Two new methods for identifying proteins based on the domain protein complexes and topological properties*

Pengli Lu† and JingJuan Yu
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, Gansu, P.R. China

Abstract

The recognition of essential proteins not only can help to understand the mechanism of cell operation, but also help to study the mechanism of biological evolution. At present, many scholars have been discovering essential proteins according to the topological structure of protein network and complexes. While some proteins still can not be recognized. In this paper, we proposed two new methods complex degree centrality (CDC) and complex in-degree and betweenness definition (CIBD) which integrate the local character of protein complexes and topological properties to determine the essentiality of proteins. First, we give the definitions of complex average centrality (CAC) and complex hybrid centrality (CHC) which both describe the properties of protein complexes. Then we propose these new methods CDC and CIBD based on CAC and CHC definitions. In order to access these two methods, different Protein-Protein Interaction (PPI) networks of Saccharomyces cerevisiae, DIP, MIPS and YMBD are used as experimental materials. Experimental results in networks show that the methods of CDC and CIBD can help to improve the precision of predicting essential proteins.

Keywords: Protein interaction network; Essential protein; Topology; Protein complex

1 Introduction

Protein is one of the main components of human life. Essential protein is defined as a protein which would result in the inability of the organism to survive when it is removed by a knockout mutation. Essential proteins are more conserved in biological evolution in comparison to non-essential proteins [1]. Not only can essential proteins help us understand the growth control system of cells, and then understand the mechanism of life, but also help the study of biological evolution mechanism [2]. Removing essential proteins can lead to fatal or infertility [3]. Determining the essentiality of proteins is of great significance to the research of system biology which provides valuable theories and methods for the diagnosis of diseases, drug design, etc. [4]. Therefore, identifying the essential protein is meaningful in biomedicine.

Previous methods for identifying essential proteins mainly used some biological experiments, including conditional knockouts [5], RNA interference [6], and single gene knockouts [7], coupled with the survival ability of infected organisms being tested. However, these biological experimental processes not only consume amounts of time and costs, but also require a lot of biological

---

*Supported by the National Natural Science Foundation of China (No.11361033) and the Natural Science Foundation of Gansu Province (No.1212RJZA029).

†Corresponding author. E-mail addresses: lupengli88@163.com (P. Lu), yujingjuanmercy@163.com (J. Yu).
resources. Nowadays, it has been a crucial research direction in the field of bioinformatics for predicting essential proteins from a large number of biological experiments by using computer technology theory and research methods.

Jeong H M et al. put forward that the essentiality of proteins is associated with the topological structure in protein interaction networks [8]. There are some species including S.cerevisiae, E.coli, C.elegans and D.melanogaster that have demonstrated the hubs in PPIs have more chance to be essential proteins [9]. Thus, we are working to investigate the importance of proteins in topologies to essential proteins. On the basis of network topology characteristics of nodes, there are many centrality measures to discover essential proteins. Some of them are global network characteristics, like betweenness centrality (BC) [11,38], eigenvector centrality (EC) [19], information centrality (IC) [20] and closeness centrality (CC) [13]. Others are local network features, such as degree centrality (DC) [10,14,15], subgraph centrality (SC) [16], local average centrality (LAC) [17] and topology potential-based method (TP) [34]. On the basis of network topology characteristics of edges, there are also some measures, including edge clustering coefficient (ECC) [35], and improved node and edge clustering coefficient (INEC) [36]. In recent years, many scholars have been working to identify proteins in combination with protein information, such as PeC which combines edge clustering coefficients with gene expression data correlation coefficients [24], esPOS which using gene expression information and subcellular localization information [21], SPP which based on sub-network partition and prioritization by integrating subcellular localization [12], extended pareto optimality consensus model (EPOC) that fuses neighborhood closeness centrality and Orthology information [39]. Go terms information can also be used to predict essential proteins such as RSG method in [25].

Apart from analyzing the essentiality of proteins from topological point of view and protein information, analyzing the characteristics from the perspective of protein complexes has become another direction of our study. Hart G T et al. found that the essential proteins are often determined by the protein complexes in which the protein is involved, rather than by a single protein [22]. Li et al. also prove that the frequency of the essential proteins appear in the complex would be more than that in the whole network [21,41]. To give examples, Luo J W et al. raised the local interaction density of binding protein complexes (LIDC) for predicting essential proteins [37]. Qin C et al. put forward the LBCC, a measure on the basis of both network topology features and protein complexes [18]. Li et al. proposed united complex centrality (UC) which combine the edge clustering coefficient and the freqencies of proteins appeared in complexes [23]. From the results of their experiences, we can see that the performances of these methods are better than using the pure topological methods.

