Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 18191–18206.
DOI: 10.3934/mbe.2023808
Received: 24 April 2023
Revised: 19 August 2023
Accepted: 23 August 2023
Published: 21 September 2023

Research article

Key protein identification by integrating protein complex information and multi-biological features

Yongyin Han [1,2,*], **Maolin Liu**[1] **and Zhixiao Wang**[1]

[1] School of Computer Science and Technology, China University of Mining and Technology, China
[2] Xuzhou College of Industrial Technology, China

* **Correspondence:** Email: 371418658@qq.com.

Abstract: Identifying key proteins based on protein-protein interaction networks has emerged as a prominent area of research in bioinformatics. However, current methods exhibit certain limitations, such as the omission of subcellular localization information and the disregard for the impact of topological structure noise on the reliability of key protein identification. Moreover, the influence of proteins outside a complex but interacting with proteins inside the complex on complex participation tends to be overlooked. Addressing these shortcomings, this paper presents a novel method for key protein identification that integrates protein complex information with multiple biological features. This approach offers a comprehensive evaluation of protein importance by considering subcellular localization centrality, topological centrality weighted by gene ontology (GO) similarity and complex participation centrality. Experimental results, including traditional statistical metrics, jackknife methodology metric and key protein overlap or difference, demonstrate that the proposed method not only achieves higher accuracy in identifying key proteins compared to nine classical methods but also exhibits robustness across diverse protein-protein interaction networks.

Keywords: Key protein; subcellular localization; GO similarity; complex participation

1. Introduction

Proteins within an organism can be classified into two categories: non key proteins and key proteins. Key proteins play crucial roles throughout the cell cycle, and their absence can result in infertility, biological dysfunction and even fatality. Furthermore, key proteins have been implicated in the pathogenesis of various diseases [1,2]. Consequently, the identification of key proteins has become a highly relevant research area within the field of bioinformatics [3–5]. Traditional experimental approaches for identifying key proteins tend to be expensive, cumbersome, inefficient and limited in scope. Conversely, key protein identification methods based on protein-protein interaction (PPI) networks [6] offer a cost-effective, efficient and reliable alternative [7].

The protein-protein interaction (PPI) network exhibits a scale-free nature, characterized by uneven internal connectivity. A small subset of nodes within the network possesses a large number of connections, often corresponding to key proteins. Consequently, topological centrality methods such as degree centrality, information centrality, betweenness centrality, feature vector centrality, subgraph centrality, local average connection centrality, and neighborhood centrality have been utilized for key protein identification [1, 2, 6]. However, the accuracy of these centrality-based methods is contingent on the quality of the PPI network, which is prone to incompleteness and includes numerous false positive and false negative data due to experimental limitations [8]. To address this challenge, several approaches have been proposed, combining biological characteristics with network analysis. For instance, Li et al. introduced the PeC method [9], which integrates the topological structure of the PPI network with gene expression profiles to identify key proteins. Peng et al. developed the UDoNC method [10], which leverages protein-domain characteristics to identify key proteins. Shang et al. proposed the DLAC method [11], which incorporates RNA sequence data to enhance the accuracy of key protein prediction. Additionally, some researchers have combined biological characteristics with random walk methods to uncover key proteins [12]. Moreover, the JDC method [13] offers a dynamic threshold approach to binarize gene expression data based on PPI network information and gene expression profiles.

Furthermore, it has been observed that key proteins tend to have a higher propensity for participation in protein complexes compared to non-key proteins. To capitalize on this characteristic, Luo et al. introduced the LIDC method [14], which predicts key proteins by considering the local interaction density and internal degree of the protein complex. Building upon this work, Qin et al. enhanced the LIDC method and proposed the LBCC method [15], which incorporates betweenness centrality to identify key proteins. The UC method [16], on the other hand, utilizes protein frequency information within the complex to identify key proteins. Shifting gears slightly, the Modality-DTA method [17] presents a novel deep learning approach for drug-target interaction prediction, leveraging the multimodal nature of both drugs and targets to enhance prediction accuracy.

