A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene Expressions, Orthology, and Subcellular Localization Information

Min Li
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Zhibei Niu
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Xiaopei Chen
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Ping Zhong
the School of Information Science and Engineering, Central South University, Changsha 410083, China.

Fangxiang Wu
the Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

Recommended Citation
Min Li, Zhibei Niu, Xiaopei Chen et al. A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene Expressions, Orthology, and Subcellular Localization Information. Tsinghua Science and Technology 2016, 21(6): 668-677.
A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene Expressions, Orthology, and Subcellular Localization Information

Authors
Min Li, Zhibei Niu, Xiaopei Chen, Ping Zhong, Fangxiang Wu, and Yi Pan
A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene Expressions, Orthology, and Subcellular Localization Information

Min Li, Zhibei Niu, Xiaopei Chen, Ping Zhong*, Fangxiang Wu, and Yi Pan

Abstract: Essential proteins are those necessary for the survival or reproduction of species and discovering such essential proteins is fundamental for understanding the minimal requirements for cellular life, which is also meaningful to the disease study and drug design. With the development of high-throughput techniques, a large number of Protein-Protein Interactions (PPIs) can be used to identify essential proteins at the network level. Up to now, though a series of network-based computational methods have been proposed, it is still a challenge to improve the prediction precision as the high false positives in PPI networks. In this paper, we propose a new method GOS to identify essential proteins by integrating the Gene expressions, Orthology, and Subcellular localization information. The gene expressions and subcellular localization information are used to determine whether a neighbor in the PPI network is reliable. Only reliable neighbors are considered when we analyze the topological characteristics of a protein in a PPI network. We also analyze the orthologous attributes of each protein to reflect its conservative features, and use a random walk model to integrate a protein’s topological characteristics and its orthology. The experimental results on the yeast PPI network show that the proposed method GOS outperforms the ten existing methods DC, BC, CC, SC, EC, IC, NC, PeC, ION, and CSC.

Key words: essential protein; reliable neighbors; GOS; orthology; subcellular localization information

1 Introduction

Proteins are the products of gene expressions and indispensable for cells life which play important roles for various biological activities[1]. Essential proteins are the products of essential genes which can lead to cell death or infertility if one of them has been removed. The identification of essential proteins and their functions[2] not only can help researchers to understand the basic needs of life, but also can provide useful information for disease study and drug design[3, 4].

In the previous studies, biological researchers generally use gene knockouts[5], RNA interference[6] or conditional knockouts[7] to predict essential proteins on a special condition. Such biological experiments are relatively tedious, time consuming, and expensive. In the past decades, many computational approaches have been proposed as complementary and alternative methods for predicting essential proteins. Especially with the development of high-throughput technologies, such as yeast two-hybrid, tandem affinity purification, and mass spectrometry, a large number of Protein-Protein Interactions (PPIs) are available, which makes...
it possible for us to discover new essential proteins at a network level\cite{8–11}. Of course, there are also some other non-network-based methods which use different biological information. For example, Gustafson et al\cite{12} identified essential proteins by using targeted genome sequencing with the basic idea that proteins whose corresponding genes have longer sequences of Open Reading Frames (ORF) tend to be essential\cite{13}. However, the network-based approaches are the most popular ones in literature.

Generally, the network-based methods for essential protein discovery can be grouped into three categories: neighborhood-based methods, path-based methods, and iterative refinement methods. The neighborhood-based methods investigate a protein’sessentiality by considering its neighbors. The simplest one of neighborhood-based methods is Degree Centrality (DC) proposed by Jeong et al\cite{1}, which is also known as “centrality-lethality” principle. DC counts the number of neighbors for each protein and ranks all the proteins in a non-increased order according to the number of their neighbors. The studies of Jeong et al\cite{1} show that the most highly connected proteins in the cell are the most important for its survival. Although there are some controversies\cite{14, 15} whether or why highly connected proteins tend to be essential, most of the researchers confirmed the relationship between degree centrality and protein essentiality\cite{11, 15–18}. For most species, there exist a number of highly connected proteins which are not essential. In our previous studies, by analyzing the highly connected non-essential proteins in yeast, we found that a few of their neighbors interact with each other and proposed a local connectivity-based method LAC\cite{15} to determine a protein’sessentiality by evaluating the relationship between a protein and its neighbors. We also used Edge Clustering Coefficient (ECC) to describe the closeness of two connected proteins by counting their common neighbors and proposed an essential protein discovery method NC\cite{14} by calculating the sum of ECC values among proteins and their neighbors. What’s more, topology potential of PPI networks was investigated to predict essential proteins\cite{16}.