Therefore, on the basis of the association with protein complexes information and topological properties, our two new novel methods complex degree centrality (CDC) and complex in-degree and betweenness definition (CIBD) are proposed. In order to describe the structural properties of protein complexes, we define CAC and CHC of a node v. Between the two indicators we put forward, one is called CDC which combine the node and its neighbors properties to describe the features for protein complexes, the other is called CIBD based on the features of protein complexes, local features and global properties in the network.

To assess the quality of CDC and CIBD methods, we apply them to different datasets of Saccharomyces cerevisiae, DIP, MIPS and YMBD. In order to obtain the performance of our proposed methods, we make comparisons by using some existing measures, including DC, BC,
LAC, SC, LBCC, EC, SoECC and UC which can gain the original paper from [10], [11], [17], [16], [18], [19], [28] and [23] respectively. In terms of the sensitivity, specificity, positive predictive value, negative predict value, F-measure, accuracy rate and the evaluation methods of “sorting-screening”, the precision-recall curves and jackknife, the results show that our two methods are more effective in determining the essentiality of proteins than existing measures.

2 Methods

2.1 Notation

An undirected simple graph $G(V, E)$ can be used to express a network of protein interaction. Proteins can be regarded as nodes set $V$ of a network and the connections between two proteins can be regarded as edges set $E$. The number of nodes and edges in a graph $G$ can be defined as $|V(G)|$ and $|E(G)|$ separately. The neighbor set of node $v$ is denoted by $N_v$, and its number can be represented as $|N_v|$. The induced subgraph of $G[S]$ is a subgraph of $G$ induced by the nodes set $S$.

2.2 Previously Proposed Centrality Measures

There are some centralities we need to understand.

- **Betweenness centrality ($BC$)** [11]
  \[
  BC(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}
  \]  
  (2.1)
  where $\sigma_{st}$ denotes the number of shortest paths between $s$ and $t$. $\sigma_{st}(v)$ denotes the number of shortest paths from $s$ to $t$ that pass through the node $v$.

- **In-degree centrality of complex ($IDC$)** [21]
  \[
  IDC(v) = \sum_{i \in \text{ComplexSet}(v)} IN - Degree(v)_i
  \]  
  (2.2)
  A subset of protein complexes that containing protein $v$ can be represented as $\text{ComplexSet}(v)$, the degree of node $v$ for the $i$th protein complex which belongs to $\text{ComplexSet}(v)$ can be represented as $IN - Degree(v)_i$.

- **LBCC method** [18]
  \[
  LBCC(v) = a \times \log Den_1(v) + b \times \log Den_2(v) + c \times \log IDC(v) + d \times \log BC(v)
  \]  
  (2.3)
  Specifically,
  \[
  Den_1(v) = \frac{2|E(H)|}{|V(H)|(|V(H)| - 1)}
  \]  
  (2.4)
  where the induced subgraph $G[N_v \cup \{v\}]$ can be represented as $H$.
  \[
  Den_2(v) = \frac{2|E(H)|}{|V(H)|(|V(H)| - 1)}
  \]  
  (2.5)
  where $M_u = \bigcup_{u \in N_v} N_u$, $H$ represents the induced subgraph $G[M_u \cup N_v \cup \{v\}]$. 

3
2.3 New Centrality: CDC and CIBD

The basic considerations of CDC and CIBD are as follows: (1) The essential proteins appear in complexes can be more frequency. (2) Both the node itself and its neighbors are critical to affect the essentiality. (3) The global topological is considered to be a factor in locating essential proteins. Consequently, we present two new definitions to judge the essentiality of proteins by combining the domain features of protein complex and the topological properties.

First, we present a new complex average central definition (CAC) for the neighbors of a node \( v \),

\[
CAC(v) = \sum_{u \in N_v} IDC(u) / |N_v| \tag{2.6}
\]

where \( \sum_{u \in N_v} IDC(u) \) represents the total values of IDC for all the neighbors of a node \( v \).