The accuracy of key protein identification based on a single biological characteristic is often compromised due to variations in space-time dimensions and the influence of different physical and chemical environments [18]. Consequently, an increasing number of researchers are exploring the integration of multiple biological characteristics to improve the accuracy of key protein mining. For instance, the TEO method [19] incorporates GO annotation information, gene expression data, and network topology to identify key proteins. Similarly, the JTBC method [20] utilizes both gene expression information and domain information in the process of mining key proteins. By leveraging these diverse biological characteristics, these methods aim to enhance the accuracy and reliability of key protein identification.

While existing methods have made progress in key protein identification, they still face certain limitations. First, these methods often overlook the crucial aspect of subcellular localization information. In reality, the importance and criticality of proteins can vary depending on their specific subcellular locations. Second, complex information plays a significant role in key protein identification. However, existing methods fail to account for the impact of proteins outside the complex that interact with proteins within the complex, thereby neglecting their potential influence on complex participation.

To address the aforementioned issues, this paper presents a novel method called CIBF (protein complex information and multi-biological features) for identifying key proteins. CIBF is designed to overcome the limitations of existing approaches by integrating complex information and multiple biological characteristics. The key contributions of this method can be summarized as follows:

1) The subcellular localization information plays a crucial role in determining the key index for various cellular locations. Then, this index is integrated with the neighborhood information of protein nodes to determine the subcellular localization centrality of each protein. By combining subcellular localization centrality with the protein's surrounding network, this method accurately assesses the protein's significance within its specific subcellular location.
2) The method introduces an edge clustering coefficient that considers the difference in public neighbor participation to quantitatively depict the interaction edge weight between proteins. Additionally, the GO similarity between protein nodes is computed using GO information. Biological characteristics are incorporated into the edge weight, enhancing its relevance. Furthermore, the method proposes a topological centrality measure with GO similarity weighting.
3) The proposed method takes into full consideration the interaction between proteins outside the complex and those inside the complex. It accurately determines the centrality of protein complex participation by integrating two key factors: the in-degree of the complex for protein nodes and the frequency of complex participation.

The identification results of key proteins in different PPI networks show that the CIBF method can effectively identify key proteins and has fine stability.

2. Materials and method

In this paper, we present a novel method for identifying key proteins by integrating complex information and multiple biological characteristics. Our approach involves a comprehensive evaluation of protein nodes, considering their centrality in subcellular location, topological centrality weighted by GO similarity, and centrality of complex participation. By examining key proteins from these diverse dimensions, we aim to enhance the accuracy of key protein mining within the protein-protein interaction (PPI) network. This integrated methodology provides a more precise and comprehensive approach to identify key proteins.

2.1. Subcellular localization centrality

Subcellular localization information identifies the location of proteins in cells, which is an important biological characteristic of proteins in space. By analyzing the subcellular localization distribution of proteins, key proteins appear more frequently in some locations than in others. Based on this phenomenon, subcellular localization information can be used to judge the spatial location centrality of proteins. There are 11 subcellular localization regions, as shown in Table 1. The key coefficient of subcellular localization region k is expressed by $csl(k)$, and its calculation method is as follows:

$$csl(k) = \frac{nep(k)}{sep} \tag{2.1}$$

Among them, $nep(k)$ represents the number of key proteins in the subcellular localization region k, and sep represents the total number of key proteins.

Table 1 shows the distribution of key proteins in 11 subcellular localization regions and the key coefficient of subcellular localization regions calculated by formula 2.1. It can be found that key proteins appear more frequently in nucleus, mitochondrion, endoplasma, cytosol and so on. In view of this, this paper calculates the spatial location centrality of protein v according to the key coefficient of subcellular location region:

$$CSL(v) = \frac{\sum_{v\in sl} csl(k)_{k\in[1,11]} \left(1 + d(v)_{k,k\in[1,11]}\right)}{n_k} \tag{2.2}$$

Among them, $csl(k)_{k\in[1,11]}$ represents the key coefficient of protein v in 11 subcellular regions, and sl represents the subcellular localization region, $d(v)_{k,k\in[1,11]}$ represents the number of neighbor proteins of protein v in subcellular localization region k, and n_k represents the number of subcellular localization regions of protein v.

Table 1. Subcellular localization and Coefficient of subcellular localization.

