BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 06/10 pp185–200
Volume 6, Number 2, June 2023
DOI: 10.26599/BDMA.2022.9020019

Continuous and Discrete Similarity Coefficient for Identifying Essential Proteins Using Gene Expression Data

Jiancheng Zhong*, Zuohang Qu, Ying Zhong, Chao Tang, and Yi Pan*

Abstract: Essential proteins play a vital role in biological processes, and the combination of gene expression profiles with Protein-Protein Interaction (PPI) networks can improve the identification of essential proteins. However, gene expression data are prone to significant fluctuations due to noise interference in topological networks. In this work, we discretized gene expression data and used the discrete similarities of the gene expression spectrum to eliminate noise fluctuation. We then proposed the Pearson Jaccard coefficient (PJC) that consisted of continuous and discrete similarities in the gene expression data. Using the graph theory as the basis, we fused the newly proposed similarity coefficient with the existing network topology prediction algorithm at each protein node to recognize essential proteins. This strategy exhibited a high recognition rate and good specificity. We validated the new similarity coefficient PJC on PPI datasets of Krogan, Gavin, and DIP of yeast species and evaluated the results by receiver operating characteristic analysis, jackknife analysis, top analysis, and accuracy analysis. Compared with that of node-based network topology centrality and fusion biological information centrality methods, the new similarity coefficient PJC showed a significantly improved prediction performance for essential proteins in DC, IC, Eigenvector centrality, subgraph centrality, betweenness centrality, closeness centrality, NC, PeC, and WDC. We also compared the PJC coefficient with other methods using the NF-PIN algorithm, which predicts proteins by constructing active PPI networks through dynamic gene expression. The experimental results proved that our newly proposed similarity coefficient PJC has superior advantages in predicting essential proteins.

Key words: Protein-Protein Interaction (PPI) network; continuous and discrete similarity coefficient; essential proteins

1 Introduction

Proteins are often deeply involved and play an irreplaceable role in biological processes[1]. Proteins are usually classified as essential and non-essential, the essential one generally exists in the complexes; the former generally exists in complexes, and their loss is prone to abnormal life activities and even the extinction of organisms[2]. Identifying essential proteins helps us understand the nature of cell life and discover human disease genes. Essential proteins play a decisive role in cell development and are closely related to the life activities of organisms. Their deficiency may be the fundamental factor of an organism's disease. To some extent, this condition can directly affect the vital functions of some cells, leading to some diseases and eventually promoting function loss or even death of the organism[3]. The prediction of essential proteins provides a further guarantee for proteomics and medical research in biological aspects[4].

• Jiancheng Zhong, Zuohang Qu, Ying Zhong, and Chao Tang are with the College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China. E-mail: jczhongcs@gmail.com; quzuohang@gmail.com; 1528450243@qq.com; 2692760250@qq.com.
• Yi Pan is with the Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences Shenzhen, Guangzhou 518055, China. E-mail: yi.pan@siat.ac.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-06-27; accepted: 2022-08-13

In the early stage of studying essential proteins, biologists mainly observed the influences of the loss of some proteins on the activity characteristics of organisms and then utilized single-gene knockout[5], RNA interference[6], conditional knockout[7], and other methods in biological experiments to judge whether these proteins are essential. Although these methods are effective, they have certain limitations, such as extended time and high cost. Some researchers use computational thinking to solve such problems. With the rapid development of high-throughput proteome technology and the continuous improvement of experimental biological data, computational thinking has become a method to identify essential proteins. Jeong et al.[8] proposed the lethality and centrality rule, which states that hub points are nodes with many degrees or those with many adjacent proteins in the network structure. Hub points are usually located in the center of the network and greatly influence the topology of the entire network. The loss of hub points (essential proteins) may be devastating to the whole network, suggesting its massive negative effect on biological activities. With the improvement of the lethality and centrality rule and the protein-protein interaction data, several centrality measures for node topologies based on protein networks have been derived. Degree Centrality (DC) refers to the degree of nodes in the network. This method is feasible and straightforward, but the predicted number of essential proteins is poor[9]. Betweenness Centrality (BC) refers to the number of shortest paths between a node and others and reflects the density of node positions; however, the calculation cost is high[10]. Closeness Centrality (CC) surveys the dependence of nodes on the information propagation of other nodes, but this method primarily depends on the network's topology[11]. Subgraph Centrality (SC) indicates the significance of a node according to the number of closed loops it forms with other nodes in the network[12]. Eigenvector Centrality (EC) measures the importance of a protein by the components of each vertex in the principal vector of the network adjacency matrix[13]. Information Centrality (IC) uses the average sum of the paths that each node passes through as the starting point to measure the essentiality of each protein node[14]. Although these measures consider the topological properties of the PPI network, they ignore some possible false negative and false positive data in PPIs, thus affecting the prediction of essential proteins. Some researchers have combined biological information to eliminate the impact of false positive data on PPI networks. Li et al.[15] and Tang et al.[16] proposed basic protein prediction methods called PeC and Weighted Degree Centratily (WDC) that combine PPI network and gene expression information. Compared with non-essential proteins, essential proteins tend to be conserved. Based on this observation, Peng et al.[17] utilized homologous information and PPI networks to predict essential proteins. Li et al.[18] used the Extended Pareto Optimality Consensus model to find the triangular structure in the PPI network and fused orthogonal information to predict essential proteins. Li et al.[19] transferred original PPI networks into weighted PPI networks by implementing the Pearson Correlation Coefficient (PCC) and combined the information on orthologous proteins, some critical network topological features, and protein functional features to predict essential proteins. Zhu et al.[20] proposed a novel iterative method to identify potential essential proteins according to topological features, gene expression data, subcellular localization, and homologous information. Zhao et al.[21] constructed a diffusion distance network to predict essential proteins by combining PPI topology characteristics with orthologous proteins and subcellular localization information of proteins. Some researchers recognized essential proteins by fusing the time series data of gene expression, constructing a protein dynamic network according to the dynamic characteristics of gene expression, and depicting the protein interaction at different times. Lichtenberg et al.[22] constructed a time series dynamic network by combining PPI interactions and gene expression data at different time points. Xiao et al.[23] proposed a time series model based on a static PPI network and constructed NF-PIN dynamic network using the three-sigma principle to eliminate gene expression noise. Li et al.[24] constructed a TS-PIN dynamic network to predict essential proteins by combining gene expression profiles and subcellular localization information. Zhang et al.[25] proposed a novel method for the identification of essential proteins by fusing the dynamic PPI networks of different time points.

The PPI network contains false positive and false negative data for protein interactions, which increases computational complexity and deteriorates the performance of existing basic protein prediction methods. Introducing gene expression data into PPI networks can solve this problem to some extent. However, the gene expression data applied in existing

methods are prone to fluctuations, reducing the accuracy of essential protein identification. In this work, the gene expression data were discretized, and the discrete similarity of gene expression profiles in the PPI network was applied to eliminate the fluctuation. A coefficient named Pearson Jaccard Coefficient (PJC) consists of continuous and discrete similarities in gene expression profiles. On the basis of the above analysis, the newly proposed similarity coefficient PJC was combined with node-based network topology centrality to improve the number of essential proteins. Experimental results proved that the proposed new similarity coefficient PJC can greatly help and improve the prediction performance of essential proteins by DC, CC, SC, IC, BC, EC, NC, PeC, and WDC methods.

2 Method

Owing to the use of different sources and instruments in biological experiments, a weak correlation between algorithms based on biometric information and protein network characteristics is found among various essential protein recognition methods. However, both algorithms aim to analyze the proteins of specific species. Therefore, high complementarity theoretically exists between the essential protein prediction algorithms based on biological characteristics and protein network characteristics. Hence, these two kinds of information could be combined to predict essential proteins.

Gene expression data provide genomic information under many different research conditions, and the PCC is often used to determine the similarity between two consecutive gene expression values. However, gene expression data have a large number of inherent noises, which can be eliminated by discretizing the gene expression spectrum using the threshold method. Given its wide usage in calculating the similarity of two discrete variables, the Jaccard similarity coefficient can be applied to measure the similarity of discretized gene expression between two proteins. The PCC and the Jaccard similarity coefficient are both similarity measures based on gene expression data; however, one is discrete and the other is continuous. Hence, these two can be simultaneously utilized to fill the gap between PPI and gene expression data.

Undirected graphs reveal two research directions: edges or points. From the perspective of edges, PeC and WDC are based on the Edge Clustering Coefficient (ECC) and PCC to add the weight values of different protein edges in adjacent fields of nodes. Zhong et al.[26] proposed a method named JDC using the ECC of the PPI network and the Jaccard similarity coefficient of gene expression data to integrate the weights of the different protein edges in the region adjacent to the nodes. Sun et al.[27] proposed a cross entropy based method for essential protein identification by using the weights of topological features of the PPI network and the PCC of gene expression data. From the perspective of nodes, we proposed a new similarity coefficient named PJC based on each protein node. The method of PJC can be divided into the following steps: (1) The continuous correlation coefficient between adjacent nodes is calculated according to the gene expression profile of each protein node. (2) Coding gene expression data are discretized by the threshold method, and the discrete similarity coefficient between each node is then obtained. (3) The network topology centrality characteristics of each node are calculated. (4) The continuous correlation coefficient of the adjacent edges of each node is weighted to obtain the sum of continuous similarity eigenvalues of the point (the same technique is applied to calculate the sum of discrete similarity eigenvalues of each node). (5) PJC consisting of the continuous and discrete similarities in gene expression data is proposed. This new coefficient is formed by combining the continuous and discrete similarity eigenvalues of each point. (6) Different protein centrality characteristics are fused by the newly proposed similarity coefficient PJC to calculate the score of each node, and the number of essential proteins is determined according to the score ranking. (7) Experimental analysis is performed, and a comparison is conducted between the centrality algorithm using the new similarity coefficient PJC and the essential protein prediction algorithm based on the same input data of network topology characteristics.

