IDENTIFICATION OF ESSENTIAL PROTEIN BASED ON SUBCELLULAR LOCALIZATION, GENE EXPRESSION AND PROTEIN COMPLEX WITH ADAPTIVE PARAMETERS

Yuan Zhu[†] ✉ *Houwang Zhang*[†] *Hanying Zheng*[†]

[†] School of Automation, China University of Geosciences, Wuhan, China

ABSTRACT

***Index Terms*—** Essential proteins, Protein-protein interaction network, Adaptive Parameters.

1. INTRODUCTION

It is well known that proteins play a decisive role in cellular life activities, and are the main components of cellular physiological metabolic pathways [1]. Among them, essential proteins are of vitally important for cellular development and survival. With the studies that some essential proteins are related to human disease genes [2], identifying essential proteins is necessary to understand the molecular mechanisms of living cells. In addition, in the wake of high-throughput technologies, plenty of protein-protein interactions (PPIs) have been available [3], which facilitates the studies of essential proteins with computational methods.

In general, protein-protein interactions are constructed to protein interaction network (PIN), which is undirected. The topological characteristics of PIN is a hotspot for most of identification studying with computational methods, which are called network-based methods.

To data, a series of efficient network-based methods have been proposed to identify essential proteins from PIN. And degree centrality (DC), known for centrality-lethality rule, is the most famous and simplest network-based method to search essential protein [4]. Some studies show that large degree proteinsin PPI network tend to be essential ones [4, 5]. Other central measurements are also used to identify proteins, such as subgraph centrality (SC) [6], eigenvector centrality (EC) [7], betweenness centrality (BC) [8], closeness centrality (CC) [9], information centrality (IC) [10], and *etc*. Furthermore, a few of edge-aided methods [11, 12, 13] have proposed in essential proteins identification from PIN, like edge clustering coefficient [12]. Li *et al*.[13] improved its performance and put forward an improved edge clustering coefficient (NC). Lately, Zhu *et al*.[14] revised the defect of NEC and proposed a new centrality method called new node and edge clustering coefficient (NNEC) which has a better combination of dual characteristics of node and edge.

Network-based methods can predict essential proteins directly without acquiring additional information, and this makes them widely used in early stage. However, due to a lot of false positives and false negative data in PPI networks, these methods are limited in identification accuracies [15]. And these methods also ignore the intrinsic biological significance of essential proteins, which results in ignoring essential proteins with low connectivity [15].

Lately, some studies have proposed that integrating some biological information into network-based methods can decrease the effect of false positives of PPI data and extremely improve the prediction accuracy of essential proteins [16, 17]. Similarly, the biological information for identification can be divided into two class based on their function: edge-aided biological information and edge-aided biological information.

Edge-aided biological information includes gene expression data, gene functional data etc. Researchers always use these data to evaluate the relevance of two nodes in a edge, which can be further used for computation of edge coefficient. Li *et al*.[17] integrated gene expression data into NC and proposed PeC method. Zhu *et al*.[14] tried to integrate gene expression information, gene functional similarity, and protein-protein sequence similarity to identify essential proteins.

Node-aided biological information includes subcellular localization information, protein complexes information etc. These kinds of information is always took as the attribute of node to compute its essence. Li *et al*.[15] integrated subcellular localization data with orthology data for the identification of essential proteins. Luo *et al*. [18] predicted essential proteins in a protein interaction network based on statistical analyses of essential proteins and protein complexes.

In this study, we propose a novel method to predict essential proteins by integrating subcellular localization information, gene expression data and protein complexes data with PPI network with adaptive parameters, named SGC.

To verify the performance of SGC, we apply SGC on the publicly-available PPI data of *Saccharomyces cerevisiae* (Yeast). And we compare the performance of SGC with DC, BC, CC, EC NC, LAC, PeC, and INNEC. The experimental results show that our proposedSGC achieves better results than other state-of-the-art methods used in this paper.

2. EXPERIMENTAL DATA

PPI network dataset of *Saccharomyces cerevisiae* (Yeast) is downloaded from DIP database [19] updated to Oct.10, 2010. There are 5093 proteins and 24,743 interactions without self-interactions and repeated interactions in this dataset. We select Yeast because its PPI data and gene essentiality data are most complete and reliable among various species.