Different from the neighborhood-based methods, the path-based methods take into account the global topological characteristics such as Betweenness Centrality (BC)\cite{17, 18}, Closeness Centrality (CC)\cite{19}, Information Centrality (IC)\cite{20}, and Subgraph Centrality (SC)\cite{21}. For each protein in a given PPI network, BC\cite{17, 18} calculates the fraction of the shortest paths going through it. CC\cite{19} summarizes the distance between target protein and all the others and gets the inverse of the distance as their score so that the larger closeness of the protein can be more essential. IC\cite{20} measures the importance of a given protein by computing the information contained in all possible paths in the network from statistical estimation. SC\cite{21} evaluates the essentiality of a given protein by calculating the weighted sum of the numbers of all closed paths starting from and ending at it.

Besides the neighborhood-based methods and the path-based methods, the iterative refinement centralities are also popular for predicting essential proteins, such as Eigenvector Centrality (EC)\cite{22} that simulates a mechanism in which each node affects all of its neighbors. Moreover, some other approaches, such as Page-Rank algorithm\cite{23}, HITs\cite{24}, Leader-Rank\cite{25}, used in complex network analysis, could also be used to predict essential proteins. CytoNCA\cite{26}, a plug-in of cytoscape, has been developed to predict essential proteins by integrating eight network centralities for both weighted and unweighted networks.

Though great progresses have been made for the network-based essential protein discovery methods, its prediction precision highly depends on the reliability of the PPI network. Unfortunately, the protein-protein interactions, especially those generated by high-throughput technologies, include high false positives. To reduce the effect of noise in the PPI networks, some researchers began to introduce other biological data, such as domains, gene expressions, protein complexes, subcellular locations, and orthology, when they investigate the essential protein discovery methods. For example, Peng et al\cite{27} proposed UDoNC by integrating domains and PPI networks. PeC\cite{28}, CEPPK\cite{10}, WDC\cite{29}, and CoEWC\cite{30} were developed to predict essential proteins by fusing gene expressions and topological characteristics of PPI networks. Kim\cite{31} proposed a machine learning method by using topological characters in GO-pruned PPI network. Harmonic Centrality (HC)\cite{32} integrates the information of protein complexes into SC, United complex Centrality (UC)\cite{33} uses protein complexes data to distinguish the contributions of different edge clustering coefficients between a pair of proteins. Of these methods, both the known complexes and the predicted complexes from different computational methods\cite{34–38} can be used.
LIDC\cite{39} was developed by the combination of Local Interaction Density and protein Complexes. CSC\cite{40} uses the in-degree of proteins in complexes. POEM\cite{41} measures the essentiality of a protein by determining the overlapping essential modules which the given protein belongs to. Localization-Specificity for Essential protein Detection (LSED)\cite{42} introduces subcellular localization information when predicting essential proteins and ION\cite{8} was developed by integrating orthologous information with the topological characteristics of PPI networks. SON\cite{43} integrates subcellular localization, orthology, and PPI networks. Zhong et al.\cite{44} collected 26 different biological and topological features and used SVM-RFE to select a feature space from them to predict essential proteins. Besides, some researchers constructed dynamic networks to reduce the effect of noise in the PPI networks by integrating dynamic gene expressions and PPI networks.\cite{45–47}. For example, Xiao et al.\cite{45} constructed an active PPI network and predicted essential proteins from the active PPI network by using different network centralities. In our previous work, we purified the PPI network by integrating gene expressions and subcellular localizations to construct a reliable network TS-PIN\cite{48}. More network-based methods and other computational approaches can be seen in a comprehensive survey by Wang et al.\cite{49}

In this paper, we propose a new neighborhood-based method GOS to identify essential proteins by integrating Gene expressions, Orthology, and Subcellular localization information. The gene expressions and subcellular localization information are used to determine whether a neighbor in the PPI network is reliable. We think that it is the unreliable neighbors that affect the prediction precision of neighborhood-based methods. Hence, we investigate a protein’s topological characteristics only considering the reliable neighbors. We also analyze the orthologous attributes of each protein to reflect its conservative features, and use a random walk model to integrate a protein’s topological characteristics and its orthology. The experimental results on the yeast PPI network show that the proposed method GOS outperforms the ten existing methods DC\cite{11}, NC\cite{14}, BC\cite{17,18}, CC\cite{19}, SC\cite{21}, IC\cite{20}, EC\cite{22}, PeC\cite{28}, ION\cite{8}, and CSC\cite{40}.

