711	38.70%
	Krogan	247	163	66.00%
	Gavin	1287	558	43.40%
	Ho	1228	427	34.80%
Union complex	YC_union	2572	906	35.20%

YC_union is a protein complex dataset which mapped in YDIP by integrating five known protein complex datasets: MIPS_216, MIPS_408, krogan, Gavin, Ho.

Table 2. The proportion of essential proteins in YC_union and not in YC_union

	proteins in YC_union	proteins not in YC_union
Number of proteins	2572	2378
Number of essential proteins	906	245
Ratio of essential protein	35.2%	10.3%

shown as Table 2. From Table 2 we can see that among the 4950 proteins in YDIP, the proportion of essential proteins in YC_union is 35.2% and the proportion of essential proteins not in YC_union is only 10.3%. Obviously, proteins in protein complexes are more likely to be essential.

It has been shown that many complexes exist a lot of overlapping part in PPI networks. That is to say, some proteins may belong to multiple protein complexes[36]. Obviously, if we wiped out these overlapping proteins, the protein complex will lost its function of this part. Thus, we divide the proteins in YC_union into two parts: proteins only appeared in one complex (overlap=1) and proteins appeared in multiple complexes(overlap>1). Here, given a protein u, its overlap denotes the number of complexes which protein u appeared in. We also count the number and the proportion of their essential proteins. As shown in Table 3, the proportion of essential proteins that appeared in multiple complexes is 43.0%, and the proportion of essential proteins that only appeared in one complex is 21.0%. Thus, essential proteins are more likely to appear in overlapping part of complexes.

A PPI network is described as an undirected graph $G = (V, E)$, which the nodes represent proteins, the edge represents an interaction between proteins. For an edge $E(i, j)$ connecting node i and node j, we pay attention to how many

Table 3. The proportion of essential proteins that only appeared in one complex and appeared in multiple complexes

	Proteins only appeared in one complex	Proteins appeared in multiple complexes
Number of proteins	910	1662
Number of essential proteins	191	715
Ratio of essential protein	21.0%	43.0%

other nodes that adjoin both i and j[28]. The edge clustering coefficient of $E(i,j)$ can be defined by the following formula:

$$ECC_{ij} = \frac{z_{i,j}}{min(k_i - 1, k_j - 1)} \tag{1}$$

where, $z_{i,j}$ is the number of triangles that include the edge actually in the network, k_i and k_j denotes the degree of node i and j, respectively, $min(k_i - 1, k_j - 1)$ is the number of triangles in which the edge $E(i,j)$ may possibly participate at most.

Then we combine the complex overlapping information and ECC, propose a new method to predict essential proteins, named UC(United complex Centrality). United complex centrality of protein i, UC(i), is defined as:

$$UC(i) = \sum_{j=1}^{n} \left(\frac{o(j) + 1}{O + 1} \times ECC_{ij} \right) \tag{2}$$

where $o(j)$ is the overlap number of protein j which interacts with protein i in YC_union, O is the biggest number of complexes which protein appeared in, ECC_{ij} is the edge clustering coefficient between protein i and protein j.

In the equation(2), if the number of complexes which protein appeared in is higher, this edge is more like to be important. Considering that protein j which interacts with protein i is not in YC_union, O and $o(j)$ should add 1 when we count the UC value of protein i.

3 Experiments and Results

The Analysis of Essential Proteins in Complexes

In the method section, we count and analyze the proportion of essential proteins in PPI network and known protein complexes, as well as the proportion of essential proteins in YC_union which is a union complex mapped from YDIP. In order to further analyze the essentiality of proteins in complexes, we divide the proteins in YC_union which mapped from YDIP into 10 groups (overlap=1, overlap=2...overlap=9 and overlap≥10). Then we calculate the proportion of proteins and essential proteins in YC_union for each group. From Fig 1, we can see the proportion of proteins in YC_union is decreased, but the proportion of essential proteins is increased. Especially, the proportion of essential proteins is 69.9% when overlap=8. When overlap=1, the proportion of proteins in YC_union reach the highest, but the proportion of essential proteins reach the lowest, only 21.0%. Apart from that, when overlap ≥10, the number of proteins is too low, so we put them into one group. But the proportion of essential proteins are still 58.9%. In conclusion, the proteins in the complexes tend to be essential proteins, especially the proteins whose overlapping times are higher.