Essential protein dataset is selected from MIPS [20], SGD [21], DEG [22] and SGDP [23]. There are 1285 essential proteins in this dataset, out of which 1167 are in PPI network. We take the 1167 proteins as essential proteins while other 3926 proteins as non-essential ones.

Subcellular localization dataset of yeast is downloaded from knowledge channel of COMPARTMENTS database [24]. It integrates several source databases (UniProtKB [25], MGD [26], SGD [27], FlyBase [28] and WormBase [29]). As a result, it contains 5095 yeast proteins and 206,831 subcellular localization records. We select this database because both its data volume is large and it is updated in a timely manner. After preprocessing, there are still 3923 proteins in PPI network which have subcellular localization information.

Protein complex sets used in this study are from the work of Luo *et al*. [18], which incorporates four real protein complex sets (CM270, CM425, CYC408, and CYC428) into one comprehensive protein complex set.

Gene expression data used in this study is from GEO [30], which samples 36 time points during each Yeast successive metabolic cycle.

3. METHODS

Our novel method, SGC, predicts essential proteins based on the information integration of two kind of node-aided biological information (subcellular localization and protein complexes), one kind of edge-aided biological information (gene expression), and network topological feature. In the following subsections, we will introduce how to use these information and integrate them to calculate a protein's essentiality.

3.1. Network Topological Feature

Network centrality (NC) is an important measure and widely used for predicting essential proteins [13]. For a protein i in the PPI network $G = (V, E)$, its network centrality $NC(i)$ is defined as the sum of edge clustering coefficients (ECC) of all edges directly connected with protein i in the graph G.

$$\begin{aligned} NC(i) &= \sum_{j \in N_i} ECC(i,j), \\ &= \sum_{j \in N_i} \frac{T(i,j)}{min(d_i - 1, d_j - 1)}, \end{aligned} \tag{1}$$

where $N(i)$ is the set of all neighbors of protein i, $T(i,j)$ denotes the number of triangles built on edge $E(i,j)$, d_i and d_j are the degrees of nodes i and j, respectively. $min(d_i - 1, d_j - 1)$ represents the maximal possible number of triangles that might include the edge $E(i,j)$ potentially.

A protein owns higher ECC values with its neighbors will have a relatively higher NC value and will tend to be an essential protein. In order to match with other biological information, here we use the normalized NC value for each protein, denoted as NNC. For a protein i, its normalized NC value NNC(i) is defined as:

$$NNC(i) = \frac{NC(i)}{max(NC(i))} \tag{2}$$

where $max(NC(i))$ denotes the maximum NC value of all the proteins in the graph G.

3.2. Node-aided Biological Information

3.2.1. Subcellular Localization Score

Study has proposed that proteins must be localized at their appropriate subcellular compartments to perform their desired functions and that's why the subcellular localization information is beneficial to the identification of essential proteins [31]. Referring to [15], we calculate the numbers of subcellular location l where the top k% proteins appear and where the bottom k% proteins appear, respectively.

Considering that the multiple counting proteins may cause more false positives, we use k= 5 in this paper as [15] sets, ie., that the top/bottom 5 % proteins are selected. And we define f_l be the frequency of l where the top k% proteins appear and h_l denotes the frequency of l where the bottom k% proteins appear as [15] defines. Subcellular Localization Correlation Coefficient LCC(l) can be calculated as follow.

$$LCC(l) = \begin{cases} 1 - \frac{h_l}{f_l} & f_l < h_l \\ \frac{h_l}{f_l} - 1 & ohterwise \end{cases} \tag{3}$$

When $f_l < h_l$, it means that more proteins of low NNC values trend to appear in the location l and it is assumed that the relationship between the location l and protein's essence is negative. On the contrary, there is a positive correlation between the location l and protein's essence $f_l \geq h_l$. When $f_l = 0$, we set LCC(l) as the maximum of $1 - \frac{h_l}{f_l}$ with $f_l \neq 0$. And when $h_l = 0$, we set LCC(l) as the maximum of $\frac{h_l}{f_l} - 1$ with $h_l = 0$. A protein may appear in multiple subcellular locations. For a protein i, its subcellular localization score SL(i) is defined as the sum of LCC(l) of all the subcellular locations it appears. Besides, we also use the normalized SL value NSL(i) for each protein i by using the following formula:

$$NSL(i) = \frac{SL(i) + max\_SL}{max(SL(i) + max\_SL)} \tag{4}$$

Where Max_SL represents the maximum value of $SL(i)$ for all the proteins in G.