Subcellular localization region	Number of key proteins(nep)	Key coefficient of subcellular localization(csl)
Plasma	61	0.0475
Cytosol	228	0.1774
Endosome	26	0.0202
Endoplasmic	152	0.1183
Extracellular	2	0.0016
Golgi	65	0.0506
Mitochondrion	193	0.1502
Nucleus	783	0.6093
Peroxisome	4	0.0331
Vacuole	26	0.0202
Cytoskeleton	99	0.0770

2.2. *Topological centrality of GO similarity weighted*

In this paper, we propose the utilization of GO similarity as a means to introduce weighted topological centrality as an additional indicator for identifying key proteins. The gene ontology (GO) framework is employed to describe the biological characteristics of genes and their corresponding products. The relationship structure within GO is organized in a tree-like structure, where nodes closer to the root encompass broader descriptions, while nodes farther away convey more specific details. Our method calculates the GO functional similarity between proteins within the PPI network. The presence of common GO annotations indicates a closer relationship and enhances the reliability of the edges in the PPI network. Additionally, recognizing that the GO functional similarity can be influenced by the number of GO annotations associated with a protein, we introduce an adjustment factor to account for this effect. The specific calculation method of GO functional similarity is:

$$AGO_{\text{sim}}(u, v) = \frac{|GO_u \cap GO_v|^2}{(|GO_u|\,\sigma_u) \times (|GO_v|\,\sigma_v)} \tag{2.3}$$

GO_u and GO_v represent GO annotation of protein u and v, σ_u and σ_v represents the corresponding adjustment factor, which punishes the protein with less GO annotations and rewards the protein with more GO annotations. The calculation method is as follows:

$$\sigma_{i=u,v} = \frac{\overline{GO}}{GO_i} \tag{2.4}$$

$\overline{GO}$ represents the average number of GO annotations in the PPI network.

In PPI network, the edge clustering coefficient (ECC) evaluates the connection strength between two proteins from the topological structure. The calculation method is as follows:

$$\text{ECC}(u, v) = \frac{z(u, v)}{\min(d_u - 1, d_v - 1)} \tag{2.5}$$

Among them, $z(u, v)$ represents the number of common neighbor nodes of protein u, v, and d_u and d_v represent the degree value of protein u, v.

The traditional edge clustering coefficient does not consider the difference of the participation degree of the public neighbors in the edge $e(u, v)$. In this paper, the public neighbor participation $\sum p_i$ is introduced to calculate the participation of different public neighbors of edge $e(u, v)$. The calculation method is as follows:

$$p_i = \frac{2}{d_i} \tag{2.6}$$

i represents the common neighbor of protein u, v, and d_i represents the degree value of protein i itself. From this, we can get the public neighbor difference edge clustering coefficient $DnECC(u, v)$, which is calculated as follows:

$$DnEC(u, v) = \frac{\sum p_i}{\min(d_u - 1, d_v - 1)} \tag{2.7}$$

$\sum p_i$ represents the sum of the participation of all common neighbor proteins of edge $e(u, v)$. The greater the value of $DnECC(u, v)$, the higher the connection strength between protein nodes.

The topological centrality weighted by GO similarity is obtained by fusing GO similarity and common neighbor difference edge clustering coefficient:

$$CBT(v) = \sum_{v=v} (AGO_{\text{sim}}(u, v) + \text{DnECC}(u, v)) \tag{2.8}$$

N represents the neighbor protein set of protein v.

2.3. Complex participation centrality

Complex information helps to identify key proteins. However, key proteins may appear inside or outside the complex. The existing methods do not consider the influence of proteins outside the complex that interact with proteins in the complex on the participation of the complex. As shown in Figure 1, although protein 8 is not inside complex A and B, it interacts with proteins 1 and 4 inside complex A, as well as proteins 5 and 6 inside complex B. The calculation of protein complex participation should be considered. In view of this, this paper distinguishes the two types of proteins

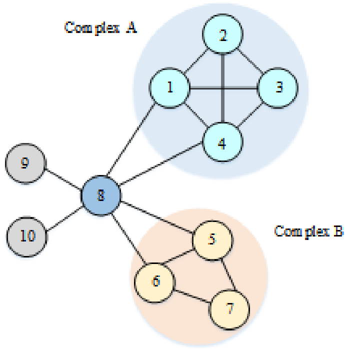

Figure 1. The relationship between protein and complex.