2 Method

The basic idea of our proposed essential protein discovery method GOS is to improve neighborhood-based methods by determining reliable neighbors. The gene expressions and subcellular location information were used to determine whether a neighbor in the PPI network is reliable. Then only the reliable neighbors are taken into account to analyze a protein’s essentiality. Finally, the topological characteristics of a protein based on the reliable neighbors are further combined with its orthology.

2.1 Determining reliable neighbors

A PPI network is usually described as an undirected graph $G(V,E)$, where $V = \{v_1, \ldots, v_n\}$ represents the proteins and $E = \{e(v_i, v_j)\}$ for all $i, j$ is the set of edges connecting two proteins $v_i$ and $v_j$. For a given protein $v$, its neighbors are all the proteins connected to it and the neighbor is denoted as $N_v$. As a PPI network is generally constructed by all PPIs collected from different labs with different environments at different times, there may be many false interactions in the PPI network. Hence, for a given protein, there may be some false neighbors. Here, we try to distinguish the reliable neighbors from those unreliable neighbors by integrating gene expressions and subcellular localization information. According to the fact that two proteins can physically interact with each other only if they are active together at least at a time point in the cell cycle and appear together at the same subcellular location, we define reliable neighbor as follows:

**Definition 1 Reliable neighbor:** For a given protein $v$, only the neighbors which physically interact with it at at least one subcellular localization $l$ and active together with it at at least one time point $t$. The reliable neighbor set of a given protein $v$ is denoted as $RN_v$.

For example in Fig. 1, protein A has eight neighbors: B, C, D, E, F, G, H, and I. Out of the eight neighbors, only four proteins (C, D, F, H) occur at the same subcellular location with protein A. Protein A is active at the time points 1, 2, 3, and 9. Proteins B, C, D, E, F, and G are active together with it at least at one same time point. As a result, three proteins C, D, and F are reliable neighbors of protein A as they are active at the time point 2 or 3 and occur at the same subcellular location Nucleus or Mitochondrion.

2.2 Network centrality based on reliable neighbors

It has been proved that ECC is effective to describe the local closeness of two connected proteins in a
PPI network and works well on the identification of protein complexes \(^{35}\) and essential proteins \(^{14}\). In this study, we also use ECC to evaluate the closeness of a protein and its reliable neighbors. Different from the original definition of ECC, here we define new reliable neighbor-based ECC as follows.

**Definition 2 Reliable neighbor-based ECC (RECC):** For an edge \(e \in E\) connecting protein \(u\) and protein \(v\), its RECC is defined as following:

\[
\text{RECC}(u, v) = \frac{|R_u \cap R_v|}{\min(|R_u| - 1, |R_v| - 1)}
\]

(1)

where \(R_u\) and \(R_v\) are the sets of reliable neighbors of protein \(u\) and protein \(v\), respectively.

Similar to NC \(^{14}\), based on the definition of RECC, we can define the Reliable Neighbor-based network Centrality (RNC) as follows.

**Definition 3 RNC:** For a given protein \(v\), its reliable neighbor-based network centrality \(\text{RNC}(v)\) is defined as the sum of RECC between it and its reliable neighbors.

\[
\text{RNC}(u, v) = \sum_{u \in R_v} \text{RECC}(u, v)
\]

(2)

### 2.3 Essential protein discovery method GOS

Considering that a protein’s conservation is highly related to its essentiality we also performed the orthology analysis as in Ref. [8]. We collected the protein orthologous information from the InParanoid database \(^{50}\), which includes 99 eukaryotes and 1 prokaryote constructed by the INPARANIOD program. The yeast PPIs used in this study was downloaded from the DIP database \(^{51}\). The final yeast PPI network contains 5093 proteins and 24 743 edges after filtering self interactions and repeats. Out of the 5093 yeast proteins, 4511 proteins have orthologous proteins in at least one reference species. Out of the 1167 essential proteins in the PPI network, 1118 have orthologous proteins in at least one reference species. The analysis and previous studies all show that a protein’s conservation is highly related to its essentiality. Hence, we further combined a protein’s conservation with its RNC.