Then, we propose complex hybrid central definition (CHC) by combining the number of complexes for a node \( v \) with complex average central definition CAC,

\[
CHC(v) = N_{\text{complex}}(v) \cdot CAC(v) \cdot IDC^2(v) \tag{2.7}
\]

where \( N_{\text{complex}}(v) \) denotes the total number of complexes for a node \( v \).

Now, based on the two definitions that we described above, we propose these two new methods for estimating the essentiality of a node \( v \). One is complex degree centrality (CDC) which combine the node with its neighbors to describe the properties for protein complexes,

\[
CDC(v) = a \cdot CAC(v) + b \cdot IDC(v) \tag{2.8}
\]

where \( a, b \) are random parameters ranging from 1 to 10. After conducting plenty of experiments, we can get the best results of the method CDC when \( a \) and \( b \) are 1 and 4, respectively.

The other is complex in-degree and betweenness definition (CIBD) which combining CHC, \( Den_2 \) and BC, where the structural property of the protein complexes is described by CHC, the local feature is described by \( Den_2 \) and the global property is described by BC. Since the values of these measures are quite different, the data is normalized by logarithmic transformation,

\[
CIBD(v) = a \cdot \log(CHC(v)) + b \cdot \log(Den_2) + c \cdot \log(BC(v)) \tag{2.9}
\]

where \( a, b \) and \( c \) are random parameters ranging from 1 to 10. Under the amounts of experiments, we can get the best results of the method CIBD when \( a, b \) and \( c \) are 1, 3 and 1, respectively.

The description of CDC and CIBD algorithms are in Table 1.

3 Experimental data and assessment methods

3.1 Experimental data

In order to analyze the performance of these two algorithms of CDC and CIBD, experiments are conducted by using the protein interaction data of Saccharomyces cerevisiae because its proteins are more complete.

Three sets of PPI network data YDIP, YMIPS and YMBD are used. The DIP dataset is marked as YDIP network [26]; The MIPS dataset is marked as YMIPS network [25]; The YMBD
Table 1: Description of CDC and CIBD algorithms

| CDC and CIBD algorithms |
|-------------------------|
| **Input**: Undirected graph $G = (V(G), E(G))$ stands for a PPI network, $C = \{C_i = (V(C_i), E(C_i))|C_i \subseteq G\}$ represents complexes. |
| **Output**: The proteins list sorted by CDC, CIBD in a descending order. |

01: For each vertex $v \in V(G)$ do
02: For each $\forall C_i \in C$ do
03: calculate $IDC(v) = IDC(v) + IN - Degree(v)$,
//where $IN - Degree(v)$ is the value of $DC(v)$ in $i_{th}$ complex
04: For each vertex $v \in V(G)$ do
05: Find the neighbor nodes $N_{v1}$ of node $v$
//where $N_{v1}$ stands for the neighbor nodes set for node $v$
06: calculate $CAC(v)$ by Equation(6)
07: For each vertex $v_2 \in N_{v1}$ do
08: Find the neighbor nodes of $N_{v2}$
//where $N_{v2}$ stands for the neighbor nodes set for node $v_2$
which $v_2 \in N_{v1}$
09: calculate $Den_2$ by Equation(5)
10: For each vertex $v \in V(G)$ do
11: calculate $CHC(v)$ by Equation(7)
12: calculate and sort $CDC(v)$ by Equation(8)
13: calculate and sort $CIBD(v)$ by Equation(9)

network comes from the Mark Gerstein Lab website. In the protein network, all self-interaction and repetitive interaction are deleted as a data preprocessing of these PPIs. Specific properties for these three networks are presented in the Table 2. In the YDIP network, there are 5093 proteins and 24743 interactions, whose clustering coefficient is about 0.0973. YMIPS network includes 4546 proteins and 12319 interactions, whose clustering coefficient is about 0.0879. YMBD network includes 2559 proteins and 11835 interactions, whose clustering coefficient is about 0.4445.

The known essential protein is derived from four databases: MIPS [40], SGD (Saccharomyces Genome Database) [33], SGDP (Saccharomyces Genome Deletion Project) [4], and DEG (Database of Essential Genes) [27]. The protein complex set is from CM270 [40], CM425 [29], CYC408 and CYC428 datasets [30,31] which can gained from [21], containing 745 protein complexes (including 2167 proteins).