As for strengthening the identification of subcellular localization, we tie the NSL score to the network topological feature NNEC proposed in [14]. We name this combined feature as NNSL.

$$NNSL(i) = NSL(i) * NNEC(i) \tag{5}$$

3.2.2. *Protein Complexes Score*

Protein complexes are stable macromolecular assemblies that perform many diverse biochemical activities essential to cell homeostasis, growth, and proliferation [18]. Proteins within these complexes appear at the same time and in the same location, and more some proteins participate in more than one complex, and at the network level, protein complexes are usually substructures in PINs [18].

For a protein i, if it appears in more protein complexes, its criticality trends to be greater. So we compute protein complexes (PC) score as follow:

$$PC(i) = |Complex(i)| \tag{6}$$

Where $Complex(i)$ denotes the sets of complexes including protein i, and $|Complex(i)|$ is just the number of complexes including protein i.

3.3. Edge-aided Biological Information

Gene expression data is a classical biological information, and always be used in the identification of essential proteins. PeC is a method combining ECC and gene expression data, which can weaken the affect of false positives of PPI network. Hence, here we use PeC to extract useful information of gene expression data. For a protein i, its PeC score $PeC(i)$ can be computed as follow:

$$\begin{aligned} PeC(i) &= \sum_{j \in N_i} ECC(i,j) * PCC(i,j), \\ PCC(i,j) &= \frac{1}{s-1} \sum_{t=1}^{s} [\frac{g(X,i) - \bar{g}(X)}{\sigma(x)}] * [\frac{g(Y,i) - \bar{g}(Y)}{\sigma(Y)}], \end{aligned} \tag{7}$$

Where N_i denotes the set of all neighbors of node i, $ECC(i,j)$ is the edge coefficient between edge $E(i,j)$, $PCC(i,j)$ is the pearson's correlation coefficient of a pair of proteins (i and j).

3.4. The SGC Algorithm with Adaptive Parameters

The sorting score of our algorithm SGC is a linear combination of the three scores: normalized network centrality $NNC(i)$, normalized subcellular localization score $NNSL(i)$, normalized protein complexes score $PC(i)$, and normalized gene expression similarity score $PeC(i)$. For a protein i, its sorting score is calculated as follows:

$$\begin{aligned} SGC(i) = ((PC(i) + NNSL(i) * a + PeC(i) * (1-a)) \\ * b + NNC(i) * (1-b) \end{aligned} \tag{8}$$

Most of the methods choose to adjust the parameter to set an optimal combination parameters when combining various information. In our algorithm, a parameter adaptive method is proposed according to the different information. The ideas are as follows: the adaptive domain of each information in identifying essential proteins is different.

For example, we use PC and NNC two features to identify essential proteins of Yeast PPI dataset respectively. Through Fig. 1, we can see that PC can capture more essential proteins when deal with proteins with prime positions in the rankings. And for proteins with lower positions in the rankings, the effect of PC is not so helpful.

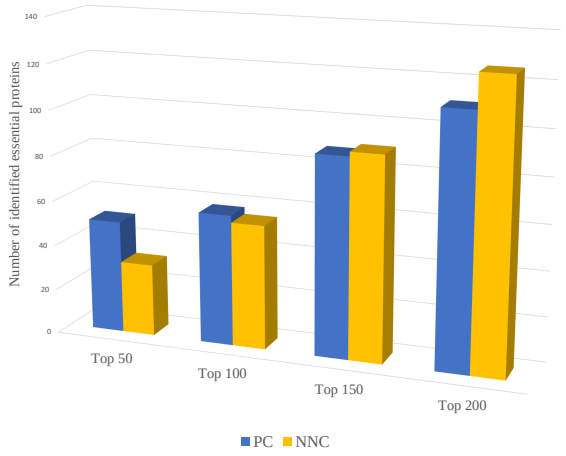

Fig. 1. The number of essential proteins predicted by PC and NNC.