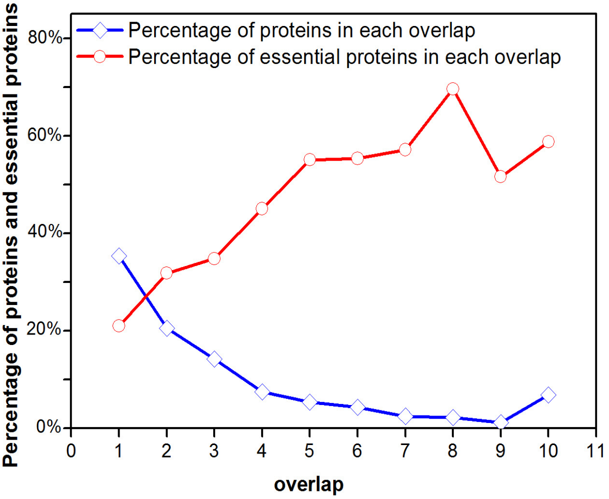

Fig. 1. Analysis about the proportion of proteins and essential proteins in YC_union for each overlap

For complex YC_union, we know it contains 2572 proteins and 906 essential proteins. In order to examine the relationship between YC_union and the essentiality of proteins is not random. We retain the size of every complex in YC_union and randomly choose proteins of YDIP to format new complex in 10 times. Then we calculate the proportion of essential proteins when proteins only appeared in one complex and proteins appeared in multiple complexes. From Fig 2, we can see that when proteins only appeared in one complex and proteins appeared in multiple complexes, the proportion of essential proteins in random complexes are all about 0.23. While the proportion is 0.43 in YC_union when proteins appeared in multiple complexes. This is higher than the case when proteins only appeared in one complex, also higher than randomly tests. Thus, there exists no random relationship between YC_union and the essentiality of proteins. Proteins are more likely to be essential proteins when they appeared in multiple complexes.

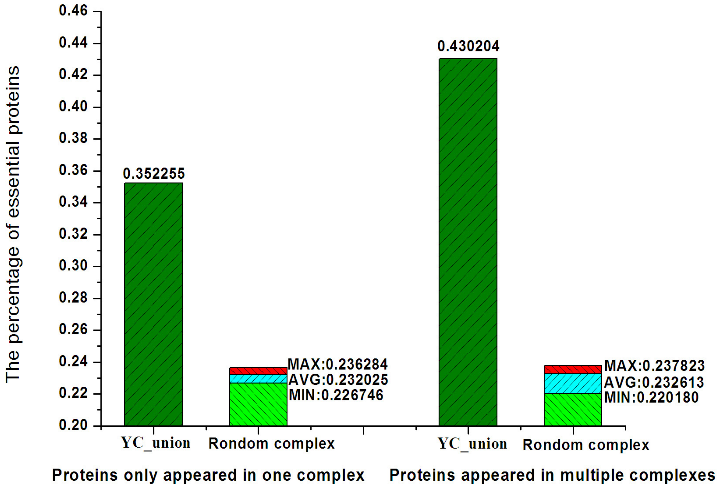

Fig. 2. The proportion of essential proteins when proteins only appeared in one complex and proteins appeared in multiple complexes. MAX means the maximum proportion of essential proteins in 10 times random complexes, AVG means the average proportion of essential proteins in 10 times random complexes, MIN means the minimum proportion of essential proteins in 10 times random complexes.

UC Compared with other Methods

In order to compare UC's performance with other methods in the prediction of essential proteins, we calculate the essentiality of every protein based on DC, IC, EC, SC, BC, CC, NC, HC and UC. As similar with previous experimental procedures[40], We do descending order to all the proteins according to the value and choose top 1%, 5%, 10%, 15%, 20%, 25% proteins as candidate essential proteins for each method. Then we calculate how many essential candidates are true essential proteins based on the above top percentage. As Fig 3 showed, UC performs better than other methods from top 1% to 25%. In top 1%(the number of proteins is 49), the true essential proteins number of every traditional centrality method is lower than 30, HC is also only 31. While the true essential proteins number of UC is 40, the identify ratio is reaching 81.6%. Especially, when we choose top 1% and 5% proteins as candidate essential proteins, the identifying performance of UC all increased 85% compared with EC and SC. With the increase of the number of the selected essential candidates, less improvement is obtained by UC compared with EC and SC. But even choose top 25% proteins as candidate essential proteins, the number of true essential proteins produced by UC still increased over 30% compared with EC and SC. Besides, comparing with the best result in above identifying measures from top 1% to top 25%, the result of UC still has great advantage in any top percentage.