Table 2: Data details of the three protein networks: YDIP, YMIPS, YMBD

| Dataset | Proteins | Interactions | Average degree | Essential proteins | Clustering coefficient |
|---------|----------|--------------|----------------|-------------------|-----------------------|
| YDIP    | 5093     | 24743        | 9.72           | 1167              | 0.0973                |
| YMIPS   | 4546     | 12319        | 5.42           | 1016              | 0.0879                |
| YMBD    | 2559     | 11835        | 9.25           | 763               | 0.4445                |

3.2 Assessment methods

According to their values of CDC, CIBD and other eight prediction measures including $DC$, $BC$, $EC$, $SC$, $LAC$, $LBCC$, $SoECC$ and $UC$, proteins are sorted from high to low orders. First, we choose some number of top proteins in sequence as predictive essential proteins and then compare them with the real essential proteins. This allows us to know the quantity of true essential proteins. Therefore, the sensitivity $(SN)$, specificity $(SP)$, F-measure $(F)$, accuracy
(ACC), positive predictive value (PPV) and negative predictive value (NPV) can be calculated [28,29].

The following are the formulas for calculating these six statistical indicators.

Sensitivity:

\[ SN = \frac{TP}{TP + FN} \]

Specificity:

\[ SP = \frac{TN}{TN + FP} \]

Positive predictive value:

\[ PPV = \frac{TP}{TP + FP} \]

Negative predictive value:

\[ NPV = \frac{TN}{TN + FN} \]

F-measure:

\[ F = \frac{2 \times SN \times PPV}{SN + PPV} \]

Accuracy:

\[ ACC = \frac{TP + TN}{P + N} \]

where \( TP \) stands for the number of true essential proteins which are correctly selected as essential proteins. \( FP \) is the number of nonessential proteins which are incorrectly selected as essential. \( TN \) is the number of nonessential proteins which are correctly selected as nonessential. \( FN \) is the number of essential proteins which are incorrectly selected as nonessential. \( P \) and \( N \) stand for the sum number of essential and nonessential proteins, respectively.

4 Results

4.1 Comparison with other previously proposed measures

In this paper, to evaluate the efficiency and accuracy of different indicators in identifying essential proteins, we follow the principle of “sorting-screening” which has described as a flow chart in Fig. 1. Then we compare CDC and CIBD methods with other eight previous measures including DC, BC, EC, SC, LAC, LBCC, SoECC and UC in the three datasets. The algorithm for LBCC was implemented according to [18] which used the same datasets as ours. Other algorithms of DC, BC, EC, SC, LAC, SoECC and UC were implemented according to references [10], [11], [19], [16], [17], [28] and [23] respectively. Besides, we can also get these algorithms by using CytoNCA [42], which is a Cytoscape app for network centrality. We have mentioned the method of BC and LBCC in the Section Previously Proposed Centrality Measures. Now we give a brief description of other six indicators.

• Degree centrality (DC) [10]

\[ DC(v) = \text{deg}(v) \]  (4.1)

where \( \text{deg}(v) \) denotes the degree of a node \( v \).
Fig. 1 “sorting-screening” method

Fig. 2 The quantity of true essential proteins determined by CDC and other eight previously methods from the YDIP network.
Fig. 3 The quantity of true essential proteins determined by \textit{CDC, CIBD} and other eight previously methods from the YMIFS network.

Fig. 4 The quantity of true essential proteins determined by \textit{CIBD} and other eight previously methods from the YMBD network.
• Local average connectivity centrality (LAC) [17]

\[
LAC(v) = \frac{\sum_{u \in N_v} \text{deg}^C_v(u)}{|N_v|}
\]

(4.2)

where \(C_v\) is the subgraph induced by the node set \(N_v\) of \(G\) and \(\text{deg}^C_v(u)\) is the number of its neighbors in \(C_v\) for a node \(u \in N_v\).

• Subgraph centrality (SC) [16]

\[
SC(v) = \sum_{k=0}^{\infty} \frac{\mu_k(v)}{k!}
\]

(4.3)

where \(\mu_k(v)\) denotes the number of closed walks of length \(k\) which starts and ends at node \(v\).

• Eigenvector centrality (EC) [19]

\[
EC(v) = \alpha_{\text{max}}(v)
\]

(4.4)

where \(\alpha_{\text{max}}\) refers to the main eigenvector corresponding to the largest eigenvalue of the network adjacency matrix \(A\), and \(\alpha_{\text{max}}(v)\) represents the \(v\)th component of \(\alpha_{\text{max}}\).