In general, the effect of biological information is more reliable when dealing with proteins with prime positions in the rankings.

Therefore, the weight should be adjusted adaptively according to the number of essential proteins needed in the sorting process. The parameter adaptive model is as follow:

$$\begin{aligned} a &= \alpha_1 * input + \beta_1, \\ b &= \alpha_2 * input + \beta_2, \end{aligned} \tag{9}$$

where input is the expected number of essential proteins. In this paper, we take $\alpha_1 = 0.49$, $beta_1 = -0.0005$, $\alpha_2 = 1$, $\beta_2 = -0.0003$. That is, the weight of biological information is greater when calculating the top essential proteins, so

is the node-aided biological information. With the increase of input, the weight of network topological feature gradually increases, and the weight of edge-aided biological information also increases gradually.

The whole procedure of our proposed approach SGC is presented in Algorithm 1.

Algorithm 1 SGC for identification of essential proteins

Input: The PPI network $G = (V, E)$, subcelluar location data, protein complexes data, gene expression data, the desired number of essential proteins n.

Output: The top n identified essential proteins sorted by SGC in a descending order.

***Step* 1 :** Calculate the value of feature NNC for each protein by using Equation (1) and (2).

***Step* 2 :** Calculate the value of feature NNSL for each protein by using Equation (3), (4) and (5).

***Step* 3 :** Calculate the value of feature PC for each protein by using Equation (6).

***Step* 4 :** Calculate the value of feature PeC for each protein by using Equation (7) using Equation (9).

***Step* 5 :** Get the adaptive parameters based on the desired number of essential proteins n.

***Step* 6 :** Incorporate NNC, NNSL, PC and PeC using the adaptive parameters using Equation (8).

***Step* 7 :** Sort proteins by the value of SGC in a descending order and output top n of sorted proteins.

4. EXPERIMENTAL RESULTS

In the experiment, we compared our method SGC with several state-of-the-art methods: BC, CC, DC, EC, LAC, PageRank, and PeC on the PPI datasets of Yeast. All methods compared in this paper adopt their default parameters. All the experiments are run on a personal computer with Windows 10 OS, Intel Core i7 2.3GHz CPU, and 16GB memory.

Similar to most of validation methods for the identification of essential proteins, we also ranked all proteins by using each essential protein identification method in a descending order. And then we selected a certain number of top ranked proteins as the essential protein candidates. Here we select the top 100, top 200, top 300, top 400, top 500, and top 600 proteins as essential candidates of Yeast PPI network.

Firstly, the precision-recall (PR) curve which is a common methodology for evaluating the performance of essential proteins identification methods is used in this paper. The comparison of our method with the other methods for predicting essential proteins on the yeast PPI network by using the PR curve is shown in Fig. 2. From Fig. 2 we can see that the PR curve of SGC obtains the better result compared to the PR curves of other methods. And our method can get the largest

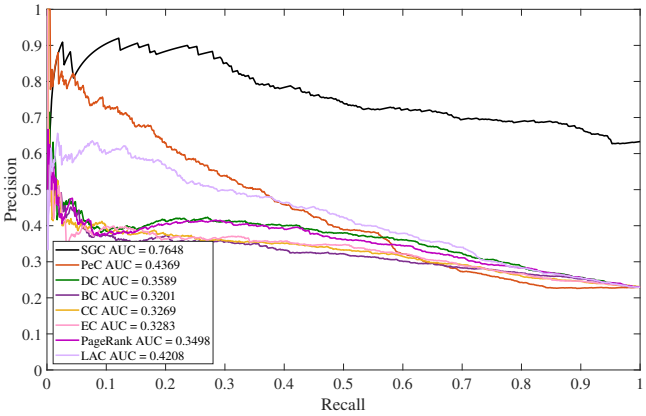

Fig. 2. Comparison of SGC, BC, CC, DC, EC, LAC, and PeC using precision-recall (PR) curve method.

AUC value, highly exceeds other methods, which also illustrates the effectiveness of our method.