inside and outside the complex to more accurately evaluate the protein complex participation, and the calculation method is as follows:

$$CPC(v) = \begin{cases} f_{\text{in}}\left(\frac{\sum d_{in-pc}}{n_{\text{in}}}\right), v \in PC \\ \frac{d_{\text{out - in}}}{n_{\text{out}}}, v \notin PC \end{cases} \tag{2.9}$$

d_{in-pc} represents the in-degree of protein v in the complex, n_{in} represents the number of times protein v appears in the complex, n_{out} represents the number of complexes connected by protein v, d_{out-in} in represents the number of connections between protein v and the protein in the complex, f_{in} represents the frequency of protein v appearing in the complex, and its calculation method is as follows:

$$f_{in} = 1 + \frac{n_{in}}{n_M} \tag{2.10}$$

n_M represents the maximum number of proteins appears in the complex.

2.4. Method description

In this paper, we combine subcellular localization centrality, GO similarity weighted topological centrality and complex participation centrality with a linear weighted model to obtain a comprehensive protein criticality evaluation method:

$$CIBF(v) = \alpha\frac{CSL(v)}{MAX(CSL)} + \beta\frac{CBT(v)}{MAX(CBT)} + (1-\alpha-\beta)\frac{CPC(v)}{MAX(CPC)} \tag{2.11}$$

α, β and (1-α-β) are used to adjust the contribution of each part to the protein criticality. The experiment part will discuss value of α, β and (1-α-β).

The CIBF method is described as follows:

3. Experimental results and analysis

3.1. Experimental data

In this paper, nine representative key protein identification methods are selected: DC [21], BC [22], SC [23], NC [24], PeC [9], LBCC [15], UC [16], TEO [19], CENC [25] making comparisons,

Table 2. PPI networks.

PPI network	Number of proteins	Number of interactions	Density
DIP	5093	24743	0.0018
Krogan	2674	7075	0.0020
MIPS	4546	12319	0.0012

to verify the effectiveness of the CIBF method. The experiment uses three PPI networks: DIP [26], Krogan [27] and MIPS [28]. Details are shown in Table 2.

GO data used in the experiment comes from gene ontology database [29] and subcellular location data from COMPARTMENTS database [30]. Key protein data for matching were integrated from DEG [31], MIPS [32], SGD [33] and SGDP [13].

Algorithm 1: CIBF Method

```
Input: PPI network G=(V, E); Subcellular localization data; GO data; Protein complex data;
       Parameter $\alpha, \beta$
Output: The rank list of protein nodes
for i=1 to n do
    Calculate CSL(i) by formula (2); //subcellular localization centrality
end
for each e∈E do
    Calculate AGOsim of e by formula (3);
    Calculate DnECC of e by formula (7);
end
for i=1 to n do
    Calculate CBT(i) by formula (8); //topological centricity of GO similarity weighted
end
for i=1 to n do
    Calculate CPC(i) by formula (9); //complex participation centrality
end
for each v in G do
    Calculate CIBF(v) by formula (11); //final centrality
end
sort the protein nodes according to the value of CIBF(v) in descending order;
return the rank list of protein nodes;
```

3.2. Evaluation metrics

This paper uses three evaluation metrics:

(1) Traditional statistical metrics. This paper uses the traditional evaluation metrics as shown in Table 3 for evaluation.

Among them, SN represents the proportion of correctly predicted key proteins in the total number of key proteins, and SP represents the proportion of correctly predicted non-key proteins in the to-

Table 3. Traditional evaluation metrics.

Evaluation metrics	Calculation method
SN	TP/(TP+FN)
SP	TN/(TN+FP)
PPV	TP/(TP+FP)
F–measure	(2*SN*PPV)/(SN+PPV)
ACC	(TP+TN)/(P+N)

tal number of non-key proteins. PPV represents the correct proportion of all key proteins predicted. F–measure is calculated from SN and PPV, which is a comprehensive measure of SN and PPV. It can more evenly evaluate the overall performance of different methods under SN and PPV metric. ACC is used to evaluate the overall accuracy of each method in identifying key proteins and non-key proteins.