A protein’s conservation is evaluated by the number of reference species in which its orthologous proteins exist. Let \(R\) be the set of reference organisms which is used to get orthologous information for the proteins in the PPI network \(G(V, E)\). For a specific reference species \(i\), we use \(X_i\) to represent the subset of node \(V\) in which its element has orthologs in organism \(s\). Let \(O(v)\) be the number of times that protein \(v \in V\) has orthologs in reference organisms.

\[
O(v) = \sum_{i=1}^{K} T_v(i)
\]

(3)

\[
T_v(i) = \begin{cases} 
1, & \text{if } v \in X_i; \\
0, & \text{otherwise}
\end{cases}
\]

(4)

where \(K\) is the number of reference species.

Then for a given protein \(v\), its orthologous score \(\text{OS}(v)\) is defined as the normalized value of \(O(v)\):

\[
\text{OS}(v) = \frac{O(v)}{\text{Max}(O(u))}, \quad u \in V
\]

(5)
Finally, a linear combination model is used to integrate RNC and the orthologous score. For a given protein \( v \), its essentiality is evaluated by \( GOS(v) \):

\[
GOS(v) = \alpha \times RNC(v) + (1 - \alpha) \times OS(v)
\]

where \( \alpha \) is a parameter to adjust the contributions of RNC and OS. When \( \alpha = 0 \), only the orthologous information is considered, and when \( \alpha = 1 \), only the topological character RNC is considered. In this study, 0.5 is used as a default value for \( \alpha \).

3 Results and Discussion

To validate the performance of proposed method GOS, we carry out a comparison between GOS and ten exciting essential protein discovery methods: DC\[1\], NC\[14\], BC\[17, 18\], CC\[19\], SC\[21\], IC\[20\], Ec\[22\], PeC\[28\], ION\[8\], and CSC\[40\]. All these methods were implemented on the yeast PPI network as the essential proteins of yeast were the most complete one and were well studied. The biological data used in this study are described as follows.

**PPI network** The yeast PPIs were downloaded from the DIP database (http://dip.mbi.ucla.edu/dip/)[51]. The final yeast PPI network was constructed by using these PPIs after filtering the repeated and self-interactions. There are 5093 proteins and 24743 interactions in the final PPI network.

**Gene expression data** The yeast gene expression data were obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo/), called GSE3431. It contains 6777 gene expression profiles at 36 time points from three consecutive metabolic cycles, each cycle contains 12 time points. The corresponding proteins of 6777 genes cover 95% of the proteins in the PPI network.

**Orthologous information** The protein orthologous information was collected from the InParanoid database with version 7 (http://inparanoid.sbc.su.se/cgi-bin/index.cgi)[37], which contains 100 whole genomes (99 eukaryotes and 1 prokaryote) constructed by the INPARANOID program.

**Subcellular localization annotation information** The protein subcellular localization annotation information of yeast was obtained from the COMpartMENTS database (http://compartments.jensenlab.org/Download)[52], which integrates the experimental-based subcellular localization annotation information from UniProtKB[53], MGI[54], SGD[55], FlyBase[56], and WormBase database[57]. In total, the yeast proteins have 11 subcellular localizations.

**Known essential proteins** The known essential proteins data were collected from four different databases: MIPS[58], SGD (http://www.yeastgenome.org/)[55], DEG[59], and SGDp’s (http://www.sequence.stanford.edu/group/yeast_deletion_project). In total, we collected 1285 essential proteins from these four databases. After mapping to the PPI network, we got 1167 known essential proteins by removing those unmapped ones.

3.1 Compare GOS with other methods

To validate the effectiveness of our proposed method GOS, we compare it with different types of other approaches. First, we compare it with two other typical neighbor-based methods DC and NC. Similar to previous experimental procedures[8, 33], we rank the proteins in descending order and choose top 100, top 200, top 300, top 400, top 500, and top 600 proteins as essential candidates for each method. Then we calculate how many candidates are true essential proteins based on the collected known essential protein set. The comparison results of DC, NC, and GOS were shown in Fig. 2. From Fig. 2, we can see that GOS outperforms DC and NC obviously when predicting no more than 600 candidates. Taking top 100 predicted essential proteins for example, 89 essential proteins are correctly identified by the GOS while 46 and 56 are correctly predicted by DC and NC, respectively. For predicting no more 600 essential candidates, GOS achieves more than 50% improvements compared with DC, and more than 20% improvements compared with NC.