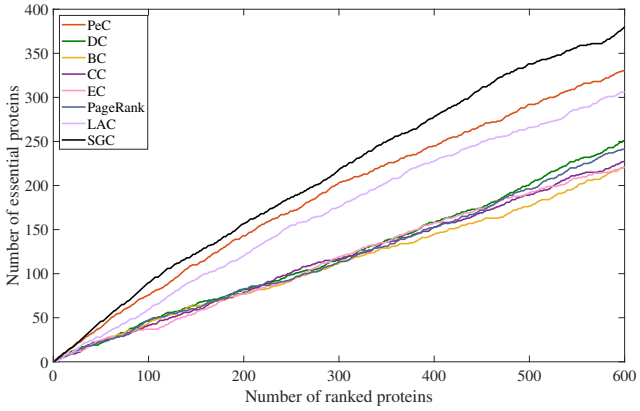

Fig. 3. Comparison of SGC, BC, CC, DC, EC, LAC, and PeC using Jackknife method.

To further evaluate the effectiveness of our method, we take the jackknife curve to compare the prediction results of our proposed method SGC with other methods. The results are shown in Fig. 3, the x-axis denotes the number of proteins ranked by each essential protein identification method and the y-axis is just the number of turely identified essential proteins of each method.

The areas under the jackknife curves can measure the performances of the method for identifying essential proteins. As shown in Fig. 3, the jackknife curve of our proposed method SGC can identify more essential proteins from the yeast PPI network compared with other methods, which also demonstrates that SGC is more effective and can get better results than other state-of-art methods.

Fig. 4 gives a more specific comparison of the results of identification of essential proteins. As shown in the Fig. 4, we can see that our method SGC can identify more essential