2.1 Continuous and discrete correlation coefficient

A new similarity coefficient containing continuous and discrete similarities named PJC based on the gene expression profile was proposed. PJC is flexible and can be superimposed on any topological centrality method to improve the identification of essential proteins. We introduced the various components of the new similarity coefficient PJC step by step and explained how this coefficient fuses node-based topology centrality.

2.1.1 Continuous correlation coefficient

The gene expression data are continuous data generated

by microarray experiments. PCC is widely used to measure the strength of the linear relationship between two objects. For two sequences of gene expressions, such as $X = (x_1, \ldots, x_l)$ and $Y = (y_1, \ldots, y_l)$, PCC could be defined as

$$\mathrm{PCC} = \frac{\sum_{i=1}^{l}(x_i - \overline{x}_i) \times (y_j - \overline{y}_j)}{\sqrt{\sum_{i=1}^{l}(x_i - \overline{x}_i^2) \times \sum_{j=1}^{l}(y_j - \overline{y}_j^2)}} \quad (1)$$

$$P(p) = \sum_{q \in D_p} \mathrm{PCC} \quad (2)$$

The PCC of each pair of interacting proteins is calculated according to the gene expression profile, and the obtained value is between -1 and 1. We defined the value of PCC as the similarity of gene expression data of protein p and protein q in PPI network cluster D. $P(p)$ is the sum of the PCC of all the edges connected to protein p.

2.1.2 Discretization of gene expression data

Gene expression data are generated by microarray or next-generation sequencing technology. This kind of high-throughput data has inevitable noise and is usually prone to large fluctuations. Genes exhibit dynamically active and inactive expression trends at different times. Sahoo[28] performed experiments on mouse B cells and conducted a Boolean analysis to understand gene regulation and gene function. In this study, threshold screening is used to discretize the gene expression data, and the characteristics of discrete data can be utilized to eliminate the influence of large fluctuation noise in the coding gene expression profiles.

The gene expression data obtained using the experimental techniques of biologists are constructed into a matrix S,

$$S = \begin{pmatrix} S_{11} & \cdots & S_{1M} \\ \vdots & \ddots & \vdots \\ S_{N1} & \cdots & S_{NM} \end{pmatrix} \quad (3)$$

where N is the number of genes, M refers to the expression cycle of genes.

$S_{p,t}$ is the expression level of the p-th gene at time t. If the expression value of $S_{p,t}$ is higher than the specified threshold, then the expression of the "active" gene is defined as "1". If the value of $S_{p,t}$ does not exceed the specified threshold, then the "inactive" gene expression is defined as "0". The screening threshold is shown in the following:

$$S'_{p,t} = \begin{cases} 1, & S_{p,t} > U(p) + \dfrac{U(p) + 2 \times \delta(p)}{1 + \delta^2(p)}; \\ 0, & S_{p,t} \leqslant U(p) + \dfrac{U(p) + 2 \times \delta(p)}{1 + \delta^2(p)} \end{cases} \quad (4)$$

where S is updated to a discretization matrix with only 0 and 1, which reflect the "active" and "inactive" states of gene expression; $U(p)$ is the mean of gene expression of the value of protein; $\delta(p)$ is the standard deviation; and $(U(p) + 2 \times \delta(p)/(1 + \delta^2(p)))$ represents the threshold for the activity of gene expression value of protein p.

2.1.3 Discrete similarity coefficient

The Jaccard similarity coefficient is generally used to measure the similarity of two discrete variables. In this work, this parameter is applied to measure the fluctuation degree of protein nodes in the PPI network and is estimated as follows:

$$J(p) = \sum_{q \in D_p} \mathrm{Jaccard}(p, q) = \sum_{q \in D_p} \frac{S_p \cap S_q}{S_p \cup S_q} \quad (5)$$

where S_p and S_q represent the discrete values of the gene expression data of protein p and protein q, respectively, and the Jaccard coefficient is between 0 and 1. We defined the value of the Jaccard coefficient as the activity expression similarity of protein p and protein q in PPI network cluster D, that is, the similarity of discretized gene expression data between the two proteins. $J(p)$ represents the superposition of the Jaccard coefficient of the domain edges in protein node p.

2.1.4 Fusion of PJC and node-base centrality

The PCC and the Jaccard similarity coefficient are both similarity measures based on gene expression data; one is discrete and the other is continuous. Hence, the combination of these two measures is appropriate for supplementing topological centrality. The discrete similarity and continuous similarity at each point are combined to obtain the new similarity coefficient PJC, and the calculation formula is as follows:

$$\mathrm{PJC}(p) = P(p) \times J(p) \quad (6)$$

Node-based network topology centrality algorithms are the basis for predicting essential proteins. We adopted nine commonly used algorithms for predicting essential proteins: DC, CC, SC, IC, EC, BC, NC, PeC, and WDC. These basic methods were then combined with the new similarity coefficient PJC to obtain the final score. The calculation formula is as follows:

$$\mathrm{PJCC}_{\mathrm{score}}(p) = \mathrm{PJC}(p) \times C_i(p) \quad (7)$$

where $C_i(p)$ is the eigenvalue of protein p with different prediction algorithms. The application of the PJC

similarity coefficient is described in detail based on the nodes in graph theory analysis, as shown in Algorithm 1.

Algorithm 1 applying the new similarity coefficient PJC mainly consists of seven steps. Step 1 is to calculate the different topological centralities of each node, and the time complexity is $O(n)$, where n represents the number of protein nodes. In Step 2, the continuous correlation coefficient between each node and its adjacent edge is calculated and superimposed, and the time complexity is $O(n^2+n)$. Step 3 is to construct the discrete gene expression matrix, and the time complexity is $O(n \times m)$, where m refers to the expression cycle of genes. In Step 4, the discrete coefficient between each node and its adjacent edge is calculated and summed, and the time complexity is $O(n^2+n)$; Step 5 is to calculate the continuous and discrete similarity coefficient PJC of each protein node, and the time complexity is $O(n)$. Step 6 combines the topological centrality of Step 1 with the PJC coefficient of Step 5 to obtain the score, and the time complexity is $O(n)$. In Step 7, the first $K\%$ proteins are selected as the essential protein output in descending order of score, and the time complexity is $O(n)$.

Algorithm 1 Application of the PJC similarity coefficient

Input: PPI network $G = (V, E)$, gene expression dataset, protein nodes p
Output: Proteins in the top $N\%$ of the $\mathrm{PJCC}_{\mathrm{score}}(p)$
Step 1: Calculate the topological eigenvalues of different proteins:
  for each $p \in V$ **do**
    Calculate $\mathrm{DC}(p), \mathrm{CC}(p), \mathrm{SC}(p), \mathrm{IC}(p), \mathrm{EC}(p), \mathrm{BC}(p), \mathrm{NC}(p), \mathrm{PeC}(p)$, and $\mathrm{WDC}(p)$.
  end for
Step 2: Calculate the sum of the continuous correlation coefficient of all the edges connected to protein p:
  for each $p \in V$ **do**
    Calculate $\mathrm{P}(p)$ according to Eq. (1)
  end for
Step 3: Construct discrete values matrix based on Eq. (3):
  for each $p \in V$ **do**
    for each $t \in M$ **do**
      If the expression level of the p-th gene at time $t >$ discrete threshold $U(p) + 2 \times \delta(p)/(1+\delta^2(p))$
        the expression level $S'_{(p,t)} = 1$
      else
        the expression level $S'_{(p,t)} = 0$
      end if
    end for
  end for
Step 4: Calculate the superposition of the discrete coefficient of the domain edges in protein node p:
  for each $p \in V$ **do**
    Calculate $J(p)$ according to Eq. (5)
  end for
Step 5: Calculate continuous and discrete similarity coefficient PJC of each protein node:
  for each $p \in V$ **do**
    Calculate $PJC(p)$ according to Eq. (6)
  end for
Step 6: Calculate the fusion score of each protein node:
  for each $p \in V$ **do**
    Calculate $\mathrm{PJCC}_{\mathrm{score}}(p)$ according to Eq. (7)
  end for
Step 7: Rank all proteins in descending order by $\mathrm{PJCC}_{\mathrm{score}}(p)$ score and output the top $K\%$ of all proteins.

3 Experimental Result and Analysis

3.1 Experimental dataset

Yeast PPI and essential protein data are the most complete and reliable among all species. Therefore, we use yeast protein as the experimental object to verify the effectiveness of our algorithm.

