RESEARCH

A novel essential protein identification method based on PPI networks and gene expression data

Jiancheng Zhong[1,2], Chao Tang[1], Wei Peng[3], Minzhu Xie[1], Yusui Sun[1], Qiang Tang[4], Qiu Xiao[1]*, Jiahong Yang[1]*

Abstract

Background: Some proposed methods for identifying essential proteins have better results by using biological information. Gene expression data is generally used to identify essential proteins. However, gene expression data is prone to fluctuations, which may affect the accuracy of essential protein identification. Therefore, we propose an essential protein identification method to calculate the similarity of "active" and "inactive" state of gene expression in a cluster of the PPI network based on gene expression and the PPI network data. Our experiments show that our method can improve the accuracy in predicting essential proteins.

Results: In this paper, we propose a new measure, named JDC, based on the PPI network data and gene expression data. The JDC method offers a dynamic threshold method to binarize gene expression data. After that, it combines the degree centrality and Jaccard similarity index to calculate the JDC score for each protein in the PPI network. We respectively perform experiments on Yeast data and E.coli data and evaluate our method by using ROC analysis, modular analysis, jackknife analysis, overlapping analysis, top analysis, and accuracy analysis. The results show that the performance of JDC is better than DC, IC, EC, SC, BC, CC, NC, PeC, and WDC. We compare JDC with both NF-PIN and TS-PIN methods, which predict essential proteins from active PPI networks constructed with dynamic gene expression.

Conclusions: We demonstrate that the new centrality measure, JDC, is more efficient than state-of-the-art prediction methods. The main ideas behind JDC are as follows: (1) Essential proteins are generally densely connected clusters in the PPI network. (2) Binarizing gene expression data can screen out fluctuations in gene expression profiles. (3) The essentiality of the protein depends on the similarity of "active" and "inactive" state of gene expression in a cluster of the PPI network.

Keywords: essential proteins; the PPI networks; Jaccard similarity index; Edge clustering coefficient.

Background

Proteins are generally involved in the life activities of organisms. Essential proteins are often found in protein complexes. Loss of essential proteins could cause lethality and even lead to the inability of the body to survive[1, 2].

Therefore, the identification of essential proteins not only helps us understand the minimal requirements for cell life but also plays a vital role in the discovery of human disease genes.

*Correspondence: xiaoqiu@hunnu.edu.cn, jiahong_yang@hunnu.edu.cn

[1] the School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China.

Full list of author information is available at the end of the article

essential proteins, such as a single gene knockout[3], RNA interference[4], and conditional knockouts[5]. Although experimental methods have achieved excellent results, it has insufficient such as time-consuming and expensive. Nowadays, a variety of biological data have been generating rapidly by high-throughput experimental technologies, such as genomics, transcriptomics, and proteomics datasets. It has become possible for researchers to identify essential proteins with computational methods. The computational methods can be classified into two categories: unsupervised and supervised machine learning methods.

Unsupervised methods usually identify essential proteins based on some essentiality-related data, including the PPI networks, cellular localization data, and gene expressing

lethality rule. Because essential proteins in the PPI network are more likely to be hubs nodes, and elimination of hubs nodes may cause the PPI network to break down. Various centrality measures for prediction of essential proteins include Degree Centrality (DC)[6], Betweenness Centrality (BC)[7], Closeness Centrality (CC)[8], Subgraph Centrality (SC)[9], Eigenvector Centrality (EC)[10], Information Centrality (IC)[11]. However, these measures only take the topological features of the PPI network and ignore false positives of the PPI network. Some researchers adopt biological information to eliminate the effect of false-positive data on the PPI network. Li and Tang et al. propose essential protein prediction methods called PeC and WDC by combining the PPI network and gene expression information[12, 13]. Compared with non-essential proteins, essential proteins tend to be conserved. According to this observation, Peng et al. adopt the orthology and PPI networks to predict essential proteins[14]. Li et al. utilize an Extended Pareto Optimality Consensus model to find the triangular structure in the PPI network and combine the orthology information for the prediction of essential proteins[15].

With the generation and improvement of multi-omics data, it has become possible to construct comprehensive dynamic networks to identify essential proteins. For better predicting essential proteins, Lichtenberg et al. build a time series dynamic network by combining gene expression data at different time points and the protein interactions data[16]. Xiao et al. propose a prediction method by constructing NF-PIN dynamic network using the time series model and 3_sigma principle to filter out the noise of gene expression[17]. Recently, Li et al. construct TS-PIN dynamic network by combining gene expression profile and subcellular localization information to predict essential proteins[18]. Li et al. introduce a sub-network partition method to predict essential proteins by using the subcellular localization information[19]. Fan et al. adopted an improved PageRank algorithm to identify essential proteins based on gene expression and subcellular localization information[20]. Lei et al. incorporate the multiple biological characteristics, including PPI network, GO annotation data, subcellular localization information, and protein complexes information, to identify essential proteins by using random walk algorithms[21]. Zhang et al. propose a method to predict essential proteins by fusing dynamic PPI networks[22].Li et al. identify essential proteins by computing each protein's topology potential[23]. Peng et al. propose the UDoNC method to predict the essential proteins[24].

On the other hand, some prediction methods have adopted supervised learning methods and used machine learning algorithms to identify essential proteins, such as SVM, Random Tree, RBF network, and Naïve Bayes. Gustafson et al. proposed using Naïve Bayes to identify essential proteins based on gene expression data and topological features in the PPI network[25]. Hwang et al. constructed an SVM classifier by using some biological features (such as ORF, ST, PHY) and some topological features (such as DC, BD, CC) of the PPI network[26]. Zhong et al. adopt the GEP method and an XGBFEMF framework to predict the essential proteins[27, 28]. Deng et al. predict essential proteins by combining Naïve Bayes classifier, C4.5 decision tree, CN2 rule, and logistical regression model[29]. Kim et al. adopt machine learning methods to predict essential proteins by using topological properties in the GO-pruned PPI network[30]. Recently, Zeng et al. design a deep learning framework for the prediction of essential proteins[31].

The methods based on PPI network and gene expression data may, to some extent, eliminate false positive and false negative of protein interaction data. However, the gene expression profile is a set of values with large fluctuations and may affect prediction performance. When studying complex biological systems, Niehrs et al. point out that the "on" and "off" of genes at different times played an important role in biological development[32]. To introduce the "on" and "off" of states of genes, we propose an essential protein prediction method based on the PPI data and gene expression data, called JDC by using the essential Degree Centrality with Jaccard similarity index. JDC can eliminate the fluctuations of gene expression data by calculating the similarity of "active" and "inactive" state of gene expression in a cluster of the PPI network. Compared with the state-of-the-art methods on Yeast data and E.coli data, our method is more accurate and has higher specificity and sensitivity.

Methods

Overview

Figure 1 illustrates an example of JDC to predict essential proteins. The JDC algorithm incorporates gene expression information and PPI network data. The whole process of JDC includes the following steps.1) ECC is used to characterize the probability of two proteins being in a cluster from a topology perspective 2) A dynamic threshold is set to binarize gene expression data for filtering out the fluctuations in gene expression profiles. 3) The Jaccard similarity index measures the similarity of two proteins that has the “active” and “inactive” state of gene expression profiles; 4) JDC scores are calculated by integrating the ECC values and Jaccard similarity index. Based on those steps, we use top rank analysis in the JDC value to verify the performance of our method.

Experimental Datasets

Yeast data was widely used in various kinds of essential proteins prediction methods, so we adopt saccharomyces cerevisiae (Bakers’ Yeast) data to evaluate the JDC method. We also used E.coli data to verify the generality of the JDC method.

The PPI data of Yeast and E.coli were from the DIP database. The PPI network of E. coli has 2727 proteins and 11803 edges after filtering the self-interactions and the repeated interactions. There were 5093 proteins and 24743 edges in the PPI network of Yeast.

Essential proteins were integrated by the four databases of MIPS [33], SGD[34], DEG[35], and SGDP[36]. In the PPI network of Yeast obtained 1167 essential proteins. Out of all 2727 proteins in the E.coli network, 254 were essential. The Gene Expression data of Yeast was downloaded from the NCBI Gene Expression Omnibus website. After pretreatment and normalization, 6777 gene products and 36 samples were obtained. Similarly, the gene expression data of E.coli was also downloaded from this website.

After removing the redundant data, the E.coli gene expression data had 7312 genes and 8 samples.

Edge clustering coefficient (ECC)

Radicchi et al. first propose the edge clustering coefficient that is an important topological feature in computational networks[37]. Wang et al. adopt the edge clustering coefficient in the yeast PPI network to predict essential proteins, which also has achieved a good detection effect[38]. The advantage of the edge clustering coefficient is to describe the clustering characteristics of PPI networks from the perspective of topology. We adopt the ECC shown in formula 1 for our method to calculate the topological attribute of the two nodes, *i* and *j*:

$$\mathrm{ECC}(i,j) = \frac{z_{i,j}^{(3)}}{\min(k_i - 1, k_j - 1)} \quad (1)$$

Where $z_{i,j}^{(3)}$ denotes the number of actual triangles formed by the edge (i,j) in PPI networks, then, the number of possible triangles determined by the minimum degree of node *i* and *j* is defined as $\min(k_i - 1, k_j - 1)$. ECC is used to describe how tightly two proteins are connected. The larger the ECC value is, the more likely two connected proteins are in the same cluster. Thus, we divide the PPI network into multiple clusters by calculating the value of the ECC of each pair of interacting proteins.

Binarization of Gene Expression Data

Gene expression data is continuous and produced from microarray experiments. However, the gene expression from high-throughput experiments are prone to large fluctuations. Sahoo et al. performed a Boolean analysis of mouse B cell gene expression data to understand gene regulation and gene function[39]. To eliminate fluctuation of gene expression, in this paper, we use a threshold strategy to covert the continuous values to the discrete state values, and then characterize gene expression data with "active" and "inactive" state.

In this paper, we select one sigma value near the mean value as the threshold for screening the "active" and "inactive" state of gene expressions. Formula 2 is the mean of gene expression data. Formula 3 is the standard deviation of gene expression, and Formula 4 is the volatility of gene expression. The threshold parameter is defined in Formula 5.

$$U(i) = \frac{\sum_{t=1}^{n} E_t^{(i)}}{n} \quad (2)$$

$$\sigma^2(i) = \frac{\sum_{t=1}^{n} \left(U(i) - E_t^{(i)}\right)^2}{n} \quad (3)$$

$$V(i) = \frac{1}{1 + \sigma^2(i)} \quad (4)$$

$$G(i) = U(i) + 2 * \sigma(i) * V(i) \quad (5)$$

Where $E_t^{(i)}$ is the expression value of protein *i* at time point *t*, $U(i)$ is the mean of expression value of protein *i*, $\sigma(i)$ is the standard deviation of expression data of protein *i*, $V(i)$ is the volatility of expression value of protein *i*, $G(i)$ is the threshold parameters of expression value of protein *i*.