• The sum of edge clustering coefficients (SoECC) [28]

\[
ECC_{v,u} = \frac{z_{v,u}}{\min(k_v - 1, k_u - 1)}
\]

(4.5)

where \(z_{v,u}\) is the number of triangles that includes the edge \(e(v,u)\) in network. \(k_v\) and \(k_u\) are the degrees of node \(u\) and node \(v\), respectively.

\[
\text{SoECC}(v) = \sum_{u \in N_v} ECC(v,u)
\]

(4.6)

where \(N_v\) denotes the set of all neighbors of node \(v\).

• United complex centrality (UC) [23]

\[
UC(v) = \sum_{u \in N_v} \left( \frac{f_u + 1}{f_M + 1} \times ECC_{v,u} \right)
\]

where \(f_u\) denotes the frequency of protein \(u\) appeared in the known protein complexes, \(f_M\) is the maximum frequency that a protein appeared in the known protein complexes.

Specifically, we compare CDC with other eight previous measures in YDIP and YMIPS networks, and compare CIBD with other eight previous measures using YMIPS and YMBD networks. Step one, we sort proteins from high to low order on the basis of their values of CDC, CIBD and other eight previous measures. Step two, we choose the top 100, 200, 300, 400, 500, and 600 proteins as predictive essential proteins, then compare them with the known essential proteins. Finally, we can get the quantity of true essential proteins among these predictive essential proteins. The experimental results of these measures are shown in Figs. 2-4.
From Fig. 2, the quantity of true essential proteins judged by CDC are 79, 152, 221, 272, 316 and 364 from the top 100 to the top 600, respectively, being the best among the seven methods in YDIP network. Besides CDC method, the method of LBCC also has well performance with 74, 135, 204, 261, 307 and 360 essential proteins correctly identified at the same level. By comparison, the true essential proteins determined by CDC method are increased by 5, 17, 17, 11, 9 and 4, respectively. Compared with other recent methods SoECC and UC, CDC also performs an excellent improvement. Moreover, the quantity of essential proteins are much more than previous method including BC, SC and EC. Although LAC has a good performance, our proposed CDC also has better results than it.

From Fig. 3, we can see that CIBD and CDC both perform better than DC, BC, SC, LAC, EC, SoECC and UC in YMIPS network, except for LBCC. The method of LBCC produces the best results at the top of 200, 500 and 600. CIBD performs the same as LBCC at the top of 100 and 300. At the top of 400, the performance of CDC and CIBD are both better than LBCC.

From Fig. 4, CIBD performs closely to the LBCC which gains the best performance at top 100, 200, 400 and 600. CIBD attains the best performance at the top of 300 and 500. We can also see these classical methods (DC, BC, SC, EC) perform not well in YMBD network. Hence, our new methods CDC and CIBD can determine much more true essential proteins in most cases.

4.2 Evaluation of six statistical methods and the precision-recall curves

To further judge these two indicators of CDC, CIBD as well as other eight identification measures, the six statistical methods mentioned in the Section Assessment methods are used. From the formulas, we can obtain some more profound meaning. The sensitivity (SN) measures the recognition ability of classifiers to identify correct essential proteins, the larger the value is, the better the classifier is. The specificity (SP) measures the recognition ability of classifiers to identify correct non-essential proteins. F-measures (F) stands for the harmonic mean of precision and sensitivity. The higher the accuracy (ACC) is, the better the classifier is. In conclusion, the values for these six statistical method can reflect the quality of indicators.

Hence, we sort proteins from high to low order on the basis of their values of these methods; Then we take the top 20 percent proteins into account as predictive essential proteins, the remaining 80 percent can be considered as candidates for nonessential proteins. Compared with the known essential protein dataset, we can obtain the values of TP, TN, FP and FN. According to the formulas, the values of these six statistical method would be calculated. On the three different networks, the comparisons among the values of CDC, CIBD and other eight measures are executed, showing in Table 3.

For YDIP network, these six statistic values for CDC are higher than other previous measures, which show that CDC has a better prediction accuracy. And the values of BC is the lowest, indicating it has poor performance. For YMIPS and YMBD networks, these six statistic values determined by CIBD are similar to LBCC which also has the ability to predict essential proteins accurately.