Then, we also compare GOS with four path-based methods BC, CC, SC, and IC. As shown in Fig. 2, when predicting no more than 600 candidates, GOS outperforms these four methods obviously too. Especially, GOS achieves more than 97% improvements compared to BC when selecting up to 400 candidates. With the increase of the number of the selected essential candidates, less improvement is obtained by GOS. However, even if the top 600 proteins are selected as essential candidates, the number of true essential proteins produced by GOS is 73% higher than that produced by BC.

The iterative refinement centrality EC is also compared with GOS. As shown in Fig. 2, the number
Min Li et al.  A Reliable Neighbor-Based Method for Identifying Essential Proteins by Integrating Gene...

of true essential proteins produced by GOS is much higher than that produced by EC. As GOS integrates biological information, we also compare it with several other network-based methods by integrating biological information. Here, PeC\(^{[28]}\) which integrates gene expressions, ION\(^{[8]}\) which integrates orthologous information, and CSC\(^{[40]}\) which integrates protein complexes were compared respectively. As shown in Fig. 2, GOS still outperforms these methods consistently even not so much improvement is obtained compared with those only network-based methods.

3.2 Evaluation using Jackknife curve and precision-recall curve

To further verify the performance of GOS and other ten network-based methods for predicting essential proteins, we use Jackknife method. The experimental results validated by Jackknife method are shown in Fig. 3. In Fig. 3, the horizontal axis represents the top \(N\) essential candidates and the vertical axis represents the accumulation quantity of the correct predictions for each method. The Area Under Curve (AUC) corresponding to each method is used to measure their performance. The bigger the area is, the better performance the method has. As shown in Fig. 3, our proposed method GOS performs better than the ten other network-based methods DC, NC, BC, CC, SC, IC, EC, PeC, ION, and CSC consistently in terms of AUC. It demonstrates that GOS is effective to predict yeast essential proteins and superior to the ten existing methods.

In addition, we calculate the precision and recall of GOS and the ten network-based methods DC, NC, BC, CC, SC, IC, EC, PeC, ION, and CSC, respectively. The precision-recall curve for each method is shown in Fig. 4. From Fig. 4 we can see that the precision-recall curve still supports that GOS outperforms the ten network-based methods DC, NC, BC, CC, SC, IC, EC, PeC, ION, and CSC for predicting yeast essential proteins.
proteins.

3.3 Effect of the parameter $\alpha$ on the results

In the above discussions, the default value $\alpha = 0.5$ is used in GOS. To analyze the effect of the parameter $\alpha$ on the results of GOS, we set $\alpha$ vary from 0 to 1 and observe the number of true essential proteins identified by GOS. The analysis results are shown in Table 1 with $\alpha$ varying from 0, 0.1, 0.2, \ldots to 1.0. From Table 1 we can see that GOS performs the worst when $\alpha = 0$ or 1. That is to say, both the orthologous information and the topological characteristics RNC contribute to the final results. When $\alpha$ varies from 0.4 to 0.7, there are not too much changes for GOS while GOS performs slightly better when $\alpha = 0.7$ for predicting no more than 100 essential candidates. However, when predicting more essential candidates, GOS with smaller $\alpha$ will perform better. In addition, we also use the precision-recall curve to show the effect of parameter $\alpha$, as shown in Fig. 5 where similar results can be observed.

Table 1 The number of true essential proteins identified by GOS correctly with different $\alpha$.

| $\alpha$ | [Top100] | [Top200] | [Top300] | [Top400] | [Top500] | [Top600] |
|----------|----------|----------|----------|----------|----------|----------|
| 0        | 82       | 142      | 198      | 255      | 300      | 351      |
| 0.1      | 83       | 149      | 214      | 276      | 333      | 371      |
| 0.2      | 84       | 155      | 222      | 284      | 347      | 392      |
| 0.3      | 86       | 163      | 229      | 290      | 338      | 383      |
| 0.4      | 86       | 166      | 230      | 290      | 341      | 388      |
| 0.5      | 89       | 169      | 222      | 286      | 335      | 382      |
| 0.6      | 92       | 168      | 226      | 283      | 328      | 375      |
| 0.7      | 93       | 167      | 223      | 282      | 326      | 369      |
| 0.8      | 89       | 160      | 220      | 279      | 326      | 365      |
| 0.9      | 85       | 157      | 218      | 275      | 324      | 362      |
| 1.0      | 75       | 150      | 216      | 272      | 327      | 358      |

4 Conclusion

It is believed that the identification of essential proteins is very useful to disease study and drug design. In this study, we have presented a new neighbor-based essential protein discovery method GOS and tested it on the yeast PPI network. We have compared GOS with two neighbor-based methods DC and NC, four path-based methods BC, CC, SC, and IC, an iterative refinement centrality EC, and three other network-based methods with the integration of different biological data PeC, ION, and CSC. The comparison results have shown GOS outperforms these ten methods for predicting yeast essential proteins. Our experimental results have also shown that the reliable neighbors can effectively reduce the effect of false positives in the PPI networks as both the reliable neighbor-based network centrality and the conservation contribute to predicting essential proteins more accurately.