G denotes a matrix constructed from gene expression data, *N* is the number of genes, and *M* is the time of proteins:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1M} \\ \vdots & \ddots & \vdots \\ s_{N1} & \cdots & s_{NM} \end{pmatrix} \quad (6)$$

Where $s_{i,t}$ is the expression level of protein *i* at time *t*. If the expression value of $s_{i,t}$ is higher than the specified threshold, the "active" gene expression is defined as "1". If the value of $s_{i,t}$ is not higher than the specified threshold $G(i)$, it is "inactive" gene expression and defined as "0". The calculation formula is as follows:

$$s'_{i,t} = \begin{cases} 1, & s_{i,t} > G(i) \\ 0, & s_{i,t} \le G(i) \end{cases} \quad (7)$$

Where $s'_{i,t}$ is the activity of protein i at time t. S is updated to the matrix with Boolean values. In this paper, the gene expression data are transformed into Boolean values that can reflect the "active" and "inactive" state of gene expression.

Jaccard similarity index

The Jaccard coefficient is generally used to measure the similarity of two discrete objects. Numanagic et al. proposed the SEDEF framework based on the Jaccard coefficient, which can accurately predict segmental duplications (SDs)[40]. Wallace et al. introduced the Jaccard coefficient into the prediction of disease-disease relationship and deduced the information of the interaction network[41]. In this paper, we compare the co-expression of two different related proteins with the Jaccard coefficient. Therefore, the Jaccard coefficient of edge (i,j) can be defined as:

$$\mathrm{Jaccard}(i,j) = \frac{S_i \cap S_j}{S_i \cup S_j} \quad (8)$$

Where S_i and S_j represent the Boolean values of the gene expression data of gene *i* and gene *j*. The Jaccard correlation coefficient should be between 0 and 1. Here, we define the value as the similarity of active expression between gene *i* and gene *j* in a cluster of PPI networks.

JDC Measure index

It has been proved that genes with similar functions often exhibit similar expression patterns, known as the "guilt-by-association" principle[42]. Based on the edge clustering coefficient (ECC) and Jaccard coefficient (Jaccard), we propose a new measurement method, which is named as the essential Degree Centrality with Jaccard similarity index (JDC). We describe the clustering degree of two proteins from topological and biological perspectives. Therefore, we define the clustering degree of an edge (i,j) in the PPI network as follows:

$$J_c(i,j) = Jaccard(i,j) * ECC(i,j) \quad (9)$$

For protein *i*, we define its JDC value as the sum of the probability that the protein and its neighbors belong to the same cluster:

$$\mathrm{JDC}(i) = \sum_{j \in D_i} Jaccard(i,j) * ECC(i,j) \quad (10)$$

Where D_i denotes all the neighborhoods of node *i*. Then, the node i and the neighbors are divided into a cluster. The values measured by JDC depend on the similarity of "active" and "inactive" state of gene expression in a cluster of PPI networks.

In this paper, we propose an essential protein identification method based on PPI data and gene expression. The advantage of this method is that the calculation is simple, and the performance of JDC is better than some state-of-the-art prediction methods.

Results and Discussion

ROC Curves and its AUC analysis

In this section, we adopt receiver operating characteristic (ROC) curves to evaluate the global performance of each method. The comparison results are shown in Figure 2.

As shown in Figure 2, the ROC curve of JDC is almost above that of other prediction methods. The area under the ROC curve (AUC) on both two datasets are 0.6996, and 0.6999, respectively, which are the highest values among all methods. The ROC results obtained by ten methods demonstrate that JDC is more suitable for predicting essential proteins.

To show that our method has better performance, we focus on comparing JDC with WDC and PeC, due to these methods with the same input data. Li and Tang have introduced the Pearson correlation coefficient to weight PPI network based on ECC, which effectively reduced false positives and false negatives in PPI network on Yeast data[12, 13]. Compared with those methods, JDC not only considers the false positive and false negative data on PPI data but also introduces the "active" and "inactive" states of gene expression. The AUC of JDC method improves more 0.0112 and 0.0665 on the yeast dataset than that of WDC and Pec, respectively. The similar results are obtained in the experimental results of E.coli dataset.

The advantage of introducing different states is to eliminate fluctuations in gene expression data, especially between two genes, a gene has a particularly high expression value and thus affects the similarity value. JDC can fully consider the co-expression state of the connected genes at multiple different moments, while WDC and Pec compare the similarity of the specific expression values of the two genes at different times.

To further compare the performance of JDC, WDC and Pec, we analyze the ROC curve based on the proteins ranked by each method at the top 20 percent. The ROC curves are shown in Figure 3. As can be seen from Figure 3, the AUC of JDC is higher than WDC and PeC both on yeast and E.coli datasets. The ROC curve for JDC can almost achieve the best on the yeast dataset, and when values of FPR are less than 0.4 on the E.coli dataset, the ROC curve for JDC also have the similar result. The results suggest that the JDC has better sensitivity than that of WDC and PeC.

Accuracy analysis

To further validate the performance of JDC method, the following criteria were used to evaluate, including sensitivity (SN), specificity (SP), false-positive rate (FPR), positive predictive value (PPV), negative predictive value (NPV), F-measure, accuracy (ACC) and Matthews correlation coefficient (MCC). Since the yeast datasets contain 5093 proteins, of which 1167 proteins are essential proteins, the top 1167 proteins ranked by each method are selected as positives. Similarly, we also take the top 254 proteins ranked by each method as positives on E.coli data.

The formula11~formula17 are as follows:

$$SN = \frac{TP}{TP + FN} \quad (11)$$

$$SP = \frac{TN}{TN + FP} \quad (12)$$

$$FPR = \frac{FP}{TN + FP} \quad (13)$$

$$PPV = \frac{TP}{TP + FP} \quad (14)$$

$$F - measure = \frac{2 * TP}{2 * TP + FP + FN} \quad (15)$$

$$ACCuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (17)$$

Where TP denotes the number of true positive proteins, FP denotes the number of false-positive proteins, TN denotes the number of true negative proteins, and FN denotes the number of false-negative proteins. In this paper, true positive is that real essential proteins are correctly predicted as essential proteins, false positive is that non-essential proteins are predicted as essential proteins, true negative is that non-essential proteins are correctly predicted as non-essential proteins, and false negative is that the essential proteins are predicted as non-essential proteins. The results on Yeast and E.coli data are in Table 1.

It can be seen from Table 1 that the values of SN, SP, PPV, NPV, $F - measure$, ACC, and MCC of JDC on Yeast data are 0.4604, 0.8403, 0.4604, 0.8403, 0.4604, 0.7535 and 0.3007 respectively. Each evaluation criterion for JDC is better than other prediction methods. Meanwhile, the values of SN, SP, PPV, NPV, $F - measure$, ACC and MCC of JDC on E.coli data are 0.2835, 0.9264, 0.2835, 0.9264, 0.2835, 0.8665 and 0.2099 respectively, which outperforms all other methods listed in Table 1. The lower the FPR is, the better the method is. The FPR value of JDC is also the lowest of all methods in the two data sets.

Jackknife Analysis

Holman et al. devised a jackknife strategy that tests the performance of ranking methods[43]. We also use this method to evaluate the JDC method and other nine essential protein prediction methods. For each prediction method, we assess the performance by calculating the sum of the true essential proteins and the number of essential proteins. Figure 3 is the jackknife curve of various methods.

The jackknife curve of ten essential protein prediction methods is plotted in Figure 6. Where the vertical axis represents the cumulative count of essential proteins, and the horizontal axis represents the predicted number of essential proteins. The jackknife curve of the JDC method is higher than that of other nine methods (DC, IC, EC, SC, BC, CC, NC, WDC, and PeC). The results from the jackknife analysis show that the performance of JDC is superior to other prediction methods in identifying essential proteins. The advantage of JDC is that it can overcome the volatility of the gene expression data.

Modularity Analysis

Hart et al. indicate that the importance of proteins is not related to themselves, but specific protein complexes[44]. Zotenko et al. further demonstrate that functional protein modules contain a large number of essential proteins[45]. To verify the conclusion, we select the top 100 proteins ranked by JDC, and constructed a small PPI network module with those proteins and their neighbor proteins. The result is shown in Figure 7. The top 100 proteins of JDC include 80 essential proteins (yellow nodes in Figure 7(a)) and 17 functional modules by Markov Cluster

procedure (MCL)[46]. For WDC, we follow a similar analysis as above, 68 essential proteins (yellow nodes in Figure 7(b)) and 14 functional modules are found. The modularity of JDC presents more obvious than that of WDC. Besides, most of the essential proteins are hubs in the network, as shown in Figure 7(a), which is consistent with views of He et al.[47].

Comparison with Dynamic Network Framework

In the previous description, we compared JDC with various essential protein prediction methods that are proposed base on the static PPI network. The experimental results show that our method can improve the accuracy of essential protein prediction. To further prove the advantage of our method, we compare it with some methods that are designed based on the dynamic PPI network. We compare JDC with both NF-PIN and TS-PIN methods. The two existing methods predict the essential protein from dynamic PPI networks constructed by using gene expression on yeast data. The results are shown in Tables 3 and Table 4.

The methods with dynamic PPI network can effectively improve the accuracy of the identification of essential proteins in DC, EC, SC, BC, CC, IC, LAC, and NC. As shown in Table 3, when the top100, top200, top300, top400, top500, and top600 proteins are selected, JDC can identify 80, 153, 224, 267, 315, and 355 essential proteins, respectively. As can be seen from Tables 3, our method is better than that of other prediction methods at the top 200, top 300, top 400, top 500, and top 600. When compared with the TS-PIN, which incorporated subcellular localization information, our method has also similar results. As shown in both Table 3 and Table 4, the exceed times of our method are 5 and 5 respectively, which indicate the JDC method is an effective prediction method for essential proteins.