(2) Jackknife Methodology metric.

It is used to evaluate the identification ability and stability of different methods for key proteins.

(3) Overlap/difference analysis of key proteins.

This evaluation metric mainly determines the performance of each method by analyzing the overlap and difference of proteins identified by different methods.

3.3. Parameter analysis

In CIBF method, α, β and (1-α-β) are used to adjust the contribution of spatial location centrality, biological topology centrality and complex participation centrality to protein criticality. This section analyzes the influence of different parameter settings on key protein identification performance through experiments. When $\alpha = 1$, only the spatial location centrality of protein is considered. When $\beta = 1$, only the biological topological centrality of protein is considered. When (1-α-β) = 1, only the centrality of protein complex participation is considered. The results on DIP, Krogan and MIPS data sets show that when $\alpha = 0.2$, $\beta = 0.4$, the number of key proteins correctly identified by CIBF method is the largest. Therefore, in the experiments of this paper, $\alpha = 0.2$, $\beta = 0.4$, (1-α-β) = 0.4.

3.4. Ablation analysis

The CIBF method involves three aspects in identifying key proteins: spatial location centrality, biological topological centrality and complex participation centrality. Through ablation experiments, this section demonstrates the identification ability of key proteins when only one or two factors are considered, providing the need for each component.

Table 4–6 show the F–measure value when only a single factor is considered in DIP, Krogan and MIPS network.

Table 4. Single factor F–measure in DIP network.

Factor	F–measure
CSL(v)	0.449
CBT(v)	0.457
CPC(v)	0.460

Table 5. Single factor F–measure in Krogan network.

Factor	F–measure
CSL(v)	0.459
CBT(v)	0.472
CPC(v)	0.467

Table 6. Single factor F–measure in MIPS network.

Factor	F–measure
CSL(v)	0.442
CBT(v)	0.453
CPC(v)	0.452

Table 7–9 show the changes in F–measure value when considering two factors in the DIP network, where the horizontal and vertical coordinates represent the proportion of each factor.

Table 7. F–measure in DIP of CSL and CBT factor.

A \ B	0	0.2	0.4	0.6	0.8	1.0
0	/	0.457	0.457	0.457	0.457	0.457
0.2	0.449	0.461	0.462	0.462	0.461	/
0.4	0.449	0.459	0.461	0.462	/	/
0.6	0.449	0.458	0.459	/	/	/
0.8	0.449	0.456	/	/	/	/
1.0	0.449	/	/	/	/	/

In Table 7, A and B represent CSL and CBT respectively.

Table 8. F–measure in DIP of CSL and CPC factor.

A \ B	0	0.2	0.4	0.6	0.8	1.0
0	/	0.460	0.460	0.460	0.460	0.460
0.2	0.449	0.464	0.466	0.467	0.464	/
0.4	0.449	0.463	0.464	0.468	/	/
0.6	0.449	0.462	0.466	/	/	/
0.8	0.449	0.459	/	/	/	/
1.0	0.449	/	/	/	/	/

These experiments demonstrate that the three aspects involved in CIBF method are helpful in improving the accuracy of key proteins identification.

3.5. Statistical metric analysis results

This section uses four statistical metrics, SN, SP, F–measure and ACC, to comprehensively evaluate the performance of CIBF method. Table 10–12 shows the SN, SP, F–measure and ACC values of

Table 9. F–measure in DIP of CBT and CPC factors

B A	0	0.2	0.4	0.6	0.8	1.0
0	/	0.460	0.460	0.460	0.460	0.460
0.2	0.457	0.471	0.472	0.470	0.468	/
0.4	0.457	0.467	0.471	0.470	/	/
0.6	0.457	0.466	0.465	/	/	/
0.8	0.457	0.464	/	/	/	/
1.0	0.457	/	/	/	/	/

different methods in three PPI networks.

Table 10. Evaluation results of DIP network.