(1) The data of Bakers' yeast and DIP' yeast are used in our study. The Bakers' yeast has two sets of PPI network data, namely, Krogan and Gavin. The PPI data of Krogan and Gavin are from the BioGRID database[29], and the data of Saccharomyces cerevisiae are obtained from the DIP database. These PPI data are then preprocessed to remove self-interaction and repeated interactions.Table 1 shows the details of these three PPIs.

(2) Essential protein data: The standard essential protein data include 1285 essential proteins, which are mainly derived from MIPS[30], SGD[31], DEG[32], and SGDP[33]. Among them, 1167 essential proteins are obtained from the yeast PPI network.

(3) Gene expression data: Yeast gene expression data are downloaded from NCBI (GSE3431) Gene Expression Omnibus website, and contain 9336 genes at 36-time points across three cell metabolic cycles. A total of 6777 gene products and 36 samples are obtained after pretreatment and normalization. About 98.88% of the proteins are covered in the Krogan data and 99.16% in the Gavin data.

Table 1 Details about Krogan, Gavin, and DIP databases.

Dataset	Number of proteins	Number of interaction edges	Number of essential proteins
Krogan	2674	7075	784
Gavin	1430	6531	617
DIP	5093	24 743	1167

3.2 Prediction performance evaluation analysis based on Receiver Operating Characteristic (ROC) curve and AUC

ROC curve analysis can reveal the performance in binary classification. Hence, the ROC curve is used to evaluate the overall performance of each method. The experimental results with the Krogan database are shown in Fig. 1. We selected the top 784 proteins, 617 proteins, and 1167 proteins in each method as thresholds for predicting the AUC of essential proteins in the Krogan, Gavin, and DIP databases.

Figure 1 shows the different centrality measures (DC, CC, SC, IC, BC, EC, NC, PeC, and WDC) in the Krogan database and the ROC curves of these nine essential protein prediction algorithms under different states. When combined with the newly proposed similarity coefficient PJC (measures are added with *PJC, for example, DC*PJC; added with *P meaning only combined the continuous correlation coefficient with the DC algorithm), the AUC values for these nine essential protein prediction methods are 0.6407, 0.6387, 0.6326, 0.6417, 0.6161, 0.6306, 0.6242, 0.6317, and 0.6830. Although the overall AUC of DC combined with the similarity coefficient PJC is slightly lower than that of the original method, the subsequent top-rank analysis shows that the new similarity coefficient PJC helps the DC algorithm to predict a great number of essential proteins and produce a reliable prediction score. IC and NC show a similar situation to DC. We only combine the continuous correlation coefficient with different prediction algorithms for comparison with PJC.

In the binary classification system, the protein node recognition algorithm based on the continuous and discrete similarity coefficient PJC and multiple biological characteristics can obtain a high true positive rate while maintaining a low false positive rate. To ensure the versatility of our method, we also compare the ROC curves of the nine prediction methods for essential proteins in the Gavin and DIP databases. The ROC values for the Gavin and DIP database are shown in Figs. 2 and 3, respectively. The AUC predicted by the new similarity coefficient PJC is improved to a certain extent compared with that of different prediction methods, except for the special cases of DC, IC, and NC. To further illustrate the superiority of the newly proposed PJC, we also compare the results of combining the topological eigenvalues of different basic centrality

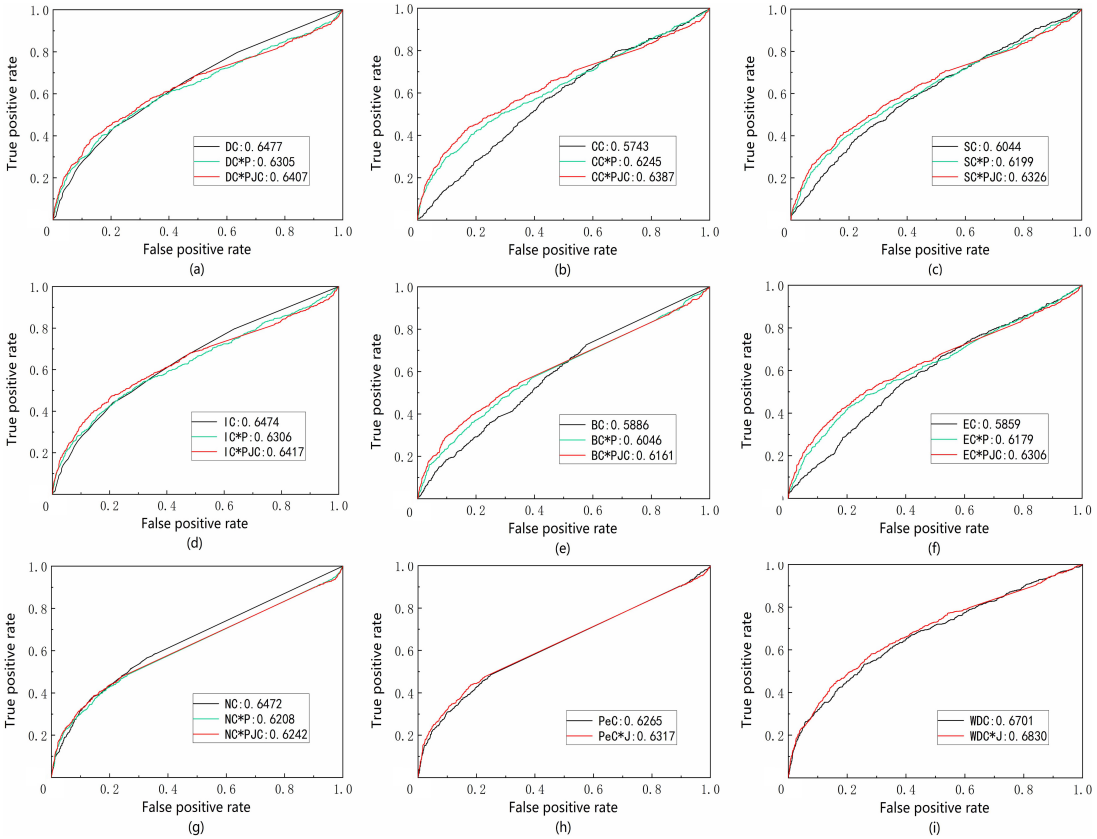


Fig. 1 ROC curves and AUC values of different prediction methods in the Krogan database.

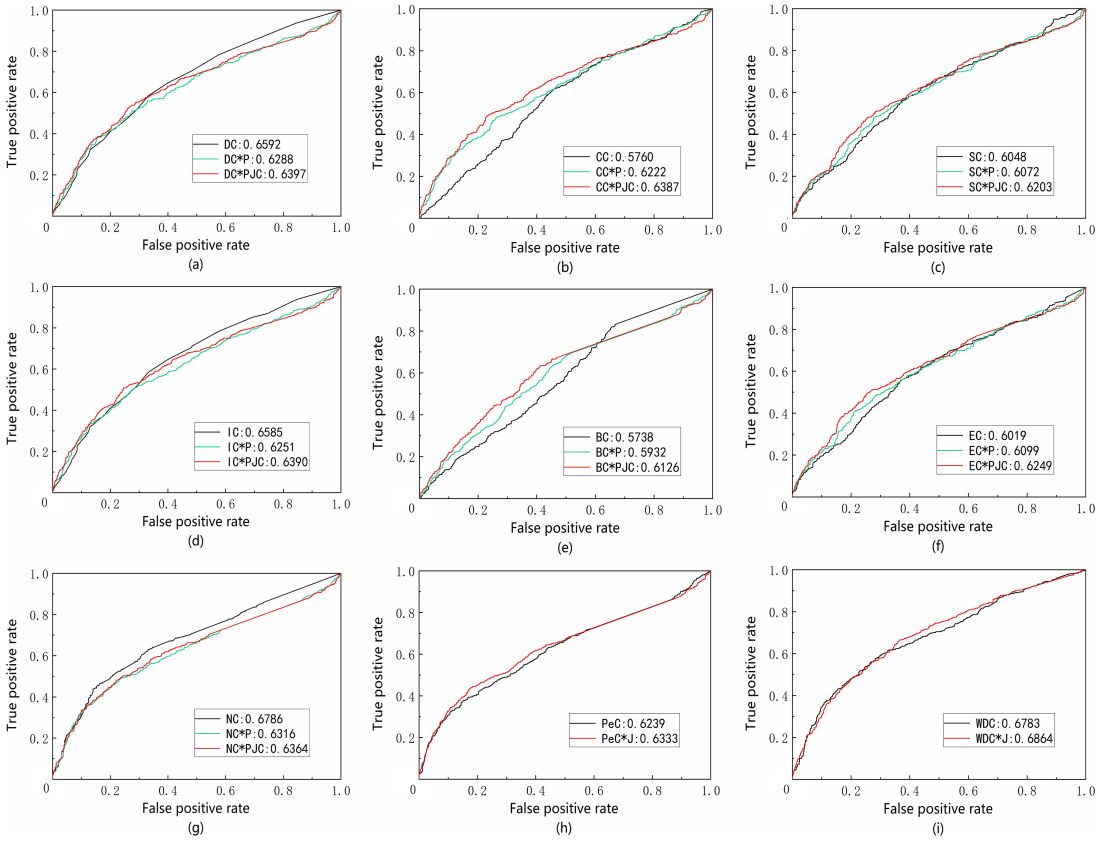


Fig. 2 ROC curves and AUC values of different prediction methods in the Gavin database.