In addition, the Precision-Recall curve, a statistical method for evaluating stability, can be
Table 3: Comparison the results of sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure (F) and accuracy (ACC) of CDC, CIBD and other eight previous algorithms.

| Dataset | Methods | SN    | SP    | PPV   | NPV   | F     | ACC   |
|---------|---------|-------|-------|-------|-------|-------|-------|
| YDIP    | DC      | 0.363 | 0.825 | 0.416 | 0.789 | 0.388 | 0.706 |
|         | BC      | 0.281 | 0.798 | 0.354 | 0.738 | 0.313 | 0.632 |
|         | LAC     | 0.408 | 0.839 | 0.467 | 0.804 | 0.435 | 0.729 |
|         | SC      | 0.335 | 0.811 | 0.36  | 0.794 | 0.347 | 0.697 |
|         | LBCC    | 0.436 | 0.853 | 0.512 | 0.817 | 0.477 | 0.749 |
|         | EC      | 0.344 | 0.814 | 0.370 | 0.796 | 0.356 | 0.701 |
|         | SoECC   | 0.40  | 0.850 | 0.463 | 0.813 | 0.428 | 0.739 |
|         | UC      | 0.391 | 0.850 | 0.458 | 0.811 | 0.422 | 0.737 |
|         | CDC     | 0.448 | 0.868 | 0.515 | 0.835 | 0.487 | 0.764 |

| YMIPS   | DC      | 0.274 | 0.821 | 0.305 | 0.797 | 0.289 | 0.699 |
|         | BC      | 0.197 | 0.796 | 0.278 | 0.716 | 0.231 | 0.629 |
|         | LAC     | 0.287 | 0.825 | 0.321 | 0.801 | 0.303 | 0.705 |
|         | SC      | 0.139 | 0.782 | 0.155 | 0.759 | 0.146 | 0.638 |
|         | LBCC    | 0.430 | 0.866 | 0.480 | 0.841 | 0.454 | 0.769 |
|         | EC      | 0.123 | 0.774 | 0.155 | 0.723 | 0.137 | 0.610 |
|         | SoECC   | 0.281 | 0.814 | 0.325 | 0.781 | 0.302 | 0.686 |
|         | UC      | 0.271 | 0.812 | 0.314 | 0.778 | 0.291 | 0.682 |
|         | CDC     | 0.461 | 0.862 | 0.503 | 0.780 | 0.421 | 0.723 |
|         | CIBD    | 0.291 | 0.840 | 0.320 | 0.724 | 0.239 | 0.637 |

| YMBD    | DC      | 0.261 | 0.868 | 0.438 | 0.749 | 0.327 | 0.696 |
|         | BC      | 0.244 | 0.861 | 0.408 | 0.743 | 0.305 | 0.686 |
|         | LAC     | 0.247 | 0.862 | 0.413 | 0.744 | 0.309 | 0.688 |
|         | SC      | 0.191 | 0.840 | 0.320 | 0.724 | 0.239 | 0.637 |
|         | LBCC    | 0.373 | 0.910 | 0.617 | 0.789 | 0.465 | 0.760 |
|         | EC      | 0.219 | 0.851 | 0.366 | 0.734 | 0.274 | 0.672 |
|         | SoECC   | 0.266 | 0.835 | 0.422 | 0.715 | 0.326 | 0.657 |
|         | UC      | 0.274 | 0.838 | 0.434 | 0.718 | 0.336 | 0.662 |
|         | CIBD    | 0.347 | 0.910 | 0.581 | 0.777 | 0.434 | 0.745 |

Fig. 5 Precision and recall curves of CDC and other eight methods for YDIP network.
Fig. 6 Precision and recall curves of CDC, CIBD and other eight methods for YMIPS network.

Fig. 7 Precision and recall curves of CIBD and other eight methods for YMBD network.
used for CDC and CIBD methods and other previous eight measures which defined as follows:

\[ Precision(n) = \frac{TP(n)}{TP(n) + FP(n)} \]

\[ Recall(n) = \frac{TP(n)}{TP(n) + FN(n)} \]

where the definitions of \( TP \), \( FP \), \( FN \) are depicted in the Assessment method Section. The results are revealed in Figs. 5-7. In YDIP network, our method of CDC has better performance than the other methods. In YMIPS and YDIP networks, the performance of CDC and CIBD are similar to the performance of LBCC.