Acknowledgment

This work was supported by the National Natural Science Foundation for Excellent Young Scholars (No. 61622213) and the National Natural Science Foundation of China (Nos. 61232001, 61370024, and 61428209).

References

[1] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Lethality and centrality in protein networks. Nature, vol. 411, no. 6833, pp. 41–42, 2001.
[2] W. Peng, M. Li, L. Chen, and L. Wang, Predicting protein functions by using unbalanced random walk algorithm on three biological networks, 2015. doi: 10.1109/TCBB.2015.2394314.
[3] M. Li, R. Zheng, Q. Li, J. Wang, F.-X. Wu, and Z. Zhang, Prioritizing disease genes by using search engine
algorithm, *Current Bioinformatics*, vol. 11, no. 2, pp. 195–202, 2016.

[4] W. Lan, J. Wang, M. Li, W. Peng, and F. Wu, Computational approaches for prioritizing candidate disease genes based on ppi networks, *Tsinghua Science and Technology*, vol. 20, no. 5, pp. 500–512, 2015.

[5] H. Lu, Research on genic function by clustering on protein network, PhD dissertation, Chinese Academy of Sciences, Beijing, China, 2006.

[6] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*. Springer, 2013.

[7] J. Xu and Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

[8] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks, *BMC Systems Biology*, vol. 6, no. 1, p. 1, 2012.

[9] M. Li, J.-X. Wang, H. Wang, and Y. Pan, Identification of essential proteins from weighted protein–protein interaction networks, *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 3, p. 1341002, 2013.

[10] M. Li, R. Zheng, H. Zhang, J. Wang, and Y. Pan, Effective identification of essential proteins based on priori knowledge, network topology and gene expressions, *Methods*, vol. 67, no. 3, pp. 325–333, 2014.

[11] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, *Proteomics*, vol. 6, no. 1, pp. 35–40, 2006.

[12] A. M. Gustafson, E. S. Snitkin, S. C. Parker, C. DeLisi, and S. Kasif, Towards the identification of essential genes using targeted genome sequencing and comparative analysis, *BMC Genomics*, vol. 7, no. 1, p. 1, 2006.

[13] Y.-C. Hwang, C.-C. Lin, J.-Y. Chang, H. Mori, H.-F. Juan, and H.-C. Huang, Predicting essential genes based on network and sequence analysis, *Molecular BioSystems*, vol. 5, no. 12, pp. 1672–1678, 2009.

[14] J. Wang, M. Li, H. Wang, and Y. 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.

[15] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, A local average connectivity-based method for identifying essential proteins from the network level, *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 143–150, 2011.

[16] M. Li, Y. Lu, J. Wang, F.-X. Wu, and Y. Pan, A topology potential-based method for identifying essential proteins from ppi networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 2, pp. 372–383, 2015.

[17] M. E. Newman, A measure of betweenness centrality based on random walks, *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005.

[18] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, High-betweenness proteins in the yeast protein interaction network, *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[19] S. Wuchty and P. F. Studler, Centers of complex networks, *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45–53, 2003.

[20] K. Stephenson and M. Zelen, Rethinking centrality: Methods and examples, *Social Networks*, vol. 11, no. 1, pp. 1–37, 1989.

[21] E. Estrada and J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.

[22] P. Bonacich, Power and centrality: A family of measures, *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[23] S. Brin and L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, *Computer Networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[24] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[25] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, Leaders in social networks, the delicious case, *PloS One*, vol. 6, no. 6, p. e21202, 2011.

[26] Y. Tang, M. Li, J. Wang, Y. Pan, and F.-X. Wu, Cytonca: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks, *Biosystems*, vol. 127, pp. 67–72, 2015.

[27] W. Peng, J. Wang, Y. Cheng, Y. Lu, F. Wu, and Y. Pan, Udono: An algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 2, pp. 276–288, 2015.