Conclusions

In this study, we proposed a new essential protein recognition algorithm JDC based on the PPI networks and gene expression data. JDC eliminates the influence of fluctuations in gene expression data by calculating the similarity of "active" and "inactive" state of gene expression in a cluster of the PPI network. Compared with the nine prediction methods with static PPI network and two prediction methods with dynamic methods, JDC is an effective essential protein prediction method.

Acknowledgements

Not applicable.

Availability of data and materials

The data and source codes are available in https://github.com/jczhongcs/JDC

Authors' contributions

JC initialized this study. TC, TQ, XQ, JH and JC discussed many times to finalized the work plan. TC and SY conducted the majority of numerical experiments. WP and MZ gave suggestions many times to modify this study. TC drafted the manuscript. Everyone read the manuscript and revised it, and agreed with the final version.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's Affiliations

[1] School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China.

[2] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, China

[3] College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, P.R. China.

[4] College of engineering and design, Hunan Normal University, Changsha 410081, China.

Corresponding author: Qiu Xiao, Jiahong Yang

Funding

This work is supported in part by the Natural Science Foundation of Hunan Province of China (No.2018JJ2262) , Hunan Provincial Science and Technology Program (No. 2018WK4001), and the Scientific Research Fund of Hunan Provincial Education Department (No.15CY007, 19A316).

References

1. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson KR, Andre B, Bangham R, Benito R, Boeke JD, Bussey H: Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. Science 1999, 285(5429):901-906.
2. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Bot NL, Moreno S, Sohrmann M: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 2003, 421(6920):231-237.
3. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucaudanila A, Anderson KR, Andre B: Functional profiling of the Saccharomyces cerevisiae genome. Nature 2002, 418(6896):387-391.
4. Cullen LM, Arndt GM: Genome-wide screening for gene function using RNAi in mammalian cells. Immunology and Cell Biology 2005, 83(3):217-223.
5. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C: Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Molecular Microbiology 2003, 50(1):167-181.
6. Hahn MW, Kern AD: Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. Molecular Biology and Evolution 2005, 22(4):803-806.
7. Joy MP, Brock A, Ingber DE, Huang S: High-Betweenness Proteins in the Yeast Protein Interaction Network. BioMed Research International 2005, 2005(2):96-103.
8. Wuchty S, Stadler PF: Centers of complex networks. Journal of Theoretical Biology 2003, 223(1):45-53.
9. Estrada E, Rodriguezvelazquez JA: Subgraph centrality in complex networks. Physical Review E 2005, 71(5):056103-056103.
10. Bonacich P: Power and Centrality: A Family of Measures. American Journal of Sociology 1987, 92(5):1170-1182.
11. Stephenson K, Zelen M: Rethinking centrality: Methods and examples☆. Social Networks 1989, 11(1):1-37.
12. Li M, Zhang H, Wang J, Pan Y: A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Systems Biology 2012, 6(1):15-15.
13. Tang X, Wang J, Pan Y: Identifying essential proteins via integration of protein interaction and gene expression data. In: bioinformatics and biomedicine: 2012. 1-4.
14. Peng WH, Wang J, Wang W, Liu Q, Wu F, Pan Y: Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Systems Biology 2012, 6(1):87-87.
15. Li G, Li M, Wang J, Li Y, Pan Y: United neighborhood closeness centrality and orthology for predicting essential proteins. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2018:1-1.
16. De Lichtenberg U, Jensen LJ, Brunak S, Bork P: Dynamic complex formation during the yeast cell cycle. Science 2005, 307(5710):724-727.
17. Xiao Q, Wang J, Peng X, Wu F, Pan Y: Identifying essential proteins from active PPI networks constructed with dynamic gene expression. BMC Genomics 2015, 16(3):1-7.
18. Li M, Ni P, Chen X, Wang J, Wu F, Pan Y: Construction of Refined Protein Interaction Network for Predicting Essential Proteins. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019, 16(4):1386-1397.
19. Li M, Li W, Wu F, Pan Y, Wang J: Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. Journal of Theoretical Biology 2018, 447:65-73.
20. Fan Y, Tang X, Hu X, Wu W, Ping Q: Prediction of essential proteins based on subcellular localization and gene expression correlation. BMC Bioinformatics 2017, 18(13):470-470.
21. Lei X, Yang X, Fujita H: Random walk based method to identify essential proteins by integrating network topology and biological characteristics. Knowledge Based Systems 2019,

167:53-67.
22. Zhang F, Peng W, Yang Y, Dai W, Song J: A Novel Method for Identifying Essential Genes by Fusing Dynamic Protein–Protein Interactive Networks. Genes 2019, 10(1):31.
23. Li M, Lu Y, Wang J, Wu F, Pan Y: A topology potential-based method for identifying essential proteins from PPI networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2015, 12(2):372-383.
24. Peng W, Wang J, Cheng Y, Lu Y, Wu F, Pan Y: UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2015, 12(2):276-288.
25. Gustafson AM, Snitkin ES, Parker SCJ, Delisi C, Kasif S: Towards the identification of essential genes using targeted genome sequencing and comparative analysis. BMC Genomics 2006, 7(1):265-265.
26. Hwang Y, Lin C, Chang J, Mori H, Juan H, Huang HC: Predicting essential genes based on network and sequence analysis. Molecular BioSystems 2009, 5(12):1672-1678.
27. Zhong J, Wang J, Peng W, Zhang Z, Pan Y: Prediction of essential proteins based on gene expression programming. BMC Genomics 2013, 14(4):1-8.
28. Zhong J, Sun Y, Peng W, Xie M, Yang J, Tang X: XGBFEMF: An XGBoost-Based Framework for Essential Protein Prediction. IEEE Transactions on Nanobioscience 2018, 17(3):243-250.
29. Deng J, Deng L, Su S, Zhang M, Lin X, Wei L, Minai AA, Hassett DJ, Lu LJ: Investigating the predictability of essential genes across distantly related organisms using an integrative approach. Nucleic Acids Research 2011, 39(3):795-807.
30. Kim W: Prediction of essential proteins using topological properties in GO-pruned PPI network based on machine learning methods. Tsinghua Science & Technology 2012, 17(6):645-658.
31. Zeng M, Li M, Fei Z, Wu F, Li Y, Pan Y, Wang J: A deep learning framework for identifying essential proteins by integrating multiple types of biological information. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019:1-1.
32. Niehrs C, Pollet N: Synexpression groups in eukaryotes. Nature 1999, 402(6761):483-487.
33. Mewes H, Frishman D, Mayer KFX, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Research 2006, 34(90001):169-172.
34. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Isseltarver L, Schroeder M, Sherlock G: Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Research 2002, 30(1):69-72.
35. Zhang R, Lin Y: DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Research 2009, 37:455-458.
36. Giaever G, Nislow C: The Yeast Deletion Collection: A Decade of Functional Genomics. Genetics 2014, 197(2):451-465.
37. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 2004, 101(9):2658-2663.
38. Wang J, Li M, Wang H, Pan Y: Identification of Essential Proteins Based on Edge Clustering Coefficient. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2012, 9(4):1070-1080.
39. Sahoo D: Boolean analysis of high-throughput biological datasets. In: 2008.
40. Numanagic I, Gokkaya AS, Zhang L, Berger B, Alkan C, Hach F: Fast characterization of segmental duplications in genome assemblies. Bioinformatics 2018, 34(17).
41. Wallace ZS, Rosenthal SB, Fisch KM, Ideker T, Sasik R: On entropy and information in gene interaction networks. Bioinformatics 2019, 35(5):815-822.
42. Wolfe CJ, Kohane IS, Butte AJ: Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 2005, 6(1):227-227.
43. Holman AG, Davis PJ, Foster JM, Carlow CKS, Kumar S: Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi. BMC Microbiology 2009, 9(1):243-243.
44. Hart GT, Lee I, Marcotte EM: A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 2007, 8(1):236-236.
45. Zotenko E, Mestre J, Oleary DP, Przytycka TM: 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 2008, 4(8).
46. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002, 30(7):1575-1584.
47. He X, Zhang J: Why Do Hubs Tend to Be Essential in Protein Networks. PLOS Genetics 2005, 2(6).