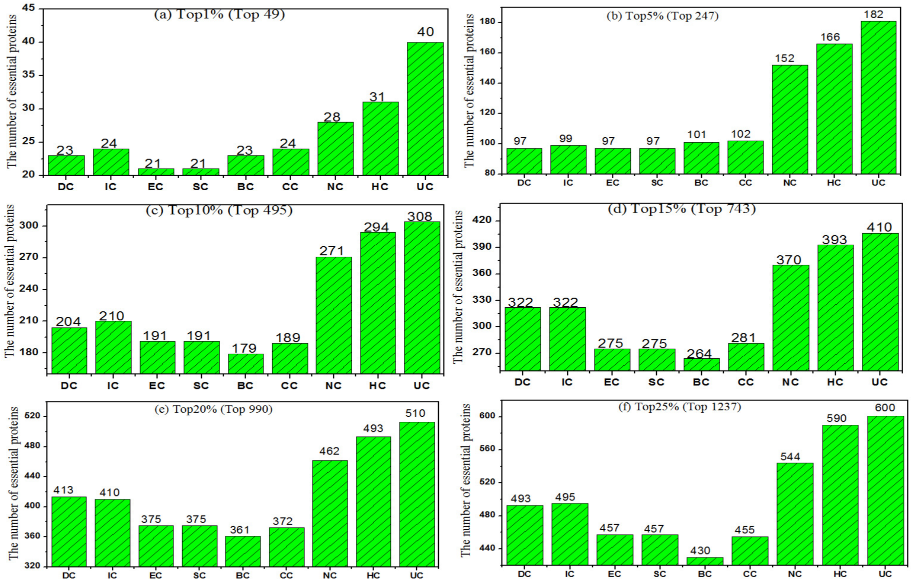

Fig. 3. The number of true essential proteins by UC, HC, DC, IC, EC, SC, BC, CC and NC when selecting the top 1%, top 5%, top 10%, top 15%, top 20%, top 25% proteins

4 Conclusion and Discussion

The identification of essential proteins based on PPI network is always a heat point in post-genome era, but the method usually concentrated on several properties of topology level, not that deeply in digging biological meaning and biological function. However, essential proteins are of great importance in biology functions. It is necessary to combine biology information in the procedure of essential proteins identifying. Previous research has shown that there are close relationship between protein complexes and essential proteins[4,14]. In this paper, we propose a new method to identify essential proteins based on information of protein complexes and topology features of PPI networks. The results show that proteins in overlapping part of complexes are more inclined to be essential proteins. Apart from that, this method improved a lot in performance of predicting essential proteins compared with traditional centrality methods(DC, BC, CC, SC, EC, IC, and NC).

In this paper, we only combine information of protein complexes and network topology characteristics to identify essential proteins. However, there are still some biological characteristics such as protein function information, domain information, gene expression information, etc closely related to essential proteins. In future, we will further analyze the essentiality of proteins by using these biological information.

References

1. Pál, C., Papp, B., Hurst, L.D.: Genomic function (communication arising): rate of evolution and gene dispensability. Nature 421(6922), 496–497 (2003)
2. Zhang, J., He, X.: Significant impact of protein dispensability on the instantaneous rate of protein evolution. Molecular Biology and Evolution 22(4), 1147–1155 (2005)
3. Liao, B.Y., Scott, N.M., Zhang, J.: Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Molecular Biology and Evolution 23(11), 2072–2080 (2006)
4. Winzeler, E.A., Shoemaker, D.D., Astromoff, A., et al.: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285(5429), 901–906 (1999)
5. Kamath, R.S., Fraser, A.G., Dong, Y., et al.: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421(6920), 231–237 (2003)
6. Kondrashov, F.A., Ogurtsov, A.Y., Kondrashov, A.S.: Bioinformatical assay of human gene morbidity. Nucleic Acids Research 32(5), 1731–1737 (2004)
7. Furney, S.J., Albá, M.M., López-Bigas, N.: Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics 7(1), 165 (2006)
8. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., et al.: Evolutionary rate in the protein interaction network. Science 296(5568), 750–752 (2002)
9. Xu, J., Li, Y.: Discovering disease-genes by topological features in human protein - protein interaction network. Bioinformatics 22(22), 2800–2805 (2006)
10. Park, D., Park, J., Park, S.G., et al.: Analysis of human disease genes in the context of gene essentiality. Genomics 92(6), 414–418 (2008)
11. Jeong, H., Mason, S.P., Barabási, A.L., et al.: Lethality and centrality in protein networks. Nature 411(6833), 41–42 (2001)
12. Estrada, E.: Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics 6(1), 35–40 (2006)
13. He, X., Zhang, J.: Why do hubs tend to be essential in protein networks? PLoS Genetics 2(6), e88 (2006)
14. Zotenko, E., Mestre, J., O'Leary, D.P., et al.: Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Computational Biology 4(8), e1000140 (2008)
15. Chua, H.N., Tew, K.L., Li, X.L., et al.: A unified scoring scheme for detecting essential proteins in protein interaction networks. In: 20th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2008, vol. 2, pp. 66–73. IEEE (2008)
16. Batada, N.N., Hurst, L.D., Tyers, M.: Evolutionary and physiological importance of hub proteins. PLoS Computational Biology 2(7), e88 (2006)
17. Seo, C.H., Kim, J.R., Kim, M.S., et al.: Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks. Bioinformatics 25(15), 1898–1904 (2009)
18. Acencio, M.L., Lemke, N.: Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics 10(1), 290 (2009)
19. Vallabhajosyula, R.R., Chakravarti, D., Lutfeali, S., et al.: Identifying hubs in protein interaction networks. PLoS One 4(4), e5344 (2009)