Method	SN	SP	F–measure	ACC
DC	0.429	0.830	0.411	0.783
BC	0.371	0.813	0.355	0.712
SC	0.400	0.822	0.359	0.725
NC	0.431	0.832	0.418	0.739
PeC	0.423	0.829	0.414	0.736
LBCC	0.466	0.841	0.454	0.755
UC	0.462	0.840	0.449	0.753
TEO	0.492	0.849	0.471	0.767
CENC	0.422	0.828	0.418	0.735
CIBF	**0.500**	**0.851**	**0.481**	**0.770**

Table 11. Evaluation results of Krogan network.

Method	SN	SP	F–measure	ACC
DC	0.406	0.754	0.408	0.652
BC	0.316	0.716	0.318	0.599
SC	0.457	0.733	0.360	0.623
NC	0.412	0.756	0.415	0.655
PeC	0.410	0.755	0.412	0.654
LBCC	0.464	0.778	0.466	0.686
UC	0.423	0.761	0.425	0.662
TEO	0.451	0.772	0.453	0.678
CENC	0.420	0.759	0.423	0.660
CIBF	**0.491**	**0.789**	**0.489**	**0.702**

It can be seen that the CIBF method has the highest accuracy in DIP, Krogan and MIPS networks, which are 0.770, 0.702 and 0.753 respectively. In DIP network, the F–measure of CIBF method is 2.1 higher than the second method TEO. In Krogan network, F–measure of CIBF method is 4.9 higher than the second method LBCC. In MIPS network, the CIBF method also has the highest F–measure

Table 12. Evaluation results of MIPS network.

Method	SN	SP	F–measure	ACC
DC	0.252	0.785	0.266	0.666
BC	0.249	0.784	0.263	0.664
SC	0.139	0.752	0.146	0.615
NC	0.281	0.793	0.297	0.679
PeC	0.314	0.803	0.331	0.693
LBCC	0.430	0.836	0.448	0.745
UC	0.348	0.812	0.365	0.709
TEO	0.427	0.835	0.446	0.744
CENC	0.321	0.805	0.340	0.696
CIBF	**0.448**	**0.841**	**0.466**	**0.753**

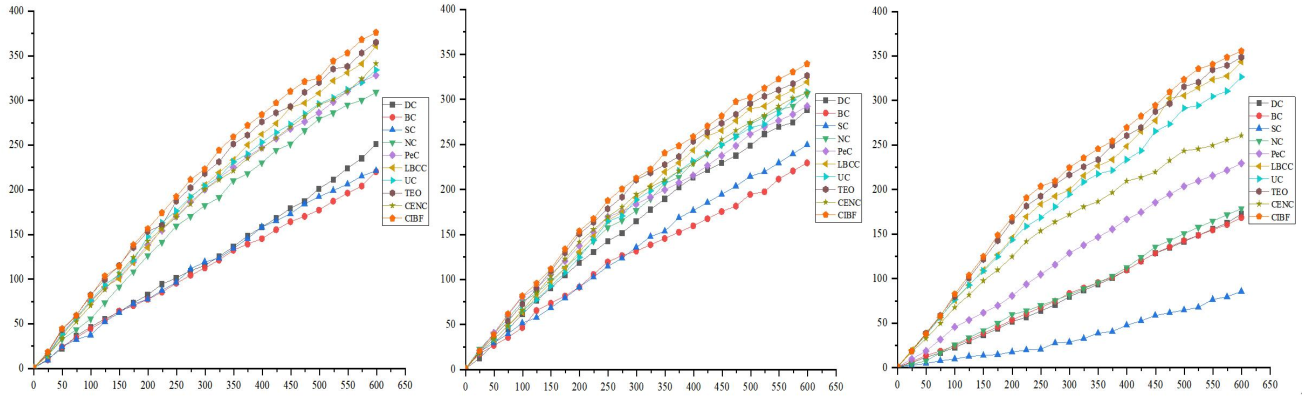

Figure 2. Jackknife results of DIP, Krogan and MIPS networks.

and ACC.

3.6. Jackknife methodology evaluation results

The jackknife methodology metric analyzes the changes in the number of key proteins correctly identified by each method as the number of true key proteins increases.