Figures

Figure 1. An illustration of JDC

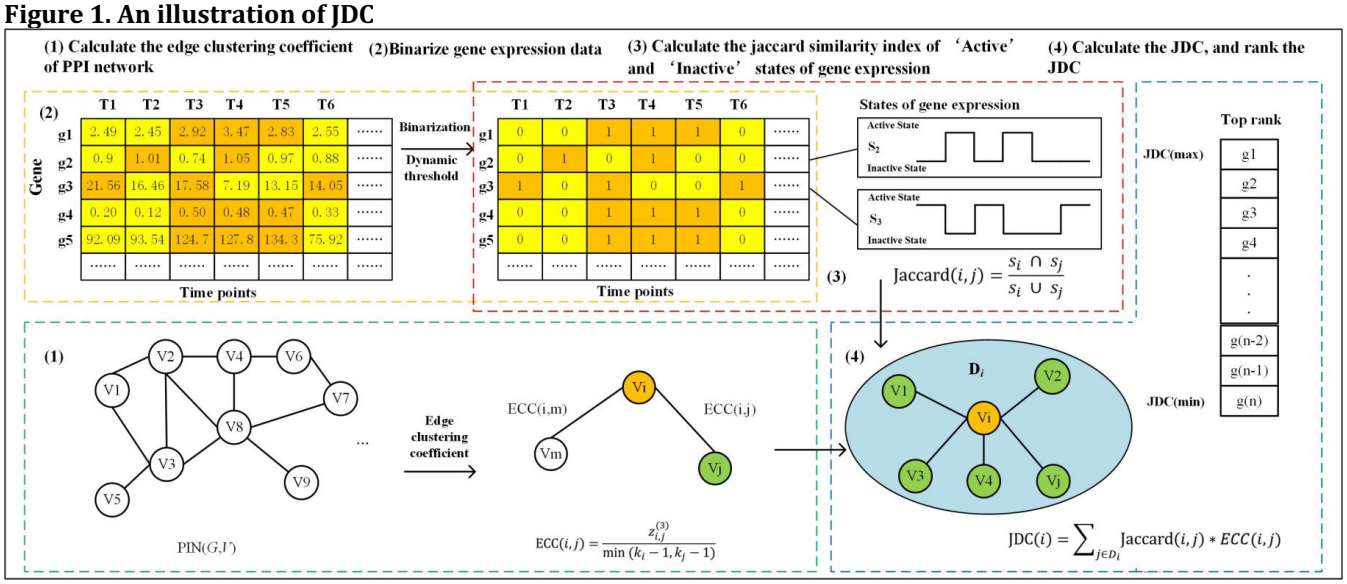
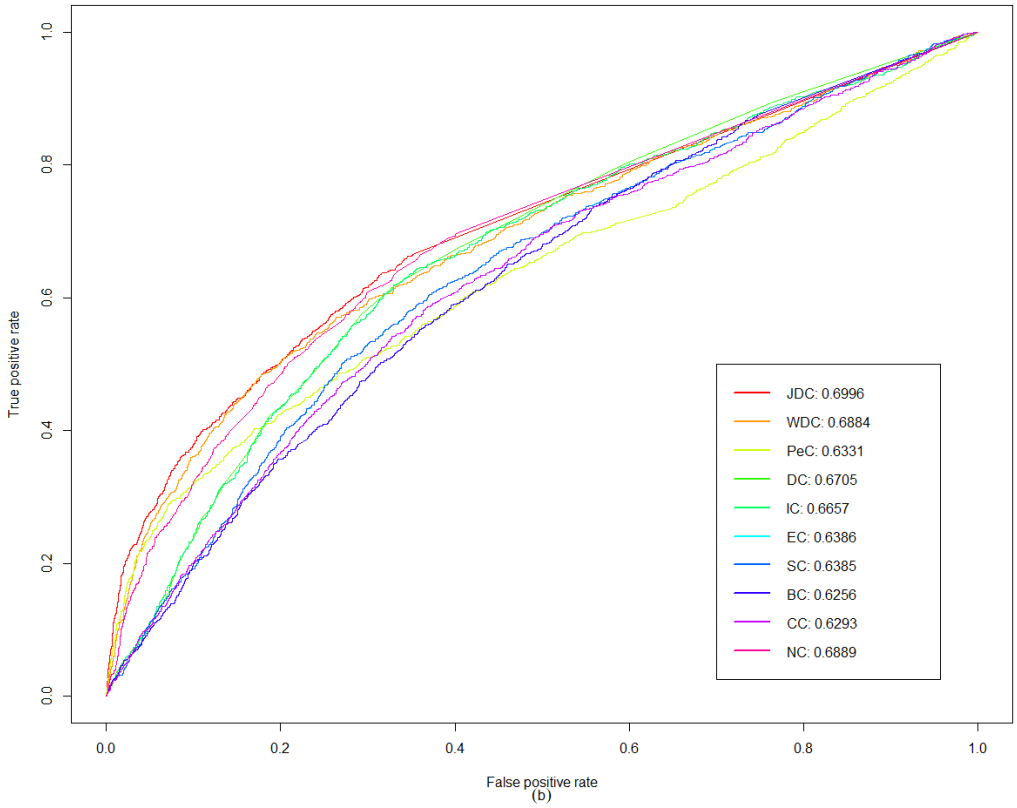


Figure 2. ROC curves and AUC values of the JDC method and other methods using the individual features. (a)Yeast data. (b)E.coli data.



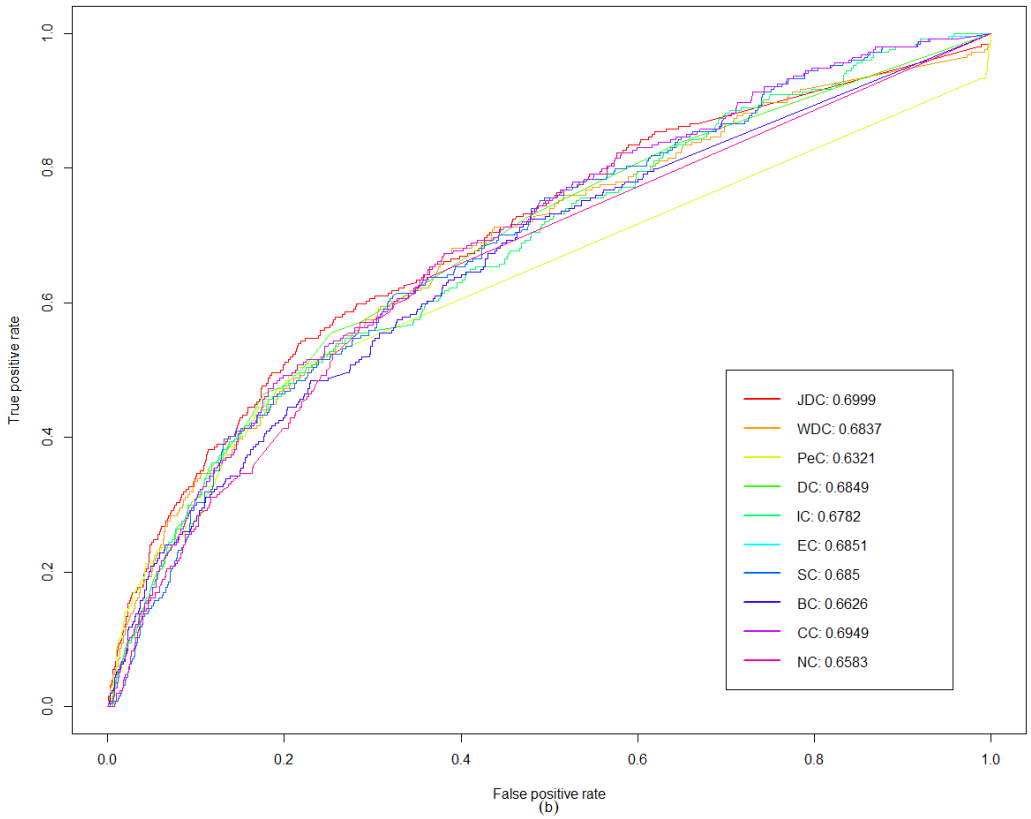
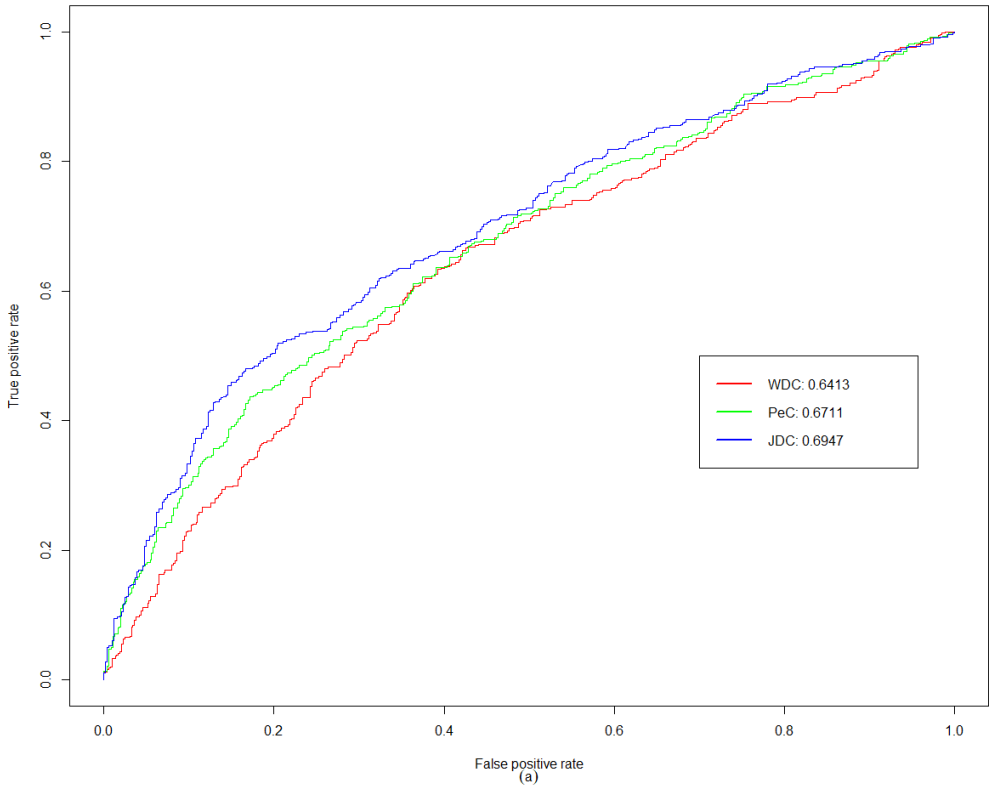


Figure 3. ROC curves and AUC values of the JDC method and other methods using the individual features in the top 20% ranked proteins. (a)Yeast data. (b)E.coli data.