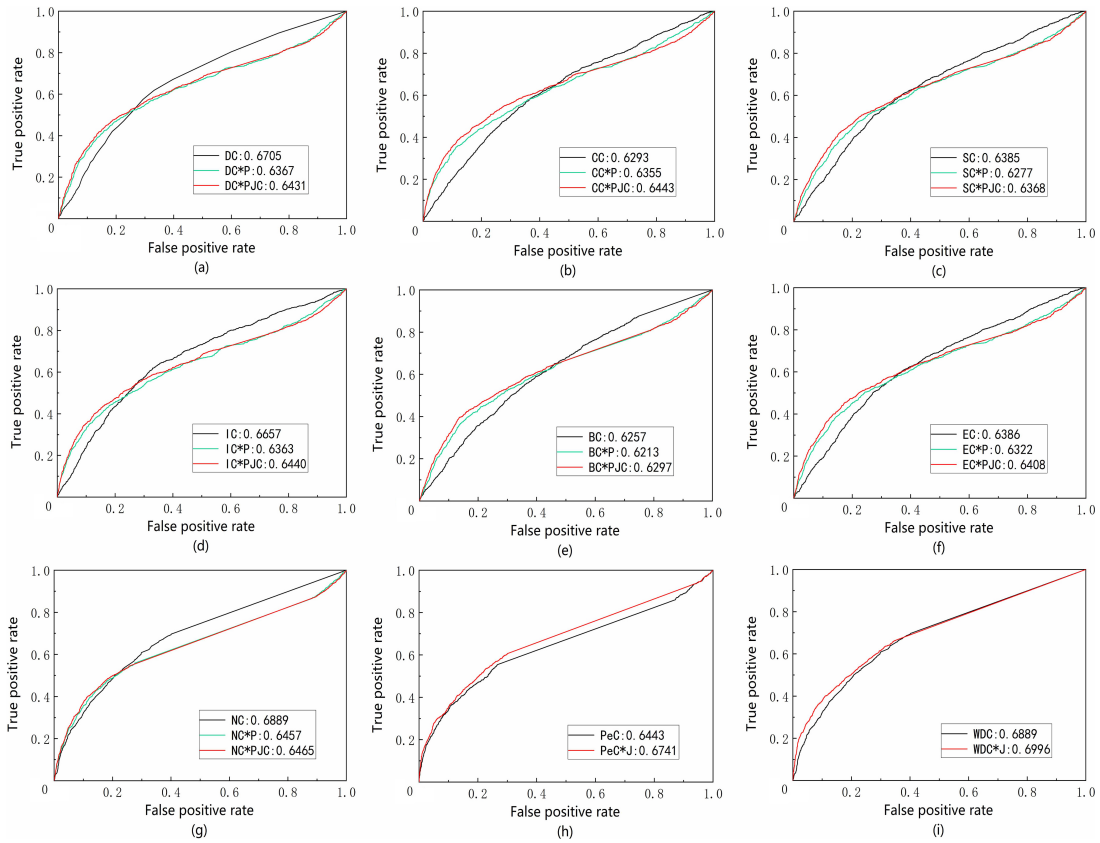


Fig. 3 ROC curves and AUC values of different prediction methods in the DIP database.

methods with PCCs. Given that PCC has already been used for PeC and WDC, it will not be applied in this study. Hence, the discrete coefficient is directly combined with PeC and WDC. According to the above analysis, the prediction algorithm integrating the similarity coefficient PJC and multibiological features exhibit superior performance. This finding further demonstrate the strong correlation between protein topology networks and prediction characteristics.

3.3 Prediction performance evaluation based on top analysis

To further verify the performance of our method, we combine the new similarity coefficient PJC with seven centrality prediction methods that have topological features (DC, IC, EC, SC, BC, CC, and NC) and prediction methods with the same input data (PeC and WDC) to achieve the top analysis. We also integrate the continuous correlation coefficient into different prediction algorithms for comparison. We select 1%, 5%, 10%, 15%, 20%, and 25% of proteins according to the descending order of each method score and determine how many of them are essential proteins.

Table 2 shows that when we select the top 1% proteins in Krogan database, the DC, BC, CC, EC, SC, IC, NC, PeC, and WDC applying the similarity coefficient PJC can identify 22, 21, 23, 22, 23, 23, 23, 23, and 22 essential proteins, respectively. According to the number of times exceeded by different prediction methods in the Over_times column of Table 2, PJC can identify a large number of essential proteins. As shown in Tables 3 and 4, the number of essential proteins identified by combining the continuous and discrete similarity coefficient PJC is substantially increased at each percentage in the Gavin and DIP databases. However, the top 15%, 20%, and 25% of the WDC algorithm do not perform well because the number of protein nodes in the Gavin database is higher than that of the edges. Hence, the proteins are less affected by interaction fluctuations.

3.4 Prediction performance evaluation based on accuracy analysis

Sensitivity (*SN*), specificity (*SP*), False Positive Rate (*FPR*), Negative Predictive Value (*NPV*), Positive Predictive Value (*PPV*), *F-measure*, accuracy (*ACC*), and Matthew Correlation Coefficient (*MCC*) are also applied for the validation of essential protein discovery methods, the definitions are as follows:

$$SN = \frac{TP}{TP + FN} \tag{8}$$

Table 2 Prediction results of nine algorithms' top analysis in the Krogan database.

Method	1%	5%	10%	15%	20%	25%	Over_times
DC	12	81	147	214	266	318	0
DC*P	20	95	170	226	274	319	0
DC*PJC	**22**	**96**	**172**	**231**	**294**	**335**	6
BC	11	62	118	159	206	248	0
BC*P	17	82	146	198	245	291	0
BC*PJC	**21**	**92**	**159**	**224**	**269**	**311**	6
CC	9	50	104	145	194	239	0
CC*P	20	96	165	225	264	317	0
CC*PJC	**23**	**96**	**173**	**239**	**289**	**341**	6
EC	20	60	109	149	203	253	0
EC*P	21	80	155	208	261	313	0
EC*PJC	**22**	**94**	**168**	**227**	**279**	**327**	6
SC	20	63	118	175	227	280	0
SC*P	21	80	156	211	260	310	0
SC*PJC	**23**	**90**	**166**	**225**	**267**	**326**	6
IC	12	81	147	214	266	318	0
IC*P	21	96	168	222	273	320	0
IC*PJC	**23**	**95**	**172**	**237**	**294**	**340**	6
NC	22	91	161	229	**286**	325	1
NC*P	22	99	176	229	272	321	0
NC*PJC	**23**	**104**	**180**	**234**	284	**336**	5
PeC	19	100	174	226	274	318	0
PeC*J	**23**	**104**	**180**	**236**	**285**	**336**	6
WDC	21	100	176	235	278	333	0
WDC*J	**22**	**102**	**182**	**239**	**300**	**355**	6

Table 3 Prediction results of nine algorithms' top analysis in the Gavin database.

Method	1%	5%	10%	15%	20%	25%	Over_times
DC	12	47	91	144	185	221	0
DC*P	12	46	94	146	192	233	0
DC*PJC	12	**53**	**98**	**147**	**194**	**234**	5
BC	7	42	77	114	143	172	0
BC*P	8	42	86	127	161	193	0
BC*PJC	**10**	**47**	**88**	**135**	**173**	**207**	6
CC	9	36	70	109	143	173	0
CC*P	13	51	101	**150**	186	222	1
CC*PJC	13	**56**	**106**	148	**189**	**230**	4
EC	13	51	87	127	156	192	0
EC*P	13	53	96	134	167	209	0
EC*PJC	13	53	**99**	**138**	**178**	**227**	4
SC	13	51	88	127	157	194	0
SC*P	13	53	96	130	162	204	0
SC*PJC	12	**54**	**99**	**133**	**173**	**218**	5
IC	12	47	91	144	185	221	0
IC*P	12	48	98	146	189	224	0
IC*PJC	**13**	**56**	**106**	**148**	**193**	**233**	6
NC	13	52	109	156	200	**252**	1
NC*P	13	53	105	**159**	203	236	1
NC*PJC	13	**55**	109	154	**205**	238	2
PeC	14	**58**	110	157	199	233	1
PeC*J	14	57	**112**	**158**	**203**	**241**	4
WDC	13	53	107	**159**	**208**	**247**	3
WDC*J	13	**58**	**108**	153	199	241	2

Table 4 Prediction results of nine algorithms' top analysis in the DIP database.