Figure 2 shows the jackknife methodology evaluation results of different methods in DIP, Krogan and MIPS networks under the TOP–600 gradient. The x axis in the figure represents the cumulative number of key proteins, and the y axis represents the number of correctly identified key proteins. From the experimental results, with the increase of the number of key proteins, the CIBF method shows higher accuracy and stability than other methods.

3.7. Overlap/Difference analysis results of key proteins

This section mainly analyzes the overlap/difference of key proteins between different methods under the TOP600 gradient. The experimental results are shown in Table 13–15. Ms represents other methods except CIBF method, |CIBF∩Ms| represents key proteins recognized by CIBF method and other methods at the same time. |CIBF-Ms| represents the proportion of true key proteins recognized by CIBF method but not by other methods. |Ms-CIBF| represents the proportion of true key proteins recognized by other methods but not by CIBF method. In DIP, Krogan and MIPS networks, the value of |CIBF-Ms| is not less than 50.7, 53.7 and 44.3. This shows that CIBF method can identify key proteins more effectively.

Table 13. Overlap/Difference of key protein on DIP network.

Method	|CIBF∩Ms|	|CIBF-Ms|	|Ms-CIBF|
DC	205	0.6481	0.2882
BC	169	0.6566	0.2701
SC	174	0.6268	0.3201
NC	271	0.6292	0.3423
PeC	269	0.6042	0.3992
LBCC	300	0.5423	0.4742
UC	290	0.5774	0.4303
TEO	318	0.5071	0.4462
CENC	294	0.5980	0.4267

Table 14. Overlap/Difference of essential protein on Krogan network.

Method	|CIBF∩Ms|	|CIBF-Ms|	|Ms-CIBF|
DC	234	0.6419	0.3287
BC	208	0.6653	0.3186
SC	199	0.6796	0.2750
NC	263	0.6048	0.4022
PeC	232	0.6075	0.3658
LBCC	254	0.5790	0.4860
UC	233	0.5993	0.4492
TEO	267	0.5375	0.4932
CENC	244	0.5595	0.4604

Table 15. Overlap/Difference of essential protein on Krogan network

Method	|CIBF∩Ms|	|CIBF-Ms|	|Ms-CIBF|
DC	143	0.5378	0.2059
BC	130	0.5444	0.1890
SC	66	0.5824	0.0962
NC	141	0.5576	0.2594
PeC	176	0.4830	0.2798
LBCC	231	0.4434	0.4047
UC	193	0.4706	0.3836
TEO	207	0.4829	0.4497
CENC	197	0.4768	0.3717

4. Discussions

4.1. Effectiveness and stability

To verify the performance of CIBF, PPI networks with various topological properties were selected for key protein identification and compared with other methods. The experimental results show that CIBF can identify more key proteins, and the performance of multiple evaluation indicators in different networks (such as F–measure, ACC, etc.) also proves that the CIBF method has good stability and effectiveness.

4.2. Limitations and deficiencies

Although CIBF method has made some progress in identifying key proteins, it still has the following defects and deficiencies:

(1) Non-central key protein recognition. CIBF and existing methods in the identification of key proteins are mainly based on the node centrality, while some key proteins have low centrality in the network, so the accuracy of the recognition of such key proteins still needs to be improved;

(2) High-quality PPI network construction. The processing environment of proteins in the organism is constantly changing, which will affect the accuracy of key protein identification, so we can build higher quality PPI network, such as the construction of dynamic PPI network fusion with fusion temporal characteristics;

(3) More effective biological feature fusion methods. The existing methods to use protein biological features is relatively simple, need to improve the effectiveness of biological feature fusion. Graph representation learning performs well in the processing of graph data, which can be considered to improve the accuracy of key protein identification.