[28] M. Li, H. Zhang, J. X. Wang, and Y. Pan, A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data, *BMC Systems Biology*, vol. 6, no. 1, p. 1, 2012.

[29] X. Tang, J. Wang, J. Zhong, and Y. Pan, Predicting essential proteins based on weighted degree centrality, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 407–418, 2014.

[30] X. Zhang, J. Xu, and W. X. Xiao, A new method for the discovery of essential proteins, *PloS One*, vol. 8, no. 3, p. e58763, 2013.

[31] W. Kim, Prediction of essential proteins using topological properties in go-pruned ppi network based on machine learning methods, *Tsinghua Science and Technology*, vol. 17, no. 6, pp. 645–658, 2012.

[32] J. Ren, J. Wang, M. Li, and F. Wu, Discovering essential proteins based on ppi network and protein complex, *International Journal of Data Mining and Bioinformatics*, vol. 12, no. 1, pp. 24–43, 2013.

[33] M. Li, Y. Lu, Z. Niu, and F. Wu, United complex centrality for identification of essential proteins from ppi networks, 2015. doi: 10.1109/TCBB.2015.2394487.
[34] B. Zhao, J. Wang, M. Li, F. X. Wu, and Y. Pan, Detecting protein complexes based on uncertain graph model, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 11, no. 3, pp. 486–497, 2014.

[35] J. Wang, M. Li, J. Chen, and Y. Pan, A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 3, pp. 607–620, 2011.

[36] M. Li, J. E. Chen, J. X. Wang, B. Hu, and G. Chen, Modifying the dpclus algorithm for identifying protein complexes based on new topological structures, BMC Bioinformatics, vol. 9, no. 1, p. 1, 2008.

[37] J. Wang, J. Zhong, G. Chen, M. Li, F. X. Wu, and Y. Pan, Clusterviz: A cytoscape app for cluster analysis of biological network, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 12, no. 4, pp. 815–822, 2015.

[38] M. Li, Y. Tang, X. Wu, J. Wang, F. X. Wu, and Y. Pan, C-deva: Detection, evaluation, visualization and annotation of clusters from biological networks, Biosystems, vol. 150, pp. 78–86, 2016.

[39] J. Luo and Y. Qi, Identification of essential proteins based on a new combination of local interaction density and protein complexes, PloS One, vol. 10, no. 6, e0131418, 2015.

[40] J. Luo and L. Ma, A new integration-centric algorithm of identifying essential proteins based on topology structure of protein-protein interaction network and complex information, Current Bioinformatics, vol. 8, no. 3, pp. 380–385, 2013.

[41] B. Zhao, J. Wang, M. Li, F. X. Wu, and Y. Pan, Prediction of essential proteins based on overlapping essential modules, IEEE Transactions on Nanobioscience, vol. 13, no. 4, pp. 415–424, 2014.

[42] X. Peng, J. Wang, J. Wang, F. X. Wu, and Y. Pan, Rechecking the centrality-lutility rule in the scope of protein subcellular localization interaction networks, PLoS One, vol. 10, no. 6, e0130743, 2015.

[43] G. Li, M. Li, J. Wang, J. Wu, F. X. Wu, and Y. Pan, Predicting essential proteins based on subcellular localization, orthology and ppi networks, BMC Bioinformatics, vol. 17, no. 8, p. 279, 2016.

[44] J. Zhong, J. Wang, W. Peng, Z. Zhang, and M. Li, A feature selection method for prediction essential protein, Tsinghua Science and Technology, vol. 20, no. 5, pp. 491–499, 2015.

[45] Q. Xiao, J. Wang, X. Peng, F. X. Wu, and Y. Pan, Identifying essential proteins from active ppi networks constructed with dynamic gene expression, BMC Genomics, vol. 16, no. Suppl 3, p. S1, 2015.

[46] M. Li, X. Wu, J. Wang, and Y. Pan, Towards the identification of protein complexes and functional modules by integrating ppi network and gene expression data, BMC Bioinformatics, vol. 13, no. 1, p. 1, 2012.

[47] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, A comparison of the functional modules identified from time course and static ppi network data, BMC Bioinformatics, vol. 12, no. 1, p. 1, 2011.

[48] M. Li, X. Chen, P. Ni, J. Wang, and Y. Pan, Identifying essential proteins by purifying protein interaction networks, in International Symposium on Bioinformatics Research and Applications, 2016, pp. 106–116.