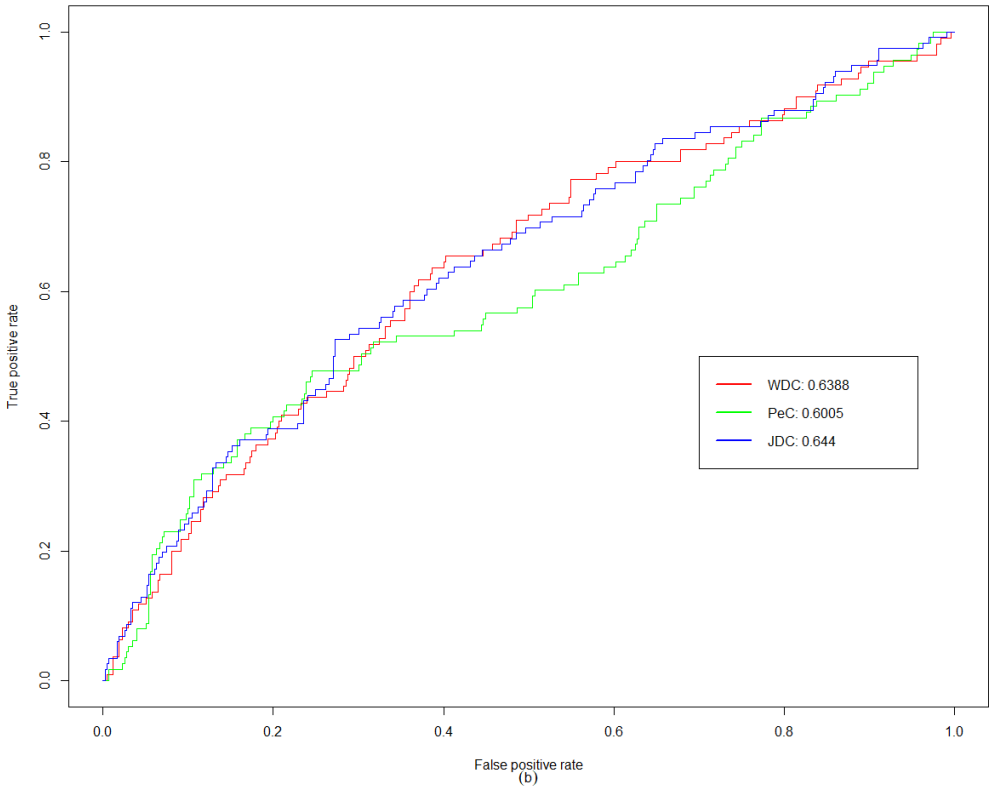
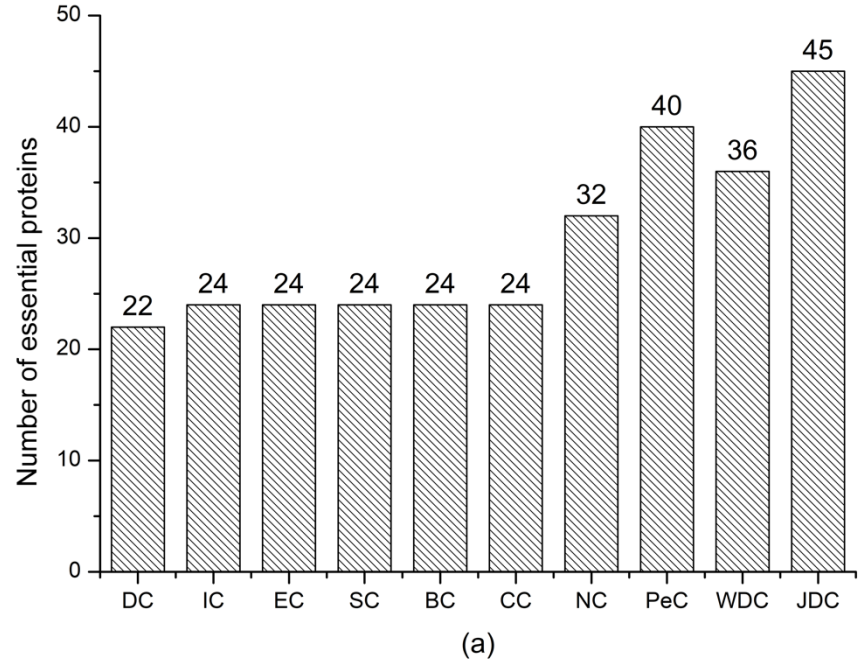
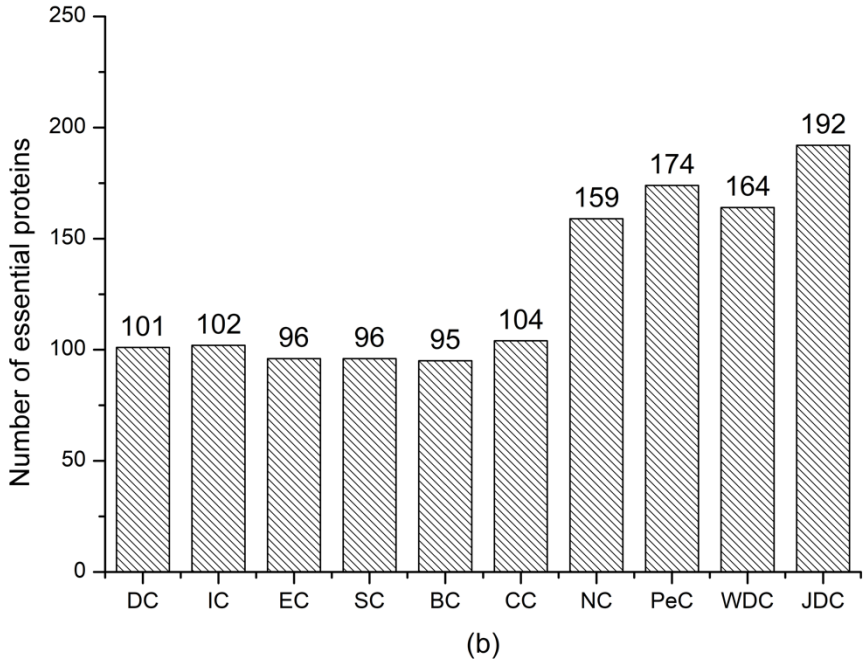


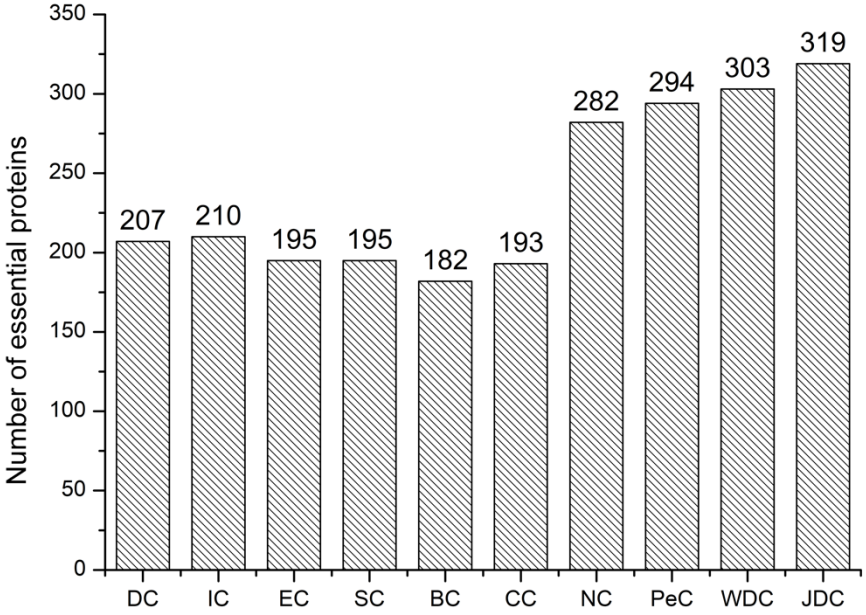
Figure 4. Compares the top 1%, 5%, 10%, 15%, 20% and 25% of essential proteins obtained by JDC with other methods in yeast data. (a)TOP1%. (b)Top5%. (c)Top10%. (d)Top15%. (e)Top20%. (f)Top25%.



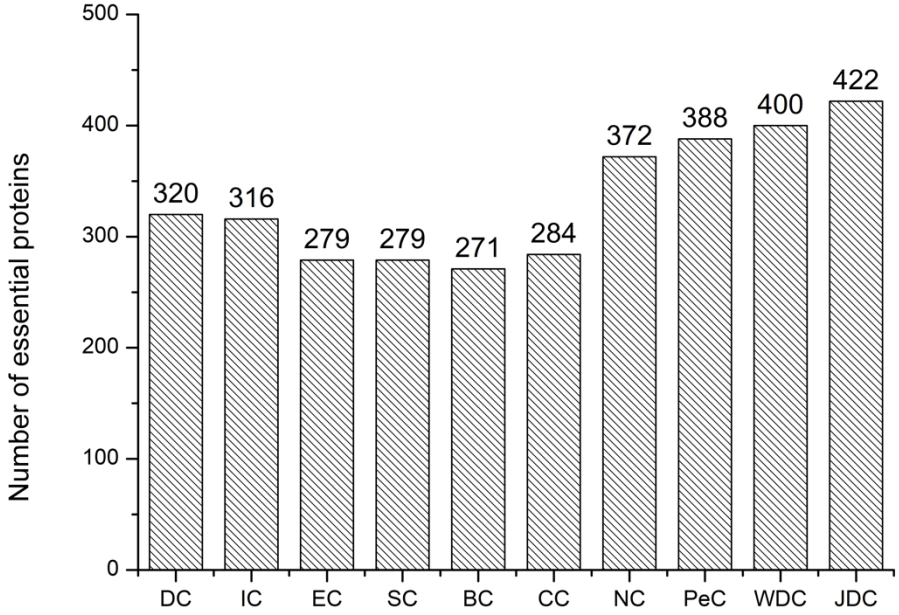
(a)



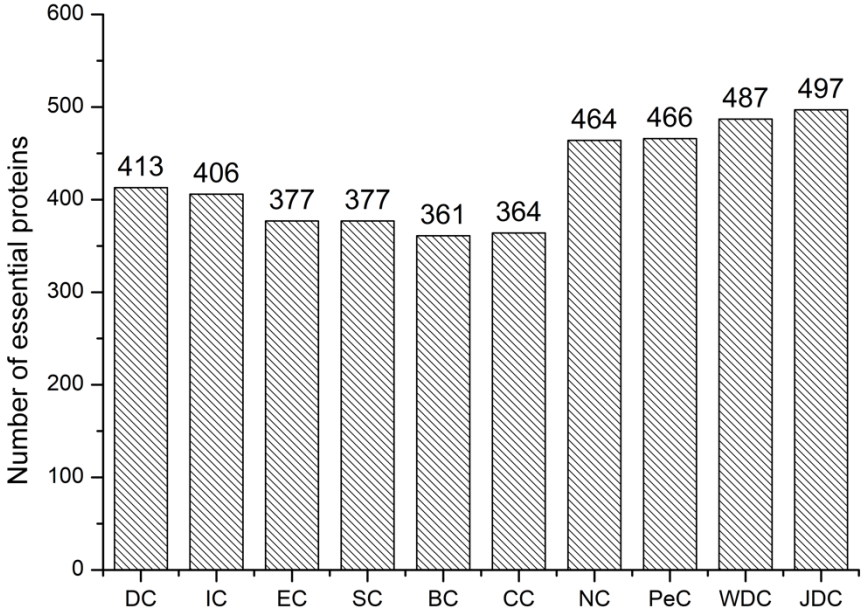
(b)


Number of essential proteins
0
50
100
150
200
250
300
350
207
210
195
195
182
193
282
294
303
319
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