20. Pang, K., Sheng, H., Ma, X.: Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network. Biochemical and Biophysical Research Communications 401(1), 112–116 (2010)
21. Ning, K., Ng, H.K., Srihari, S., et al.: Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. BMC Bioinformatics 11(1), 505 (2010)
22. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry, 35–41 (1977)
23. Joy, M.P., Brock, A., Ingber, D.E., et al.: High-betweenness proteins in the yeast protein interaction network. BioMed Research International 2005(2), 96–103 (2005)
24. Wuchty, S., Stadler, P.F.: Centers of complex networks. Journal of Theoretical Biology 223(1), 45–53 (2003)
25. Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. Physical Review E 71(5), 056103 (2005)
26. Bonacich, P.: Power and centrality: A family of measures. American Journal of Sociology, 1170–1182 (1987)
27. Stephenson, K., Zelen, M.: Rethinking centrality: Methods and examples. Social Networks 11(1), 1–37 (1989)
28. Wang, H., Li, M., Wang, J., Pan, Y.: A new method for identifying essential proteins based on edge clustering coefficient. In: Chen, J., Wang, J., Zelikovsky, A. (eds.) ISBRA 2011. LNCS (LNBI), vol. 6674, pp. 87–98. Springer, Heidelberg (2011)
29. Li, M., Wang, J., Chen, X., et al.: A local average connectivity-based method for identifying essential proteins from the network level. Computational Biology and Chemistry 35(3), 143–150 (2011)
30. Tang, X., Wang, J., Zhong, J., Pan, Y.: Predicting essential proteins based on weighted degree centrality. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2014)
31. Li, M., Zheng, R., Zhang, H., Wang, J., Pan, Y.: Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. Methods (2014)
32. Kim, W.: Prediction of essential proteins using topological properties in GO-pruned PPI network based on machine learning methods. Tsinghua Science and Technology 17(6), 645–658 (2012)
33. Sprinzak, E., Sattath, S., Margalit, H.: How reliable are experimental protein - protein interaction data? Journal of Molecular Biology 327(5), 919–923 (2003)
34. Hart, G.T., Lee, I., Marcotte, E.M.: A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8(1), 236 (2007)
35. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences 100(21), 12128–12128 (2003)
36. Mewes, H.W., Amid, C., Arnold, R., et al.: MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Research 32(suppl. 1), D41–D44 (2004)
37. Xenarios, I., Rice, D.W., Salwinski, L., et al.: DIP: the database of interacting proteins. Nucleic Acids Research 28(1), 289–291 (2000)
38. Gavin, A.C., Aloy, P., Grandi, P., et al.: Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084), 631–636 (2006)
39. Krogan, N.J., Cagney, G., Yu, H., et al.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084), 637–643 (2006)

40. Ho, Y., Gruhler, A., Heilbut, A., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868), 180–183 (2002)
41. Issel-Tarver, L., Christie, K.R., Dolinski, K., et al.: Saccharomyces Genome Database. Methods in Enzymology 350, 329 (2002)
42. Zhang, R., Lin, Y.: DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Research 37(suppl. 1), D455–D458 (2009)
43. `http://www-sequence.stanford.edu/group/yeast_deletion_project` (Saccharomyces Genome Deletion Project)
44. Li, M., Wang, J., Wang, H., Pan, Y.: Essential proteins discovery from weighted protein interaction networks. In: Borodovsky, M., Gogarten, J.P., Przytycka, T.M., Rajasekaran, S. (eds.) ISBRA 2010. LNCS, vol. 6053, pp. 89–100. Springer, Heidelberg (2010)
45. Ren, J., Wang, J., Li, M., Wang, H., Liu, B.: Prediction of essential proteins by integration of PPI network topology and protein complexes information. In: Chen, J., Wang, J., Zelikovsky, A. (eds.) ISBRA 2011. LNCS (LNAI), vol. 6674, pp. 12–24. Springer, Heidelberg (2011)