4.3 Evaluation of jackknife methodology

Holman et al. developed the jackknife methodology which is an effective universal prediction method [32]. The X-axis represents the quantity of selected predictive essential proteins after sequencing, and the Y-axis represents the quantity of true essential proteins in the selected proteins. The area under the curve reflects the performance of each method. The larger the area under the curve is, the better the centrality is.

First, according to the predicted value, proteins are sorted in descending order. And then we choose predictive essential proteins of top 600 for each dataset. Last, the jackknife curve is drawn based on the accumulation quantity of real essential proteins.

From Fig. 8, it can be seen that the prediction efficiency of CDC is higher than that of other centrality measures on the YDIP network. From Fig. 9, it is shown that CDC and CIBD exhibit performances resemble to that of LBCC and better than those of all the other methods including DC, BC, LAC, SC and EC, SoECC and UC on the YMIPS network. From the
Fig. 9 The performances of CDC, CIBD and other eight centrality measures on the YMIPS network are evaluated by a jackknife methodology.

Fig. 10 The performances of CIBD and other eight centrality measures on the YMBD network are evaluated by a jackknife methodology.
YMBD network, we can get the same results as shown in Fig. 10. Consequently, the jackknife curves reveal that our methods CDC and CIBD both are effective approaches for predicting essential proteins.

5 Conclusion

Identifying essential proteins in protein networks is an indispensable point in the post-genomic era. Improving the recognition rate of essential proteins is a challenging task. At present, plenty of centrality algorithms have been proposed to determine the essentiality of proteins, most of them focus on the analysis and mining of node topology characteristics. In this paper, on the basis of the combination of the local features of protein complexes and topological properties, two new methods are proposed which named as CDC and CIBD. We apply them to different datasets YDIP, YMIPS and YMBD. Then we compare the quantity of true essential proteins predicted by CDC, CIBD and other eight proposed methods, containing DC, BC, LAC, SC, LBCC, EC, SoECC and UC. The results show that CDC and CIBD perform well in most cases. By using the methods of the six statistical, the precision-recall curve and jackknife, we can find that our proposed methods of CDC and CIBD have the ability to improve the accuracy in predicting essential proteins. In future work, deepening the mining of protein biological function and biological significance can be another direction to find the essential proteins.

References

[1] Fraser H B, Hirsh A E, et al., Evolutionary Rate in the Protein Interaction Network, Science, 296(5568):750-752, 2002.

[2] Xu B, Guan J, Wang Y, et al., Essential protein detection by random walk on weighted protein-protein interaction networks, IEEE/ACM Trans Comput Biol Bioinform, PP(99):1-1, 2017.

[3] Winzeler E A, Shoemaker D D, Astromoff A, Liang H, Anderson K, Andre B, et al., Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis, Science, 285 (5429):901-906, 1999.

[4] Wang Y, Sun H, Du W, Blanzieri E, Viero G, Xu Y, et al., Identification of essential proteins based on ranking edge-weights in protein-protein interaction networks, PloS One, 9(9):e108716, 2014.

[5] Roemer T, Jiang B, Davison J, Ketela T, Veillette K, et al., Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery, Mol Microbiol, 50:167-181, 2003.

[6] Cullen L M, Arndt G M, Genome-wide screening for gene function using RNAi in mammalian cells, Immunol Cell Biol, 83:217-223, 2005.

[7] Giaever G, Chu A M, Ni L, et al., SGD: Functional profiling of the saccharomyces cerevisiae genome, Nature, 418(6896):387-391, 2002.

[8] Jeong H M, Mason S P, Albert B, et al., Lethality and centrality in protein networks, Nature, 411:41-42, 2001.

[9] Zhao B H, Wang J X, Li M, et al., Prediction of Essential Proteins Based on Overlapping Essential Modules, IEEE Transactions on Nanobioscience, 13(4):415-424, 2014.