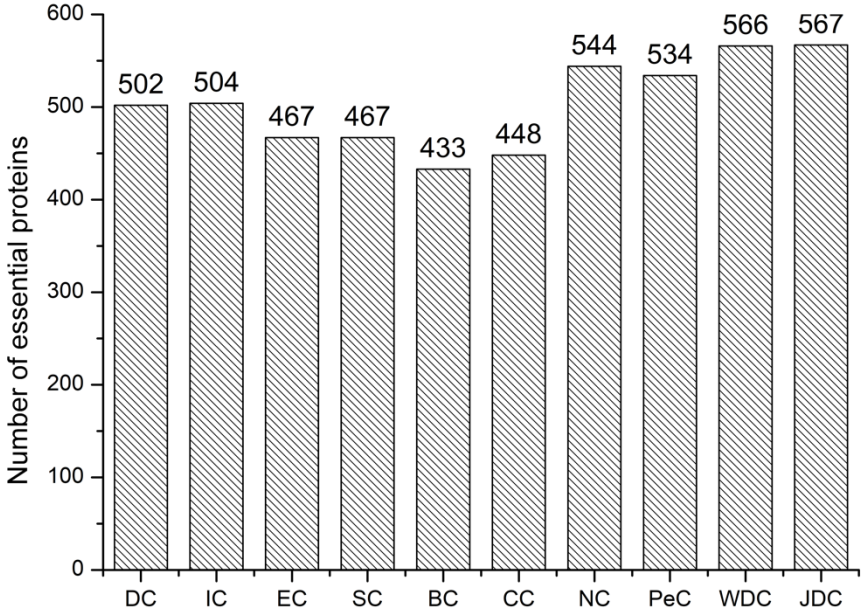
(c)


Number of essential proteins
0
100
200
300
400
500
320
316
279
279
271
284
372
388
400
422
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(d)

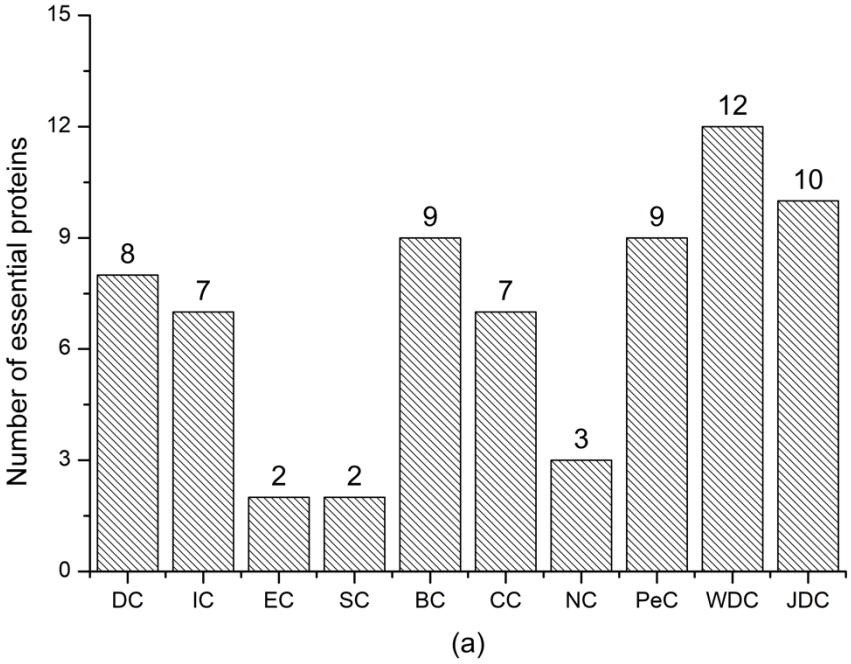

Number of essential proteins
600
500
400
300
200
100
0
413
406
377
377
361
364
464
466
487
497
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(e)

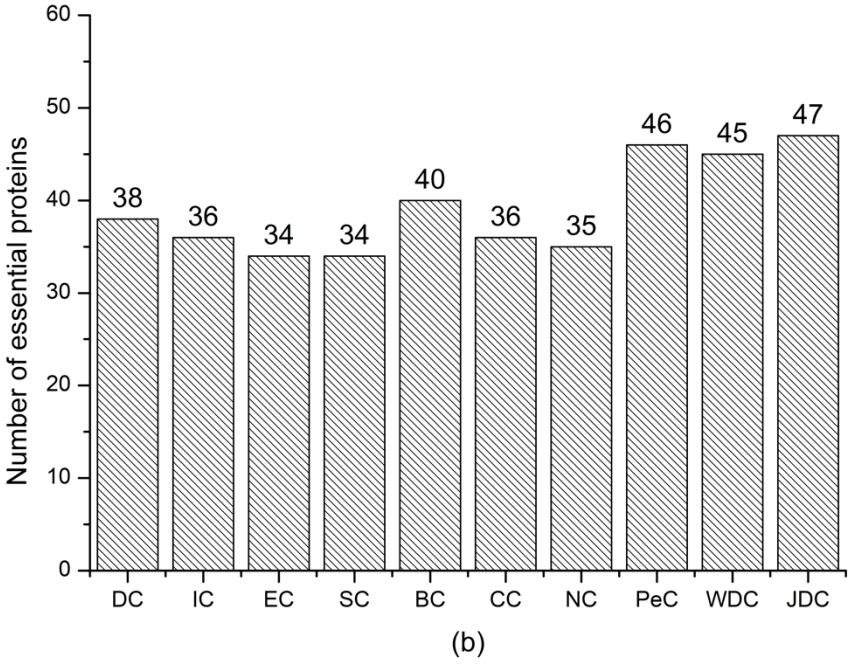

Number of essential proteins
600
500
400
300
200
100
0
502
504
467
467
433
448
544
534
566
567
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(f)

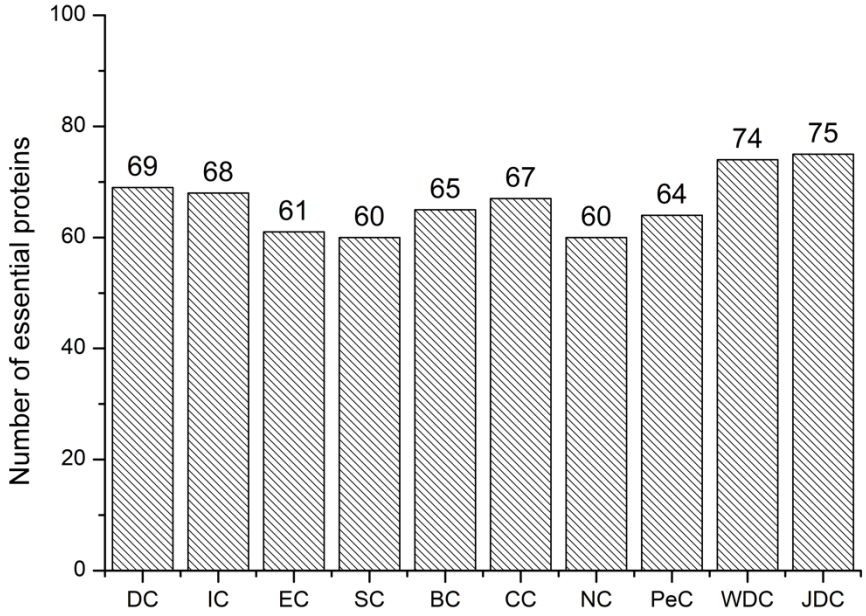
Figure 5. Compares the top 1%, 5%, 10%, 15%, 20% and 25% of essential proteins obtained by JDC with other methods in E.coli data. (a)TOP1%. (b)Top5%. (c)Top10%. (d)Top15%. (e)Top20%. (f)Top25%.



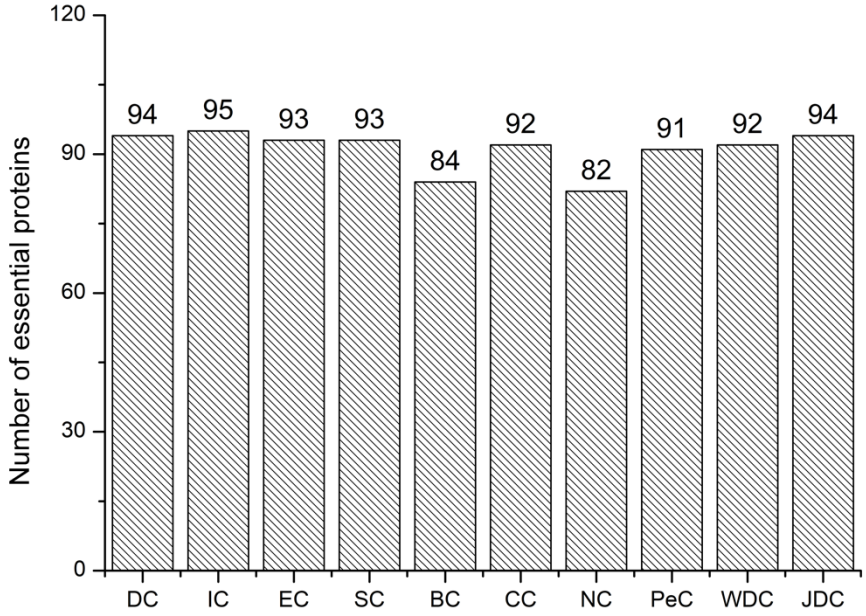
(a)



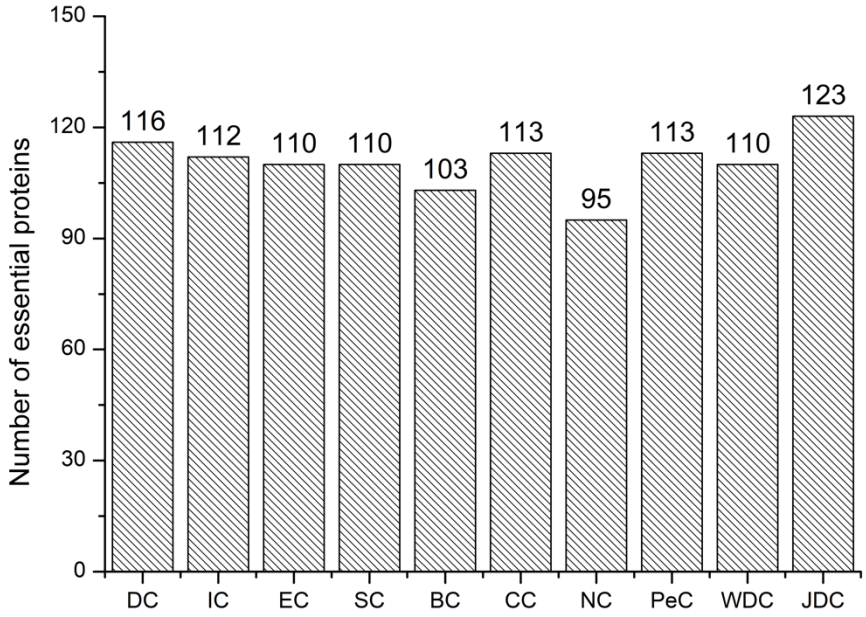
(b)