Method	1%	5%	10%	15%	20%	25%	Over_times
DC	22	101	207	320	413	502	0
DC*P	31	145	274	380	463	535	0
DC*PJC	**32**	**156**	**289**	**391**	**481**	**548**	6
BC	24	95	182	271	361	433	0
BC*P	26	122	246	345	438	497	0
BC*PJC	**29**	**140**	**263**	**370**	**464**	**522**	6
CC	24	104	193	284	364	448	0
CC*P	36	162	280	378	449	513	0
CC*PJC	36	**169**	**295**	**397**	**474**	**535**	5
EC	24	96	195	279	377	467	0
EC*P	29	137	261	357	451	514	0
EC*PJC	**33**	**154**	**285**	**379**	**468**	**542**	6
SC	24	96	195	279	377	467	0
SC*P	26	135	248	336	433	507	0
SC*PJC	**28**	**149**	**266**	**366**	**462**	**532**	6
IC	24	102	210	316	406	504	0
IC*P	36	159	280	376	455	521	0
IC*PJC	36	**168**	**295**	**401**	**477**	**536**	5
NC	32	159	282	372	464	544	0
NC*P	**36**	**169**	295	397	483	554	2
NC*PJC	35	168	**304**	**409**	**492**	**562**	4
PeC	40	174	294	388	466	543	0
PeC*J	**44**	**188**	**317**	**395**	**481**	**553**	6
WDC	36	164	303	400	487	566	0
WDC*J	**45**	**192**	**319**	**422**	**497**	**567**	6

$$SP = \frac{TN}{TN + FP} \tag{9}$$

$$FPR = \frac{FP}{TN + FP} \tag{10}$$

$$NPV = \frac{TN}{TN + FP} \tag{11}$$

$$PPV = \frac{TP}{TP + FP} \tag{12}$$

$$F\text{-}measure = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{13}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN)}} \times \frac{1}{\sqrt{(TN + FP) \times (TN + FN)}} \tag{15}$$

where *TP* represents the number of true positive proteins, which are truly essential proteins that are correctly predicted to be essential. *FP* denotes the number of false positive proteins, which are non-essential proteins that are wrongly predicted to be essential. *TN* is the number of true negative proteins, which are non-essential proteins that are accurately predicted to be non-essential. *FN* refers to the number of false-negative proteins, which are truly essential proteins that are mistakenly predicted to be non-essential. The results on Krogan, Gavin, and Yeast data are shown in Tables 5, 6, and 7, respectively.

As shown in Table 5, the *SN* values of DC, BC, CC, EC, SC, IC, PeC, and WDC combined with the new similarity coefficient PJC on the Krogan database are 0.4694, 0.4355, 0.4681, 0.4617, 0.4521, 0.4764, 0.4687, and 0.4962, respectively. Each evaluation criterion of the centrality method using the new similarity coefficient PJC is better than the original prediction methods, except for the NC algorithm. Table 5 shows that NC using PJC slightly underperforms relative to the original NC in predicting the number of top essential proteins. For the Gavin database in Table 6, IC and NC show the same trend as the NC of the Krogan database. One reason is that we focus on the number of essential proteins predicted at the top of the list. In the previous top analysis, PJC combined with NC and IC predicts more protein quantities than the original. Another reason is that the Gavin database has many redundancy nodes and the edge effect is relatively small. The similarity

Table 5 ***SN*, *SP*, *FPR*, *PPV*, *NPV*, *F-measure*, *ACC*, and *MCC* of various methods for the total ranked proteins in the Krogan database.**

Method	*SN*	*SP*	*FPR*	*PPV*	*NPV*	*F-measure*	*ACC*	*MCC*
DC	0.4554	0.7741	0.2259	0.7741	0.4554	0.4554	0.6806	0.2294
DC*P	0.4521	0.7726	0.2274	0.7730	0.4515	0.4518	0.6788	0.2246
DC*PJC	**0.4694**	**0.7799**	**0.2201**	**0.7799**	**0.4694**	**0.4694**	**0.6889**	**0.2493**
BC	0.3673	0.7376	0.2624	0.7376	0.3673	0.3673	0.6290	0.1049
BC*P	0.4145	0.7571	0.2429	0.7571	0.4145	0.4145	0.6567	0.1717
BC*PJC	**0.4355**	**0.7657**	**0.2343**	**0.7661**	**0.4349**	**0.4352**	**0.6690**	**0.2012**
CC	0.3533	0.7317	0.2683	0.7317	0.3533	0.3533	0.6208	0.0851
CC*P	0.4457	0.7700	0.2300	0.7704	0.4452	0.4454	0.6750	0.2156
CC*PJC	**0.4681**	**0.7794**	**0.2206**	**0.7794**	**0.4681**	**0.4681**	**0.6881**	**0.2475**
EC	0.3737	0.7402	0.2598	0.7402	0.3737	0.3737	0.6328	0.1139
EC*P	0.447	0.7705	0.2295	0.7709	0.4464	0.4467	0.6758	0.2174
EC*PJC	**0.4617**	**0.7767**	**0.2233**	**0.7767**	**0.4617**	**0.4617**	**0.6844**	**0.2385**
SC	0.4082	0.7545	0.2455	0.7545	0.4082	0.4082	0.6530	0.1627
SC*P	0.4337	0.7651	0.2349	0.7651	0.4337	0.4337	0.6679	0.1988
SC*PJC	**0.4521**	**0.7726**	**0.2274**	**0.7730**	**0.4515**	**0.4518**	**0.6788**	0.2246
IC	0.4528	0.7730	0.2270	0.7730	0.4528	0.4528	0.6791	0.2258
IC*P	0.4515	0.7725	0.2275	0.7725	0.4515	0.4515	0.6784	0.2240
IC*PJC	**0.4764**	**0.7827**	**0.2173**	**0.7831**	**0.4758**	**0.4761**	**0.6930**	**0.2589**
NC	**0.4592**	**0.7757**	**0.2243**	**0.7757**	**0.4592**	**0.4592**	**0.6829**	**0.2348**
NC*P	0.4503	0.7720	0.2280	0.7720	0.4503	0.4503	0.6776	0.2222
NC*PJC	0.4566	0.7746	0.2254	0.7746	0.4566	0.4566	0.6814	0.2312
PeC	0.4592	0.7757	0.2243	0.7757	0.4592	0.4592	0.6829	0.2348
PeC*J	**0.4687**	**0.7795**	**0.2205**	**0.7799**	**0.4681**	**0.4684**	**0.6885**	**0.2481**
WDC	0.4732	0.7815	0.2185	0.7815	0.4732	0.4732	0.6911	0.2547
WDC*J	**0.4962**	**0.7910**	**0.2090**	**0.7910**	**0.4962**	**0.4962**	**0.7046**	**0.2872**

Table 6 ***SN*, *SP*, *FPR*, *PPV*, *NPV*, *F-measure*, *ACC*, and *MCC* of various methods for the total ranked proteins in the Gavin database.**

Method	*SN*	*SP*	*FPR*	*PPV*	*NPV*	*F-measure*	*ACC*	*MCC*
DC	0.5673	0.6716	0.3284	0.6716	0.5673	0.5673	0.6266	0.2388
DC*P	0.5592	0.6654	0.3346	0.6654	0.5592	0.5592	0.6196	0.2246
DC*PJC	**0.5754**	**0.6777**	**0.3223**	**0.6777**	**0.5754**	**0.5754**	**0.6336**	**0.2531**
BC	0.4700	0.5978	0.4022	0.5978	0.4700	0.4700	0.5427	0.0678
BC*P	0.5154	0.6322	0.3678	0.6322	0.5154	0.5154	0.5818	0.1476
BC*PJC	**0.5446**	**0.6544**	**0.3456**	**0.6544**	**0.5446**	**0.5446**	**0.6070**	**0.1989**
CC	0.4992	0.6199	0.3801	0.6199	0.4992	0.4992	0.5678	0.1191
CC*P	0.5332	0.6458	0.3542	0.6458	0.5332	0.5332	0.5972	0.1790
CC*PJC	**0.5559**	**0.6630**	**0.3370**	**0.6630**	**0.5559**	**0.5559**	**0.6168**	**0.2189**
EC	0.5284	0.6421	0.3579	0.6421	0.5284	0.5284	0.5930	0.1704
EC*P	0.5365	0.6482	0.3518	0.6482	0.5365	0.5365	0.6000	0.1847
EC*PJC	**0.5494**	**0.6581**	**0.3419**	**0.6581**	**0.5494**	**0.5494**	**0.6112**	**0.2075**
SC	0.5300	0.6433	0.3567	0.6433	0.5300	0.5300	0.5944	0.1733
SC*P	0.5397	0.6507	0.3493	0.6507	0.5397	0.5397	0.6028	0.1904
SC*PJC	**0.5511**	**0.6593**	**0.3407**	**0.6593**	**0.5511**	**0.5511**	**0.6126**	**0.2103**
IC	**0.5673**	**0.6716**	**0.3284**	**0.6716**	**0.5673**	**0.5673**	**0.6266**	**0.2388**
IC*P	0.5430	0.6531	0.3469	0.6531	0.5430	0.5430	0.6056	0.1961
IC*PJC	0.5640	0.6691	0.3309	0.6691	0.5640	0.5640	0.6238	0.2331
NC	**0.5964**	**0.6937**	**0.3063**	**0.6937**	**0.5964**	**0.5964**	**0.6517**	**0.2902**
NC*P	0.5575	0.6642	0.3358	0.6642	0.5575	0.5575	0.6182	0.2217
NC*PJC	0.5640	0.6691	0.3309	0.6691	0.5640	0.5640	0.6238	0.2331
PeC	0.5365	0.6482	0.3518	0.6482	0.5365	0.5365	0.6000	0.1847
PeC*J	**0.5559**	**0.6630**	**0.3370**	**0.6630**	**0.5559**	**0.5559**	**0.6168**	**0.2189**
WDC	**0.5981**	**0.6950**	**0.3050**	**0.6950**	**0.5981**	**0.5981**	**0.6531**	**0.2930**
WDC*J	0.5867	0.6863	0.3137	0.6863	0.5867	0.5867	0.6434	0.2731