[49] J. Wang, W. Peng, and F. X. Wu, Computational approaches to predicting essential proteins: A survey, PROTEOMICS-Clinical Applications, vol. 7, nos. 1&2, pp. 181–192, 2013.

[50] G. Östlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Missina, S. Roopra, O. Frings, and E. L. Sonnhammer, Inparanoid 7: New algorithms and tools for eukaryotic orthology analysis, Nucleic Acids Research, vol. 38, no. suppl 1, pp. D196–D203, 2010.

[51] K. Ning, H. K. Ng, S. Srihari, H. W. Leong, and A. I. Nesvizhskii, Examination of the relationship between essential genes in ppi network and hub proteins in reverse nearest neighbor topology, BMC Bioinformatics, vol. 11, no. 1, p. 1, 2010.

[52] X. He and J. Zhang, Why do hubs tend to be essential in protein networks?, PLoS Genet., vol. 2, no. 6, e88, 2006.

[53] D. Edwards, Plant Bioinformatics: Methods and Protocols, Springer, 2008.

[54] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, The human disease network, Proceedings of the National Academy of Sciences, vol. 104, no. 21, pp. 8685–8690, 2007.

[55] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al., Sgd: Saccharomyces genome database, Nucleic Acids Research, vol. 26, no. 1, pp. 73–79, 1998.

[56] S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, et al., Flybase: Enhancing drosophila gene ontology annotations, Nucleic Acids Research, vol. 37, no. suppl 1, pp. D555–D559, 2009.

[57] L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spith, Wormbase: Network access to the genome and biology of caenorhabditis elegans, Nucleic Acids Research, vol. 29, no. 1, pp. 82–86, 2001.

[58] H.-W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, S. Tweedie, M. Ashburner, K. Falls, P. Leyland, M. Schroeder, et al., Sgd: Saccharomyces genome database, Nucleic Acids Research, vol. 30, no. 1, pp. 31–34, 2002.

[59] R. Zhang and Y. Lin, Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes, Nucleic Acids Research, vol. 37, no. suppl 1, pp. D455–D458, 2009.
Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently a professor at the School of Information Science and Engineering, Central South University, Changsha, China. Her main research interests include bioinformatics and systems biology.

Zhibei Niu is a master student in computer science at Central South University. She received the bachelor degree from Central South University, China, in 2014. Her main research interests include biological network analysis and essential protein discovery.

Xiaopei Chen is a master student in computer science at Central South University. Her main research interests include biological network analysis and essential protein discovery. She received the bachelor degree from Central South University, China, in 2013.

Ping Zhong received the PhD degree in communication engineering from Xiamen University, China, in 2011. She is currently a lecturer with the Department of Computer Science and technology at Central South University. Her research interests include machine learning, data mining, and networks protocol design. She is a member of ACM and IEEE.

Fangxiang Wu received the BSc and MSc degrees in applied mathematics, both from Dalian University of Technology, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from Northwestern Polytechnical University, Canada, in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan, Canada, in 2004. Currently, he is working as an associate professor of bioengineering with the Department of Mechanical Engineering and graduate chair of the Division of Biomedical Engineering at the University of Saskatchewan, Canada. His current research interests include systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, and applications of control theory to biological system.

Yi Pan is a Regents’ Professor of computer science and an Interim Associate Dean and Chair of Biology at Georgia State University, USA. Dr. Pan joined Georgia State University in 2000 and was promoted to full professor in 2004, named a Distinguished University Professor in 2013 and designated a Regents’ Professor (the highest recognition given to a faculty member by the University System of Georgia) in 2015. He served as the Chair of Computer Science Department from 2005-2013. He is also a visiting Changjiang Chair Professor at Central South University, China. Dr. Pan received the BEng and MEng degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and the PhD degree in computer science from the University of Pittsburgh, USA, in 1991. His profile has been featured as a distinguished alumnus in both Tsinghua Alumni Newsletter and University of Pittsburgh CS Alumni Newsletter. Dr. Pan’s research interests include parallel and cloud computing, wireless networks, and bioinformatics. Dr. Pan has published more than 330 papers including over 180 SCI journal papers and 60 IEEE/ACM Transactions papers. In addition, he has edited/authored 40 books. His work has been cited more than 6500 times. Dr. Pan has served as an editor-in-chief or editorial board member for 15 journals including 7 IEEE Transactions. He is the recipient of many awards including IEEE Transactions Best Paper Award, 4 other international conference or journal Best Paper Awards, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized many international conferences and delivered keynote speeches at over 50 international conferences around the world.