Number of essential proteins
100
80
60
40
20
0
69
68
61
60
65
67
60
64
74
75
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

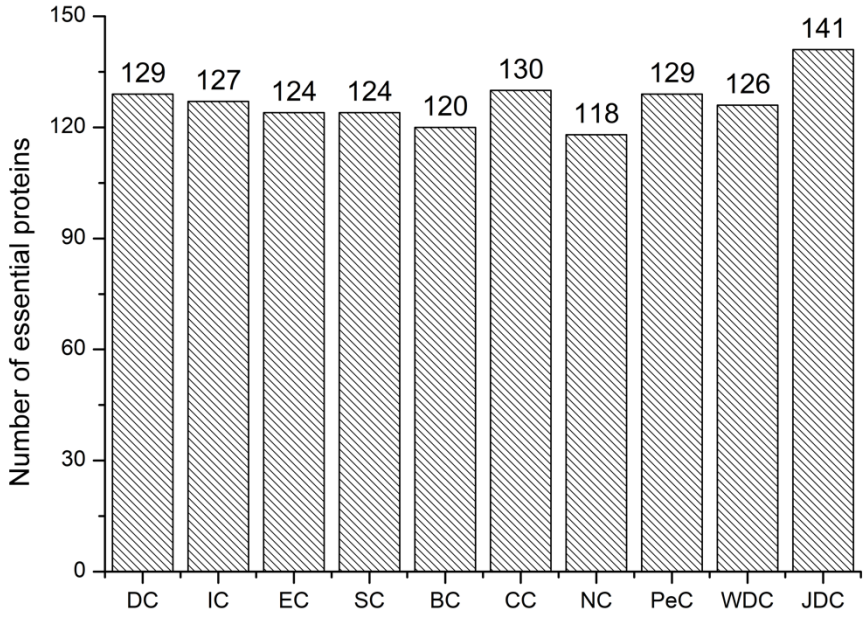
(c)


Number of essential proteins
120
90
60
30
0
94
95
93
93
84
92
82
91
92
94
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(d)

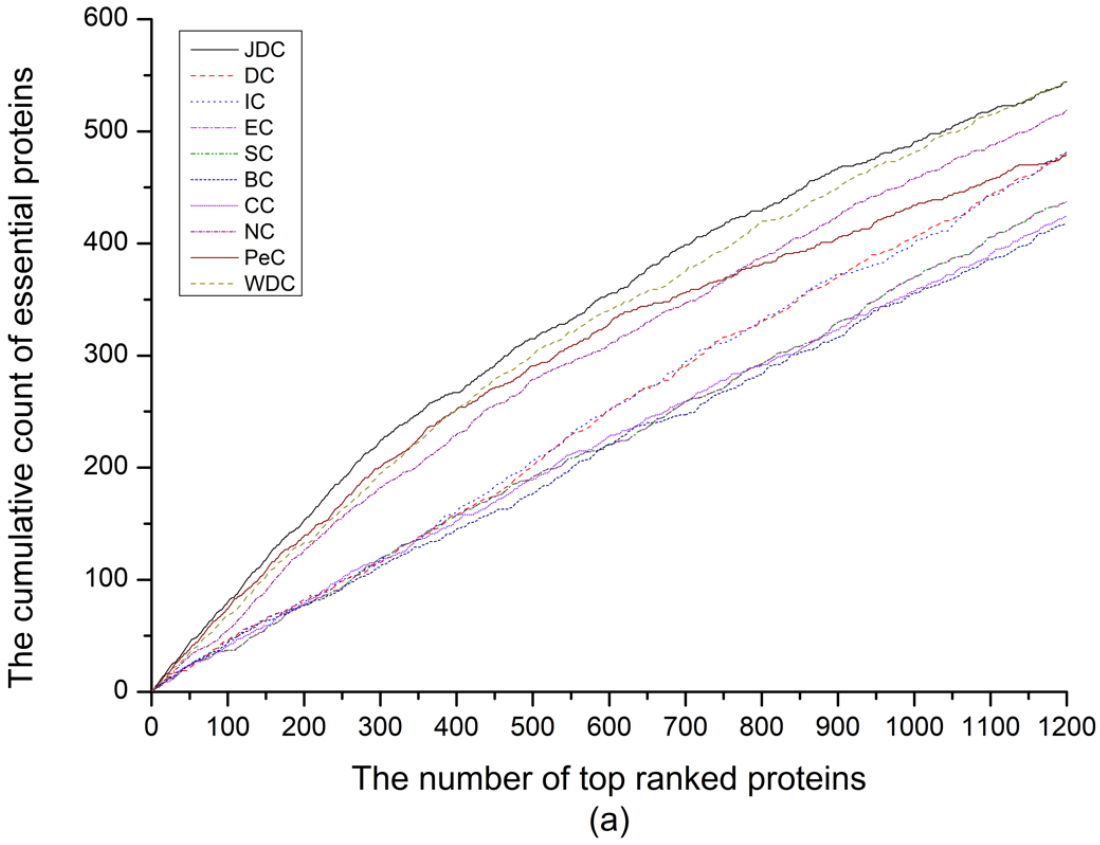

Number of essential proteins
150
120
90
60
30
0
116
112
110
110
103
113
95
113
110
123
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(e)

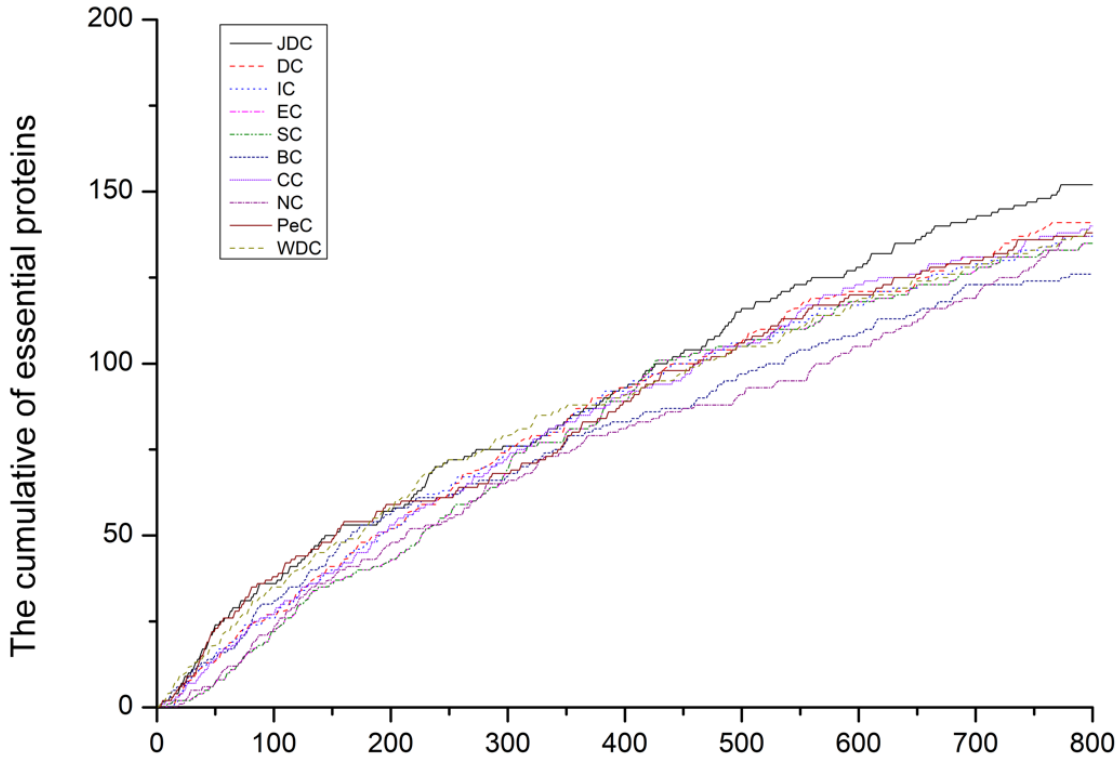

Number of essential proteins
150
120
90
60
30
0
129
127
124
124
120
130
118
129
126
141
DC
IC
EC
SC
BC
CC
NC
PeC
WDC
JDC

(f)

Figure 6. Jackknife curve of various prediction methods. (a) Yeast data. (b) E.coli data