Table 7 ***SN*, *SP*, *FPR*, *PPV*, *NPV*, *F-measure*, *ACC*, and *MCC* of various methods for the total ranked proteins in the DIP database.**

Method	*SN*	*SP*	*FPR*	*PPV*	*NPV*	*F-measure*	*ACC*	*MCC*
DC	0.4002	0.8217	0.1783	0.8217	0.4002	0.4002	0.7251	0.2219
DC*P	0.4310	0.8309	0.1691	0.8309	0.4310	0.4310	0.7393	0.2619
DC*PJC	**0.4473**	**0.8357**	**0.1643**	**0.8357**	**0.4473**	**0.4473**	**0.7467**	**0.2830**
BC	0.3505	0.8069	0.1931	0.8069	0.3505	0.3505	0.7023	0.1574
BC*P	0.4087	0.8242	0.1758	0.8242	0.4087	0.4087	0.7290	0.2330
BC*PJC	**0.4310**	**0.8309**	**0.1691**	**0.8309**	**0.4310**	**0.4310**	**0.7393**	**0.2619**
CC	0.3548	0.8082	0.1918	0.8082	0.3548	0.3548	0.7043	0.1630
CC*P	0.4165	0.8265	0.1735	0.8265	0.4165	0.4165	0.7326	0.2430
CC*PJC	**0.4422**	**0.8342**	**0.1658**	**0.8342**	**0.4422**	**0.4422**	**0.7444**	**0.2763**
EC	0.3676	0.8120	0.1880	0.8120	0.3676	0.3676	0.7102	0.1796
EC*P	0.4147	0.8260	0.1740	0.8260	0.4147	0.4147	0.7318	0.2408
EC*PJC	**0.4404**	**0.8337**	**0.1663**	**0.8337**	**0.4404**	**0.4404**	**0.7436**	**0.2741**
SC	0.3676	0.8120	0.1880	0.8120	0.3676	0.3676	0.7102	0.1796
SC*P	0.3676	0.8120	0.1880	0.8120	0.3676	0.3676	0.7102	0.1796
SC*PJC	**0.4336**	**0.8316**	**0.1684**	**0.8316**	**0.4336**	**0.4336**	**0.7404**	**0.2652**
IC	0.4010	0.8220	0.1780	0.8220	0.4010	0.4010	0.7255	0.2230
IC*P	0.4259	0.8293	0.1707	0.8293	0.4259	0.4259	0.7369	0.2552
IC*PJC	**0.4439**	**0.8347**	**0.1653**	**0.8347**	**0.4439**	**0.4439**	**0.7451**	**0.2786**
NC	0.4353	0.8321	0.1679	0.8321	0.4353	0.4353	0.7412	0.2674
NC*P	0.4482	0.8360	0.1640	0.8360	0.4482	0.4482	0.7471	0.2841
NC*PJC	**0.4559**	**0.8383**	**0.1617**	**0.8383**	**0.4559**	**0.4559**	**0.7506**	**0.2941**
PeC	0.4362	0.8324	0.1676	0.8324	0.4362	0.4362	0.7416	0.2686
PeC*J	**0.4499**	**0.8365**	**0.1635**	**0.8365**	**0.4499**	**0.4499**	**0.7479**	**0.2863**
WDC	0.4353	0.8321	0.1679	0.8321	0.4353	0.4353	0.7412	0.2674
WDC*J	**0.4584**	**0.8390**	**0.1610**	**0.8390**	**0.4584**	**0.4584**	**0.7518**	**0.2975**

coefficient PJC has an excellent value of eight evaluation criteria in the DIP database, as shown in Table 7. A low *FPR* indicates that the method is good at predicting. In Tables 5–7, the numbers in boldface refer to relative optimal result.

3.5 Prediction performance evaluation based on Jackknife analysis

Jackknife methodology is used to evaluate and predict the performance differences of essential proteins. For each prediction method, we evaluate its performance by calculating the sum of the actual essential protein and the predicted number of essential proteins. The Jackknife curve evaluation and analysis results of Krogan, Gavin, and DIP databases are shown in Figs. 4, 5, and 6, respectively.

We ranked the prediction results of each method in descending order. Figures 4 and 6 show that DC, IC, EC, SC, BC, CC, NC, WDC, and PeC have higher curves and larger AUC areas when combined with the newly proposed similarity coefficient PJC, indicating that PJC can improve the accuracy of essential protein prediction. However, Fig. 5 reveals that the areas of AUCs of PeC and WDC only slightly increase upon the combination with PJC. A possible explanation is that these two prediction algorithms may be relatively stable and have minimal dependence on our correlation coefficients. Further improvement of this simple fusion coefficient consisting of continuous and discrete similarities in gene expression data will be investigated in a future study.

3.6 Prediction performance evaluation based on NF-PIN dynamic network

In the previous chapter, the newly proposed similarity coefficient PJC is fused with various protein prediction methods based on a static PPI network, and the experimental results show that PJC can improve the accuracy of predicting essential proteins. To further prove the superiority of PJC, we incorporate it into the dynamic NF-PIN network, which also uses gene expression data, and then fuse it with node-base topological centrality. Meanwhile, the topological centrality method of the dynamic NF-PIN network is used for comparison, and the DIP database is selected for the experiment.

The coefficient PJC consisting of continuous and discrete similarities based on gene expression data in a dynamic network can also help improve the number of predicting essential proteins to some certain extent. Figure 7 shows that combined with the similarity

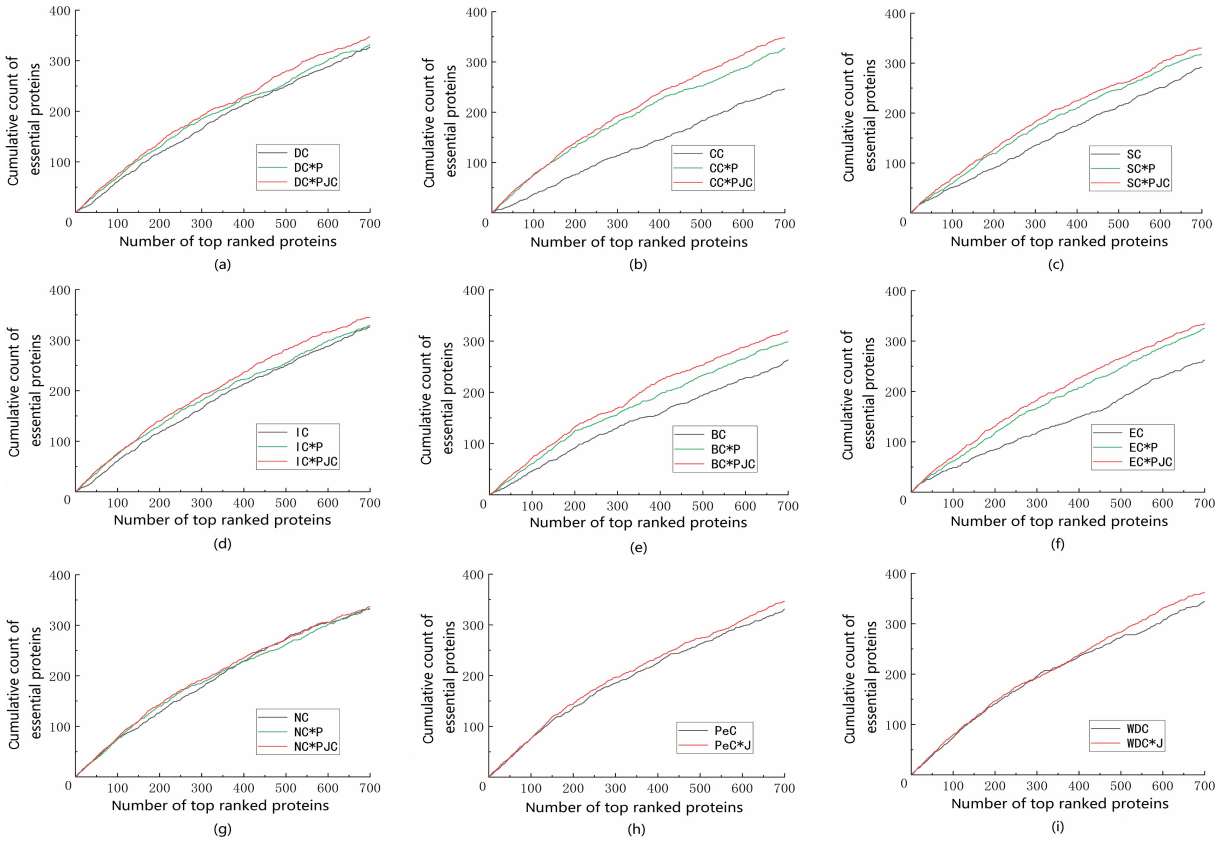


Fig. 4 ROC curves and AUC values of different prediction methods in the Krogan database.