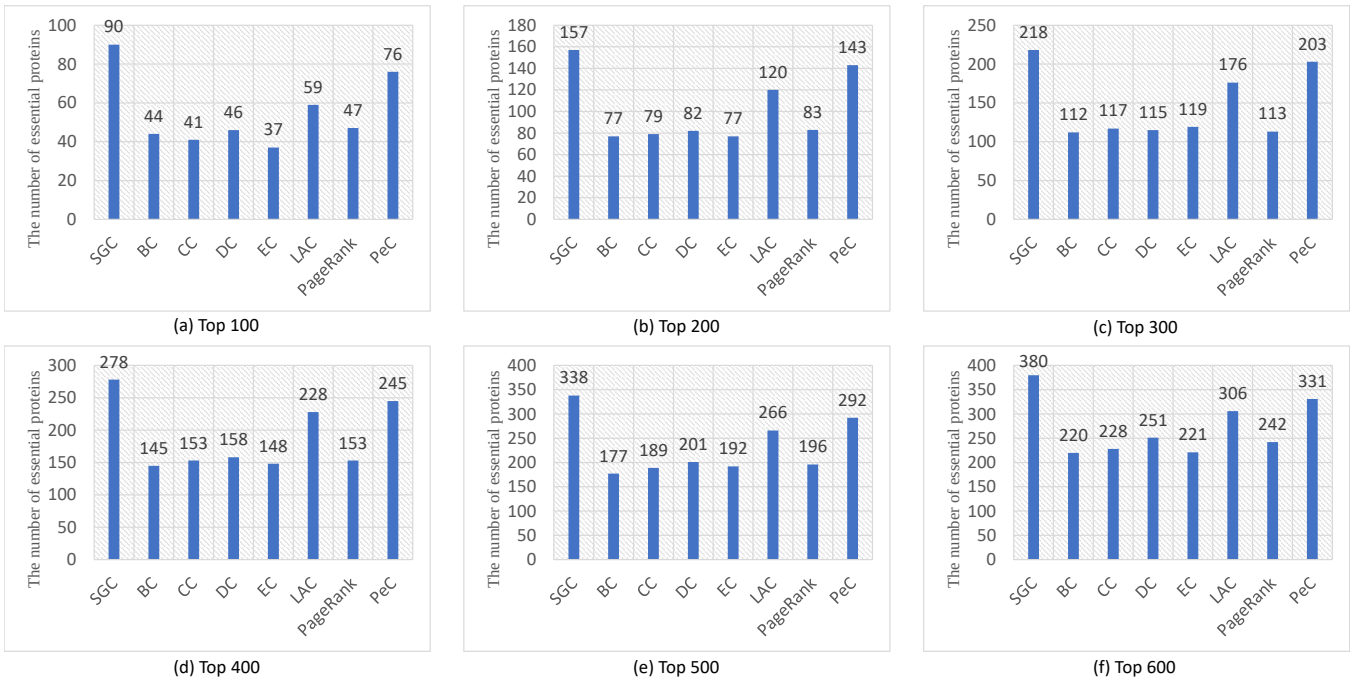

Fig. 4. The number of essential proteins predicted by SGC, BC, CC, DC, EC, LAC, and PeC. (a)-(f) show the results of these methods when select top 100 to 600 ranked proteins as candidate essential proteins.

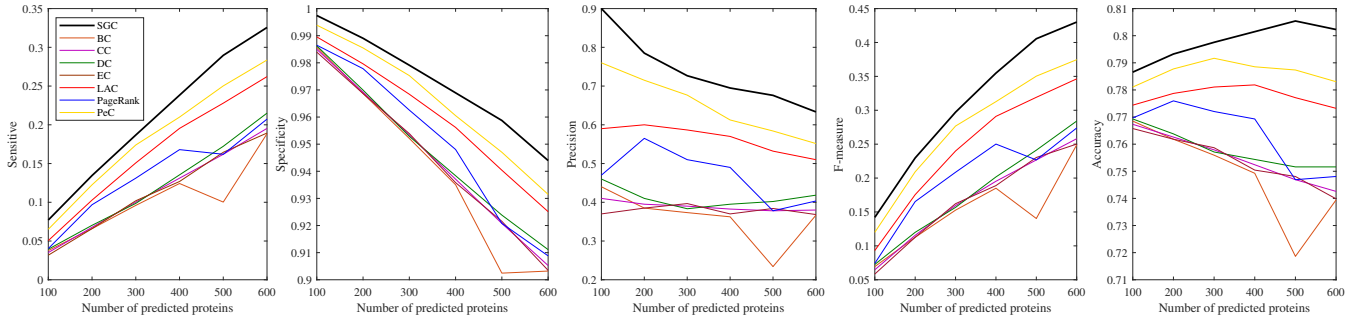

Fig. 5. Comparison of sensitive, specificity, precision, F-measure, and accuracy obtained by SGC, BC, CC, DC, EC, LAC, and PeC on the Yeast PPI dataset.

proteins compared with the other methods. The number of true essential proteins identified by SGC is both higher than other methods in the top 100, top 200, top 300, top 400, top 500, and top 600 proteins. And by observing the results of the top 100 proteins, we can see that SGC can obtain a prediction precision of 90%, which is much higher than other methods.

To further interpret the advantages of our method, we choose sensitive, specificity, Precision, F-measure, and accuracy 5 metrics to evaluate all the methods. Fig 5 shows the results of 5 evaluation metrics (sensitive, specificity, Precision, F-measure, and accuracy) obtained by all identification methods on the PPI network of Yeast. By observing Fig 5, We can see that our proposed SGC outperforms rest methods in terms of all 5 evaluation metrics.

5. CONCLUSION

6. REFERENCES

[1] Bin Xu, Jihong Guan, Yang Wang, and Zewei Wang, "Essential protein detection by random walk on weighted protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

[2] Simon J Furney, M Mar Albà, and Núria López-Bigas, "Differences in the evolutionary history of disease genes affected by dominant or recessive mutations," *BMC genomics*, vol. 7, no. 1, pp. 165, 2006.

[3] Min Li, Peng Ni, Xiaopei Chen, Jianxin Wang, Fangxiang Wu, and Yi Pan, "Construction of refined protein interaction network for predicting essential proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

[4] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41, 2001.

[5] Xionglei He and Jianzhi Zhang, "Why do hubs tend to be essential in protein networks?," *PLoS genetics*, vol. 2, no. 6, pp. e88, 2006.

[6] Ernesto Estrada and Juan A Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, no. 5, pp. 056103, 2005.

[7] Phillip Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[8] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang, "High-betweenness proteins in the yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[9] Stefan Wuchty and Peter F Stadler, "Centers of complex networks," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45–53, 2003.