[10] Hahn M W, Kern A D, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Molecular Biology and Evolution, 22(4):803-806, 2005.
[11] Freeman L C, A set of measures of centrality based on betweenness, Sociometry, 40(1):35-41, 1977.
[12] Li M, Li W, Wu F X, et al., Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information, Journal of Theoretical Biology, 2018.
[13] Wuchty S, Stadler P F, Centers of complex networks, Journal of Theoretical Biology, 223(1):45-53, 2003.
[14] Lin C C, Juan H F, Hsiang J T, Hwang Y C, Mori H, Huang H C, Essential core of protein-protein interaction network in escherichia coli, Journal of Proteome Research, 8(4):1925-1931, 2009.
[15] Liang H, Li W H, Gene essentiality, gene duplicability and protein connectivity in human and mouse, Trends in Genetics, 23(8):375-378, 2007.
[16] Estrada E, Juan A, Subgraph centrality in complex networks, Physical Review E, 71(5):1-9, 2005.
[17] Li M, Wang J, Chen X, et al., A local average connectivity-based method for identifying essential proteins from the network level, Computational Biology and Chemistry, 35(3):143-150, 2011.
[18] Qin C, Sun Y, Dong Y, A new method for identifying essential proteins based on network topology properties and protein complexes, PLOS ONE, 11(8):e0161042, 2016.
[19] Bonacich P, Power and centrality: a family of measures, American Journal of Sociology, 92(5):1170-1182, 1987.
[20] Stephenson K, Zelen M, Rethinking centrality: methods and examples, Soc Networks, 11:1-37, 1989.
[21] Zhang Z P, Ruan J S, Gao J Z, et al., Predicting essential proteins from protein-protein interactions using order statistics, Journal of Theoretical Biology, 480:274-283, 2019.
[22] Hart G T, Lee I, Marcotte E M. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. Bmc Bioinformatics, 8(1):236-0, 2007.
[23] Li M, Lu Y, Niu Z, et al., United complex centrality for identification of essential proteins from PPI networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(2):370-380, 2017.
[24] Li M, Zhang H H, Fei Y P, Essential protein discovery method based on integration of PPI and gene expression data, Journal of Central South University, 44(3):1024-1029, 2013.
[25] Lei X, Zhao J, et al., Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets, Knowledge-Based Systems, 2018.
[26] Xenarios I, Lukasz S, et al., DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Research, 30(1):303-305, 2002.
[27] Zhang R, Liu Y, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes, Nucleic Acids Res, 37(suppl 1):D455-D458, 2009.
[28] Wang J, Li M, Wang H, Pan Y, Identification of essential proteins based on edge clustering coefficient, Transactions on Computational Biology and Bioinformatics, 9(4):1070-1080, 2012.
[29] Friedel C C, Krumsieck J, Zimmer R, International Conference on Research in Computational Molecular Biology. Springer-Verlag, 2008.
[30] Pu S, Wong J, Turner B, Cho E, Wodak S J, Up-to-date catalogues of yeast protein complexes, Nucleic Acids Research, 37(3):825-831, 2009.
[31] Pu S, Vlasblom J, Emili A, et al., Identifying functional modules in the physical interactome of saccharomyces cerevisiae, Proteomics, 7(6):944-960, 2010.
[32] Holman A G, Davis P J, Foster J M, et al., Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi, Bmc Microbiology, 9(1):1-14, 2009.
[33] Cherry J M, Adler C, Ball C A, et al., SGD: saccharomyces genome database, Nucleic Acids Research, 26(1):73-79, 1998.
[34] Li M, Lu Y, Wang J, et al., A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(2):372-383, 2015.

[35] Radicchi F, Castellano C, Cecconi F, et al., Defining and identifying communities in networks, Proceedings of the National Academy of Sciences of the United States of America, 101(9):2658-2663, 2003.

[36] Zhu Y, Wu C, Identification of essential proteins using improved node and edge clustering coefficient, Proceedings of the 37th Chinese Control Conference, 2018.

[37] Luo J W, Qi Y, Identification of essential proteins based on a new combination of local interaction density and protein complexes, PLOS ONE, 10(6):e0131418, 2015.

[38] Joy M P, Brock A, Ingber D E, et al., High-betweenness proteins in the yeast protein interaction network, Journal of Biomedicine and Biotechnology, 2005(2):96, 2014.

[39] Li G, Li M, Wang J, et al., United neighborhood closeness centrality and orthology for predicting essential proteins, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018.

[40] Mewes H W, Amid C, Arnold R, et al., MIPS: analysis and annotation of proteins from whole genomes, Nucleic Acids Research, 34(Database issue):169-72, 2004.

[41] Pereira-Leal J B, Benjamin A, Peregrin-Alvarez J M, et al., An Exponential Core in the Heart of the Yeast Protein Interaction Network, Molecular Biology and Evolution, 2015.

[42] Tang Y, Li M, Wang J, et al., CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks, Bioystems, 127:67-72, 2015.