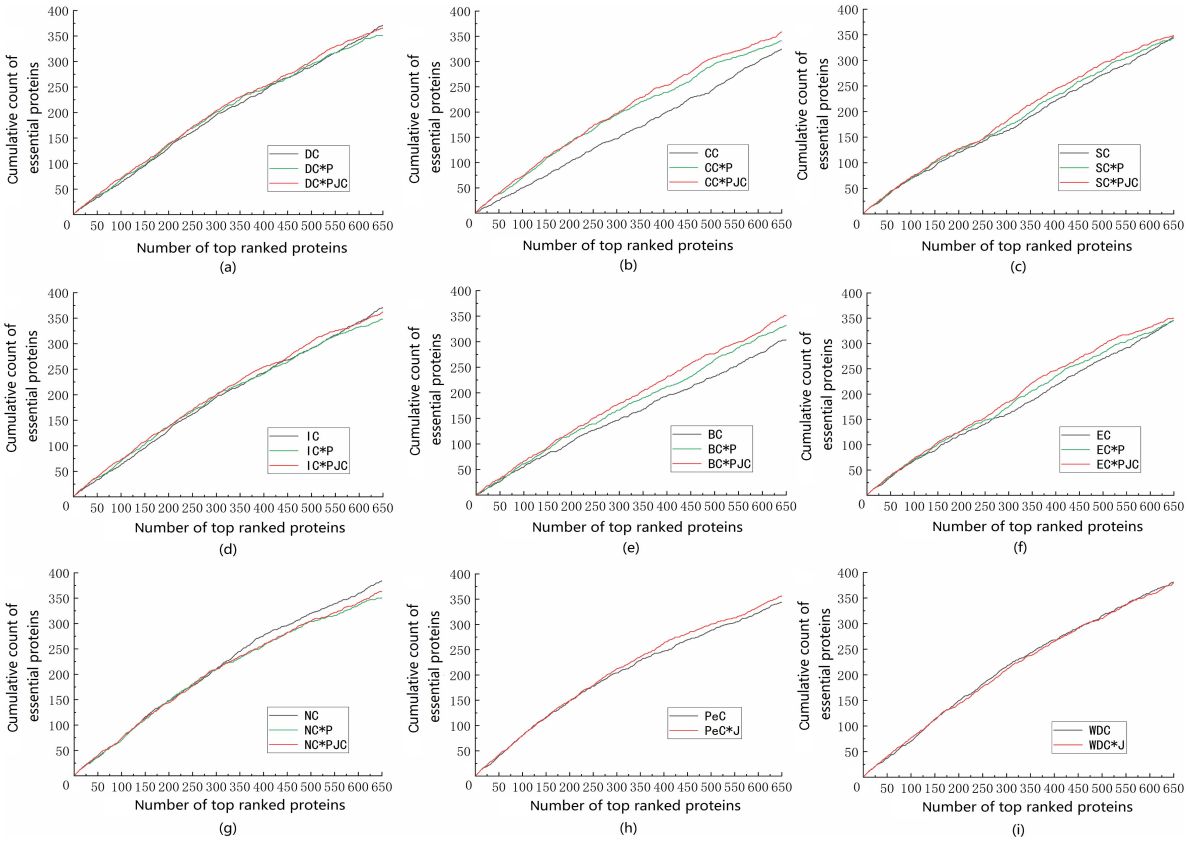


Fig. 5 ROC curves and AUC values of different prediction methods in the Gavin database.

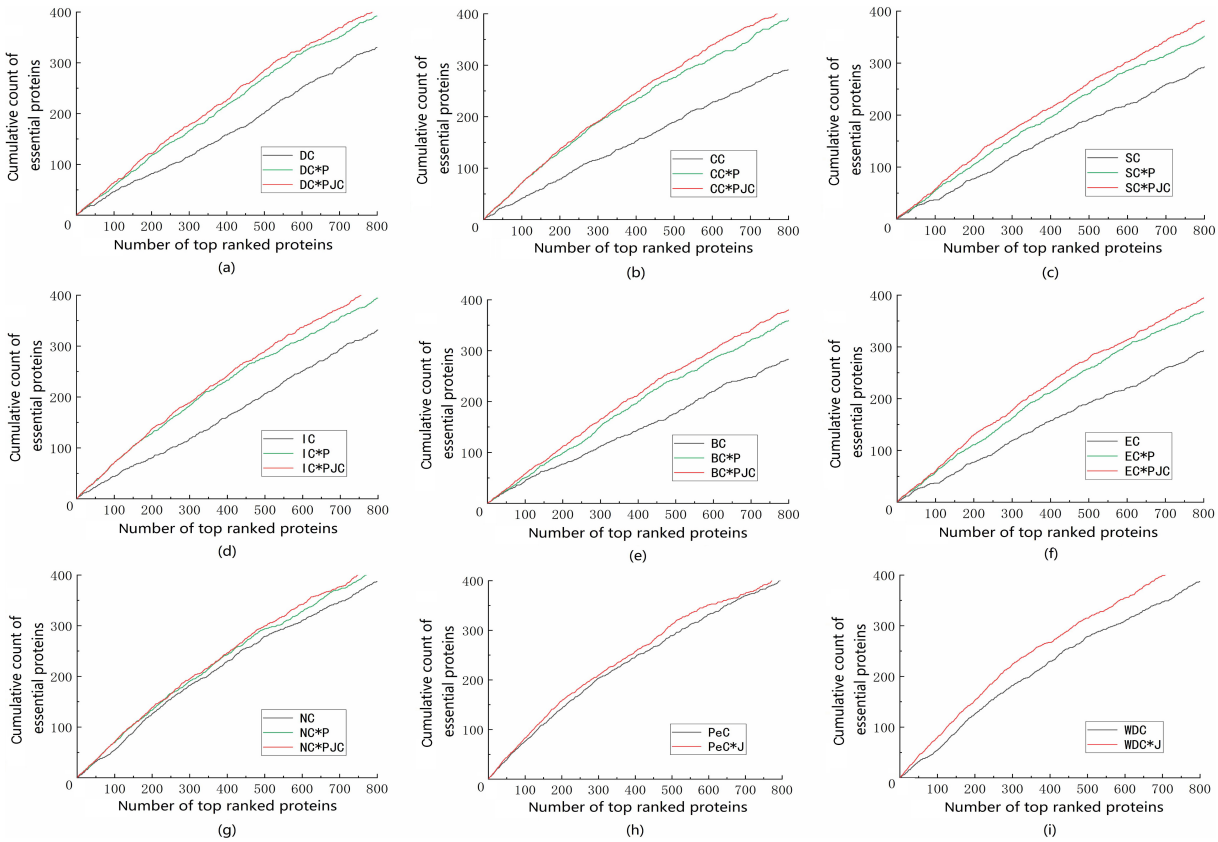


Fig. 6 ROC curves and AUC values of different prediction methods in the DIP database.

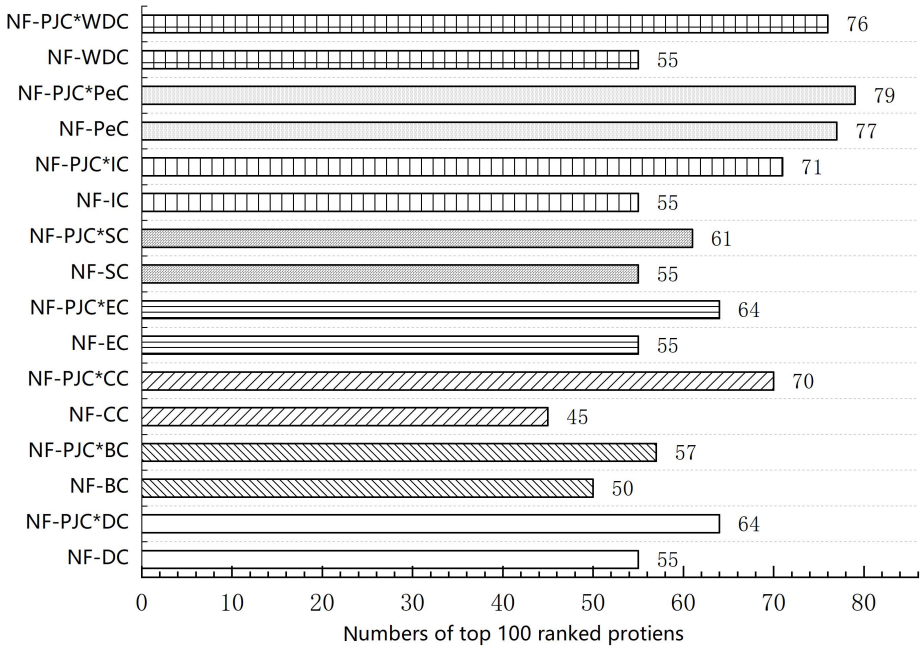


Fig. 7 Histogram of top 100 essential proteins in NF-PIN dynamic network in DIP database.

coefficient PJC in NF-PIN dynamic network, the quantity of the top 100 essential proteins in DC, BC, CC, EC, SC, IC, PeC, and WDC is 64, 57, 70, 64, 61, 71, 79, and 76, respectively. All of these values imply a certain improvement compared with the topological centrality in the NF-PIN network. Whether in static or dynamic networks, the combination of continuous and discrete similarity coefficients of gene expression data is feasible and reliable.