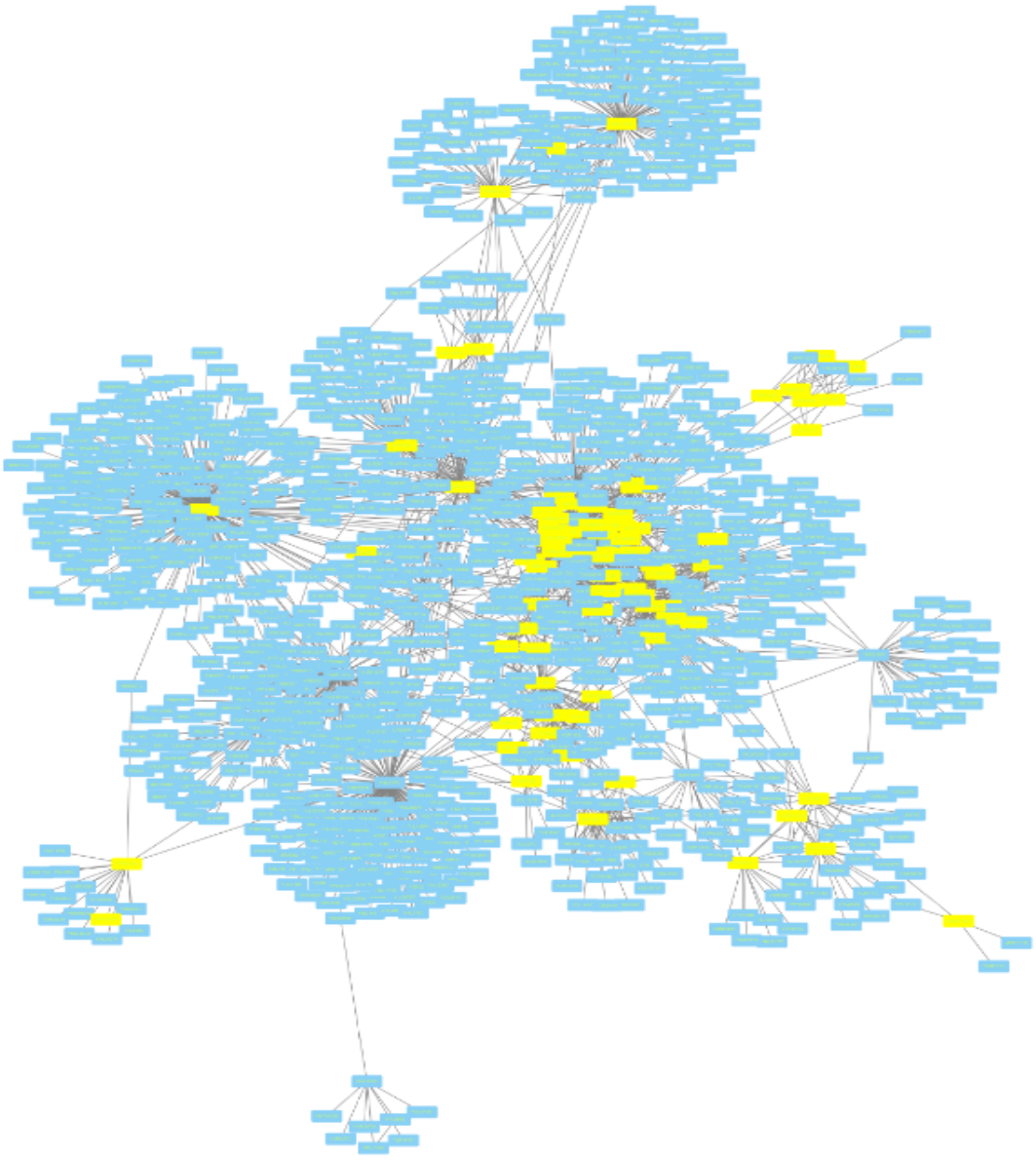


(a)

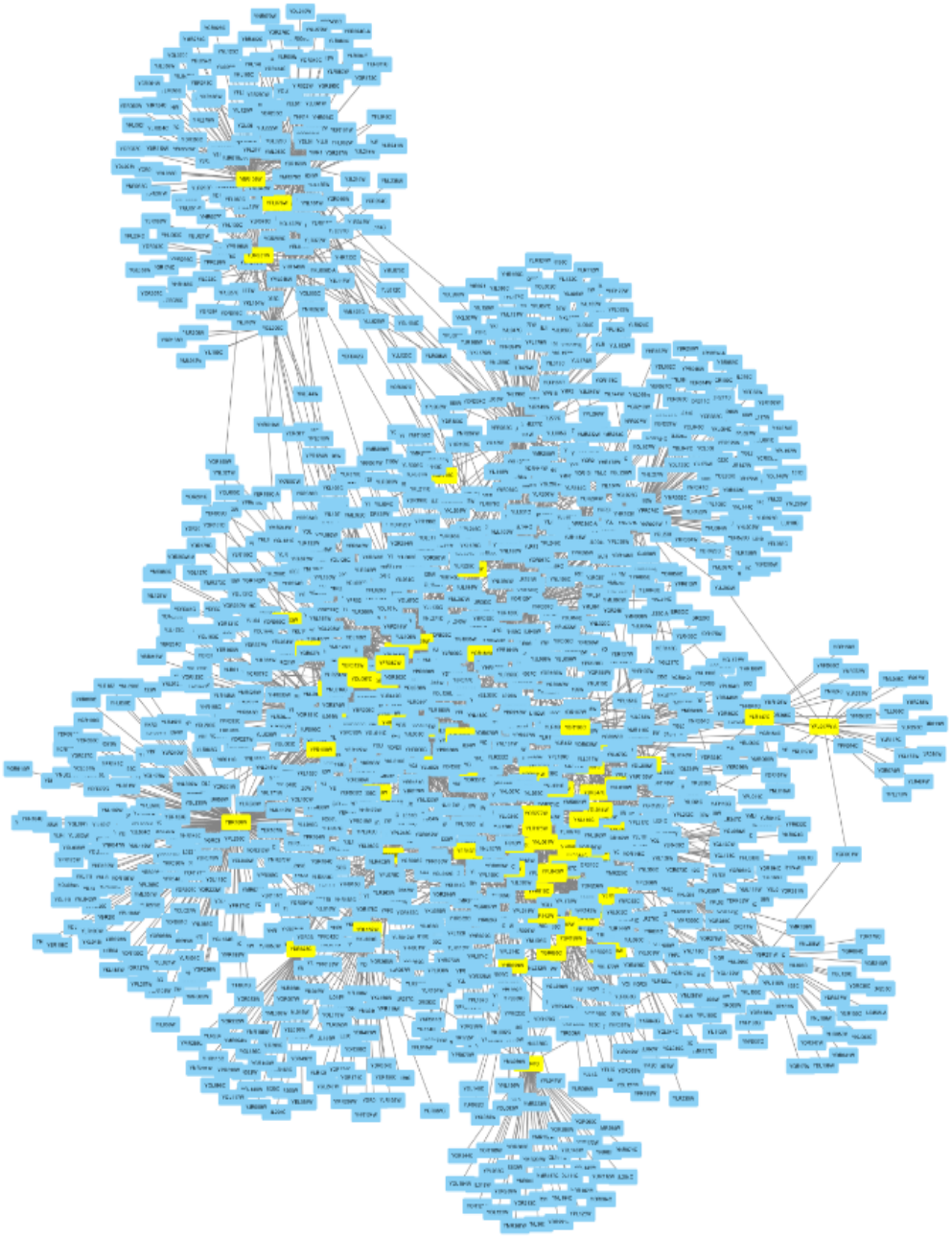


(b)

Figure 7. The modularity of interactions among the top 100 essential proteins predicted by JDC and WDC



(a) JDC



(b) WDC

Tables

Table 1. SN, SP, FPR, PPV, NPV, F-measure, ACC and MCC of Various Methods on Total Ranked Proteins

Yeast data								
Methods	**SN**	**SP**	**FPR**	**PPV**	**NPV**	**F-measure**	**ACC**	**MCC**
JDC	**0.4604**	**0.8403**	**0.1597**	**0.4604**	**0.8403**	**0.4604**	**0.7535**	**0.3007**
DC	0.4002	0.8217	0.1783	0.4002	0.8217	0.4002	0.7251	0.2219
BC	0.3505	0.8069	0.1931	0.3505	0.8069	0.3505	0.7023	0.1574
CC	0.3548	0.8082	0.1918	0.3548	0.8082	0.3548	0.7043	0.163
SC	0.3676	0.812	0.188	0.3676	0.812	0.3676	0.7102	0.1796
EC	0.3676	0.812	0.188	0.3676	0.812	0.3676	0.7102	0.1796
IC	0.401	0.822	0.178	0.401	0.822	0.401	0.7255	0.223
NC	0.4353	0.8321	0.1679	0.4353	0.8321	0.4353	0.7412	0.2674
PeC	0.4036	0.8227	0.1773	0.4036	0.8227	0.4036	0.7267	0.2263
WDC	0.4576	0.839	0.161	0.458	0.8388	0.4578	0.7516	0.2967

E.coli data								
Methods	**SN**	**SP**	**FPR**	**PPV**	**NPV**	**F-measure**	**ACC**	**MCC**
JDC	**0.2835**	**0.9264**	**0.0736**	**0.2835**	**0.9264**	**02835**	**0.8665**	**02099**
DC	0.2559	0.9236	0.0764	0.2559	0.9236	0.2599	0.8614	0.1795
BC	0.2441	0.9224	0.0776	0.2441	0.9224	0.2441	0.8592	0.2665
CC	0.2441	0.9224	0.0776	0.2441	0.9224	0.2441	0.8592	0.1665
SC	0.2283	0.9207	0.0793	0.2283	0.9207	0.2283	0.8562	0.1491
EC	0.2283	0.9207	0.0793	0.2283	0.9207	0.2283	0.8562	0.1491
IC	0.2559	0.9236	0.0764	0.2559	0.9236	0.2559	0.8614	0.1795
NC	0.2165	0.9195	0.0805	0.2165	0.9195	0.2165	0.8541	0.1361
PeC	0.2441	0.9204	0.0776	0.2441	0.9224	0.2441	0.8592	0.1665
WDC	0.2689	0.922	0.078	0.2689	0.922	0.2689	0.859	0.1909

Table 2. The overlapping relationships between JDC and nine other prediction measures for the top 100 proteins

Centrality	$JDC \cap C_i$	Non-essential proteins of C_i in $\lvert C_i - JDC \rvert$	Non-essential proteins of JDC in $\lvert C_i - JDC \rvert$	Percentage of essential proteins of C_i in $\lvert C_i - JDC \rvert$	Percentage of essential proteins of JDC in $\lvert C_i - JDC \rvert$
DC	16	46	15	45.24%	82.14%
IC	17	46	18	44.58%	78.31%
EC	8	61	18	33.70%	80.43%
SC	8	61	18	33.70%	80.43%
BC	15	49	18	42.35%	78.82%
CC	13	52	17	40.23%	80.46%
NC	36	34	14	46.88%	78.13%
PeC	67	12	8	63.64%	75.76%
WDC	55	20	12	55.56%	73.33%

Table 3. Accurate analysis of the number of essential proteins predicted by various central methods in the dynamic network of NF-PIN with JDC

Centrality	Top100	Top200	Top300	Top400	Top500	T600	Exceed times
JDC	80	**153**	**224**	**267**	**315**	**355**	**5**
NF-DC	55	111	167	221	261	303	0
NF-EC	55	110	157	202	239	276	0
NF-SC	55	116	161	204	239	276	0
NF-BC	50	97	133	188	226	254	0
NF-CC	45	87	122	161	193	230	0
NF-IC	55	111	167	221	261	303	0
NF-LAC	**82**	141	198	243	280	322	1
NF-NC	80	147	197	252	290	324	0

Table 4. Accurate analysis of the number of essential proteins predicted by various central methods in the dynamic network of TS-PIN with JDC

Centrality	Top100	Top200	Top300	Top400	Top500	T600	Exceed times
JDC	80	**153**	**224**	**267**	**315**	**355**	**5**
TS-DC	71	143	198	250	297	347	0
TS-EC	71	143	209	259	300	334	0
TS-SC	78	144	210	266	308	351	0
TS-BC	55	117	165	215	252	287	0
TS-CC	55	114	173	221	273	326	0
TS-IC	71	143	198	247	297	347	0
TS-LAC	**85**	138	196	246	300	350	1
TS-NC	82	142	200	253	301	350	0