[10] Karen Stephenson and Marvin Zelen, "Rethinking centrality: Methods and examples," *Social networks*, vol. 11, no. 1, pp. 1–37, 1989.

[11] Yan Wang, Huiyan Sun, Wei Du, Enrico Blanzieri, Gabriella Viero, Ying Xu, and Yanchun Liang, "Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks," *PloS one*, vol. 9, no. 9, pp. e108716, 2014.

[12] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[13] Jianxin Wang, Min Li, Huan Wang, and Yi Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1070–1080, 2012.

[14] Zhu Yuan and Wu Chong, "Identification of essential proteins using improved node and edge clustering coefficient," in *2018 37th Chinese Control Conference (CCC)*, 2018.

[15] Gaoshi Li, Min Li, Jianxin Wang, Jingli Wu, Fangxiang Wu, and Yi Pan, "Predicting essential proteins based on subcellular localization, orthology and ppi networks.," *BMC Bioinformatics*, vol. 17, no. 8, pp. 279, 2016.

[16] Min Li, Jianxin Wang, Huan Wang, and Yi Pan, "Essential proteins discovery from weighted protein interaction networks," in *International Symposium on Bioinformatics Research and Applications*. Springer, 2010, pp. 89–100.

[17] Min Li, Hanhui Zhang, Jian-xin Wang, and Yi Pan, "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, pp. 15, 2012.

[18] Jiawei Luo and Yi Qi, "Identification of essential proteins based on a new combination of local interaction density and protein complexes," *PLOS ONE*, vol. 10, no. 6, 2015.

[19] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sulmin Kim, and David Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[20] Hanswerner Mewes, Dmitrij Frishman, Klaus F X Mayer, Martin Munsterkotter, Octave Noubibou, Philipp Pagel, Thomas Rattei, Matthias Oesterheld, Andreas Ruepp, and Volker Stumpflen, "Mips: analysis and annotation of proteins from whole genomes in 2005.," *Nucleic Acids Research*, vol. 34, no. 90001, pp. 169–172, 2006.

[21] J Michael Cherry, Caroline Adler, Catherine A Ball, Stephen A Chervitz, Selina S Dwight, Erich T Hester, Yankai Jia, Gail Juvik, Taiyun Roe, Mark Schroeder, et al., "Sgd: Saccharomyces genome database," *Nucleic Acids Research*, vol. 26, no. 1, pp. 73–79, 1998.

[22] Ren Zhang and Yan Lin, "Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Research*, vol. 37, pp. 455–458, 2009.

[23] "Saccharomyces genome deletion project," [http://yeastdeletion.stanford.edu//].1:32 2020/5/26 Accessed 20 June 2012.

[24] "Compartments," [http://compartments.jensenlab.org]. Accessed 28 Dec 2014.

[25] Michele Magrane, "Uniprot knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, no. 2011, 2011.

[26] Janan T Eppig, Judith A Blake, Carol J Bult, James A Kadin, and Joel E Richardson, "The mouse genome database (mgd): comprehensive resource for genetics and genomics of the laboratory mouse," *Nucleic Acids Research*, vol. 40, pp. 881–886, 2012.

[27] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al., "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, pp. 700–705, 2012.

[28] Peter Mcquilton, Susan E St Pierre, and Jim Thurmond, "Flybase 101 – the basics of navigating flybase," *Nucleic Acids Research*, vol. 40, pp. 706–714, 2012.

[29] Todd W Harris, Igor Antoshechkin, Tamberlyn Bieri, Darin Blasiar, Juancarlos Chan, Wen J Chen, Norie De La Cruz, Paul H Davis, Margaret Duesbury, Ruihua Fang, et al., "Wormbase: a comprehensive resource for nematode research," *Nucleic Acids Research*, vol. 38, no. 2, pp. 463–467, 2010.

[30] Edgar Ron, Domrachev Michael, and Alex E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Research*, , no. 1, pp. 1, 2002.

[31] Xiaoqing Peng, Jianxin Wang, Jun Wang, Fangxiang Wu, and Yi Pan, "Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks.," *PLOS ONE*, vol. 10, no. 6, 2015.