4 Conclusion

In this study, we proposed a new coefficient named PJC consisting of continuous and discrete similarities based on gene expression data. The discrete similarity coefficient was obtained by discretizing the gene expression data and then fused with the continuous correlation coefficient of the gene expression profile to obtain the similarity coefficient PJC. PJC eliminates the influence of large noise fluctuations on PPI network data and gene expression data. We analyzed biological characteristics and the topology centrality of protein networks in the essential protein algorithms and showed that they can be highly complementary. Therefore, the newly proposed similarity coefficient PJC can be flexibly applied to PPI network topology centrality to improve the identification efficiency of essential proteins. We described the similarity coefficient PJC in detail and carried out experiments using Krogan, Gavin, and DIP's PPI network in yeast datasets. ROC analysis, jackknife analysis, top analysis, and accuracy analysis revealed that node-base topology centrality fused with the new similarity coefficient PJC has superior advantages in predicting essential proteins. The topological centrality method fused with PCC was also added for comparison and also showed high accuracy and precision.

Acknowledgment

This work was supported by the Shenzhen KQTD Project (No. KQTD20200820113106007), China Scholarship Council (No. 201906725017), the Collaborative Education Project of Industry-University cooperation of the Chinese Ministry of Education (No. 201902098015), the Teaching Reform Project of Hunan Normal University (No. 82), and the National Undergraduate Training Program for Innovation (No. 202110542004).

References

[1] P. R. Graves and T. A. J. Haystead, Molecular biologist's guide to proteomics, *Microbiol. Mol. Biol. Rev.*, vol. 66, no. 1, pp. 39–63, 2002.

[2] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, et al., Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis, *Science*, vol. 285, no. 5429, pp. 901–906, 1999.

[3] S. Asur, D. Ucar, and S. Parthasarathy, An ensemble framework for clustering protein-protein interaction networks, *Bioinformatics*, vol. 23, no. 13, pp. i29–i40, 2007.

[4] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. H. Yang, X. C. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, et al., Interaction network containing conserved

and essential protein complexes in Escherichia coli, *Nature*, vol. 433, no. 7025, pp. 531–537, 2005.

[5] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, et al., Functional profiling of the Saccharomyces cerevisiae genome, *Nature*, vol. 418, no. 6896, pp. 387–391, 2002.

[6] L. M. Cullen and G. M. Arndt, Genome-wide screening for gene function using RNAi in mammalian cells, *Immunol. Cell Biol.*, vol. 83, no. 3, pp. 217–223, 2005.

[7] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots, C. Marta, et al., Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery, *Mol. Microbiol.*, vol. 50, no. 1, pp. 167–181, 2003.

[8] 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.

[9] M. W. Hahn and A. D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol. Biol. Evol.*, vol. 22, no. 4, pp. 803–806, 2005.

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

[11] S. Wuchty and P. F. Stadler, Centers of complex networks, *J. Theor. Biol.*, vol. 223, no. 1, pp. 45–53, 2003.

[12] E. Estrada and J. A. Rodríguez-Velázquez, Subgraph centrality in complex networks, *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.*, vol. 71, no. 5Pt2, p. 056103, 2005.

[13] P. Bonacich, Power and centrality: A family of measures, *Am. J. Sociol.*, vol. 92, no. 5, pp. 1170–1182, 1987.

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

[15] M. Li, H. 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 Syst. Biol.*, vol. 6, p. 15, 2012.

[16] X. W. Tang, J. X. Wang, and Y. Pan. Identifying essential proteins via integration of protein interaction and gene expression data, in *Proc. 2012 IEEE Int. Conf. on Bioinformatics and Biomedicine*, Philadelphia, PA, USA, 2012, pp. 1–4.

[17] W. Peng, J. X. Wang, W. P. Wang, Q. Liu, F. X. Wu, and Y. Pan, Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks, *BMC Syst. Biol.*, vol. 6, p. 87, 2012.

[18] G. S. Li, M. Li, J. X. Wang, Y. H. Li, and Y. Pan, United neighborhood closeness centrality and orthology for predicting essential proteins, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 4, pp. 1451–1458, 2020.

[19] S. Y. Li, Z. P. Chen, X. He, Z. Zhang, T. Pei, Y. H. Tan, and L. Wang, An iteration method for identifying yeast essential proteins from weighted PPI network based on topological and functional features of proteins, *IEEE Access*, vol. 8, pp. 90792–90804, 2020.

[20] X. Y. Zhu, Y. C. Zhu, Y. H. Tan, Z. P. Chen, and L. Wang, An iterative method for predicting essential proteins based on multifeature fusion and linear neighborhood similarity, *Front. Aging Neurosci.*, vol. 13, p. 799500, 2021.

[21] B. H. Zhao, X. Han, X. E. Liu, Y. C. Luo, S. Hu, Z. H. Zhang, and L. Wang, A novel method to predict essential proteins based on diffusion distance networks, *IEEE Access*, vol. 8, pp. 29385–29394, 2020.

[22] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork, Dynamic complex formation during the yeast cell cycle, *Science*, vol. 307, no. 5710, pp. 724–727, 2005.

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

[24] M. Li, P. Ni, X. P. Chen, J. X. Wang, F. X. Wu, and Y. Pan, Construction of refined protein interaction network for predicting essential proteins, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 4, pp. 1386–1397, 2019.

[25] F. Y. Zhang, W. Peng, Y. F. Yang, W. Dai, and J. R. Song, A novel method for identifying essential genes by fusing dynamic protein-protein interactive networks, *Genes*, vol. 10, no. 1, p. 31, 2019.

[26] J. C. Zhong, C. Tang, W. Peng, M. Z. Xie, Y. S. Sun, Q. Tang, Q. Xiao, and J. H. Yang, A novel essential protein identification method based on PPI networks and gene expression data, *BMC Bioinformatics*, vol. 22, no. 1, p. 248, 2021.

[27] W. M. Sun, L. Wang, J. X. Peng, Z. Zhang, T. R. Pei, Y. H. Tan, X. Y. Li, and Z. P. Chen, A cross-entropy-based method for essential protein identification in yeast protein-protein interaction network, *Curr. Bioinf.*, vol. 16, no. 4, pp. 565–575, 2021.

[28] D. Sahoo, Boolean analysis of high-throughput biological datasets, PhD dissertation, Stanford University, Palo Alto, CA, USA, 2008.

[29] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, BioGRID: A general repository for interaction datasets, *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D535–D539, 2006.

[30] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H. W. Mewes, et al., The MIPS mammalian protein-protein interaction database, *Bioinformatics*, vol. 21, no. 6, pp. 832–834, 2005.

[31] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, et al., Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO), *Nucleic Acids Res.*, vol. 30, no. 1, pp. 69–72, 2002.

[32] R. Zhang and Y. Lin, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes, *Nucleic Acids Res.*, vol. 37, no. suppl_1, pp. D455–D458, 2009.

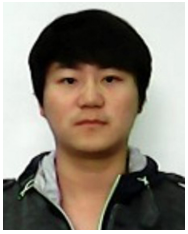
[33] G. Giaever and C. Nislow, The yeast deletion collection: A decade of functional genomics, *Genetics*, vol. 197, no. 2, pp. 451–465, 2014.

Jiancheng Zhong received the BEng (computer science and technology) and MEng (circuits and systems) degrees from Hunan Normal University, China in 2004 and 2007, respectively, and the PhD degree (computer application technology) from Central South University, Changsha, China in 2016. He is currently working as an associate professor and supervisor of master's students at the College of Information Science and Engineering, Hunan Normal University. His main research interests include bioinformatics and proteomics.

Yi Pan received the BEng (computer engineering) and MEng (computer engineering) degrees from Tsinghua University, Beijing, China in 1982 and 1984, respectively, and the PhD degree (computer science) from the University of Pittsburgh, USA in 1991. He is currently working as a dean and chair professor at the Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences Shenzhen, Guangdong, China. He has published over 250 academic papers in SCI-indexed journals, including more than 100 papers in top IEEE/ACM transactions/journals. His publications have been cited more than 19 000 times, and his current H-index is 88. He has been awarded IEEE Distinguished Achievement Award, IEEE Distinguished Service Award, IEEE Transactions Best Paper Award, IEEE and other international conference best paper awards many times, IBM Professor Award four times, Andrew Mellon Award, and other awards. His research interests include parallel and distributed processing systems, Internet technology, and bioinformatics.

Zuohang Qu received the BEng degree (electronic information science and technology) from Hunan University of Arts and Science, China in 2019. She is currently a master student at Hunan Normal University. Her research interests include machine learning, deep learning, and bioinformatics.

Ying Zhong is currently an undergraduate student at Hunan Normal University. His main research interest is bioinformatics.



Chao Tang received the BEng degree from Chuzhou University, China in 2017. He is currently a master student at Hunan Normal University, Changsha, China. His research interests include bioinformatics and machine learning.