5. Conclusion

Identification of key proteins in the PPI network is not only helpful to analyze biological tissue structure and important to predict pathogenic genes and discover drugs. In this paper, we propose an key protein identification method that combines complex information and multiple biological characteristics. This method comprehensively evaluates the importance of proteins from the perspectives of subcellular localization centrality, GO function centrality, and complex participation centrality of proteins. The experimental results in different PPI networks show that the CIBF method can effectively identify key proteins and has good stability.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. L. yan Wang, Z. Zhang, Y. Li, Y. Wan, B. Xing, Integrated bioinformatic analysis of rna binding proteins in hepatocellular carcinoma, *Aging (Albany NY)*, **13** (2020), 2480–2505. https://doi.org/10.18632/aging.202281
2. X. Wang, J. Zhao, Targeted cancer therapy based on acetylation and deacetylation of key proteins involved in double-strand break repair, *Cancer Manag. Res.*, (2022), 259–271. https://doi.org/10.2147/CMAR.S346052
3. Y. Yue, C. Ye, P.-Y. Peng, H.-X. Zhai, I. Ahmad, C. Xia, et al., A deep learning framework for identifying essential proteins based on multiple biological information, *BMC Bioinform.*, **23** (2022), 318. https://doi.org/10.1186/s12859-022-04868-8
4. Y. Liu, W. Chen, Z. He, Essential protein recognition via community significance, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **18** (2021), 2788–2794. https://doi.org/10.1109/TCBB.2021.3102018
5. L. Shen, J. Zhang, F. Wang, K. Liu, Predicting essential proteins based on integration of local fuzzy fractal dimension and subcellular location information, *Genes*, **13** (2022), 173. https://doi.org/10.3390/genes13020173
6. X.-J. Lei, Y. Gao, L. Guo, Mining protein complexes based on topology potential weight in dynamic protein-protein interaction networks, *Acta Electon. Sin.*, **46** (2018), 145. https://doi.org/10.3969/j.issn.0372-2112.2018.01.020
7. T. Tang, X. Zhang, Y. Liu, H. Peng, B. Zheng, et al., Machine learning on protein–protein interaction prediction: models, challenges and trends, *Brief. Bioinform.*, **24** (2023), bbad076. https://doi.org/10.1093/bib/bbad076
8. M. Li, H. Zhang, J.-x. Wang, Y. Pan, A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data, *BMC Syst. Biol.*, **6** (2012), 1–9. https://doi.org/10.1186/1752-0509-6-15
9. W. Peng, J. Wang, Y. Cheng, Y. Lu, F. Wu, Y. Pan, Udonc: An algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **12** (2014), 276–288. https://doi.org/10.1109/TCBB.2014.2338317
10. X. Shang, Y. Wang, B. Chen, Identifying essential proteins based on dynamic protein-protein interaction networks and rna-seq datasets, *Sci. China Inform. Sci.*, **59** (2016), 1–11. https://doi.org/10.1007/s11432-016-5583-z
11. M. LI, X.-t. WANG, H.-m. LUO, X.-m. MENG, J.-x. WANG, Progress on random walk and its application in network biology, *Acta Electon. Sin.*, **46** (2018), 2035. https://doi.org/10.3969/j.issn.0372-2112.2018.08.033

12. M. Li, Y. Lu, Z. Niu, F.-X. Wu, United complex centrality for identification of essential proteins from ppi networks, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **14** (2015), 370–380. https://doi.org/10.1109/TCBB.2015.2394487

13. J. Zhong, C. Tang, W. Peng, M. Xie, Y. Sun, Q. Tang, et al., A novel essential protein identification method based on ppi networks and gene expression data, *BMC Bioinform.*, **22** (2021), 1–21. https://doi.org/10.1186/s12859-021-04175-8

14. C. Qin, Y. Sun, Y. Dong, A new method for identifying essential proteins based on network topology properties and protein complexes, *PloS One*, **11** (2016), e0161042. https://doi.org/10.1371/journal.pone.0161042

15. G. Yu, G. Fu, J. Wang and H. Zhu, Predicting protein function via semantic integration of multiple networks, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **13** (2015), 220–232. https://doi.org/10.1109/TCBB.2015.2459713

16. J. Luo, Y. Qi, Identification of essential proteins based on a new combination of local interaction density and protein complexes, *PloS One*, **10** (2015), e0131418. https://doi.org/10.1371/journal.pone.0131418

17. X. Yang, Z. Niu, Y. Liu, B. Song, W. Lu, L. Zeng, et al., Modality-dta: Multimodality fusion strategy for drug–target affinity prediction, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **20** (2022), 1200–1210. https://doi.org/10.1109/TCBB.2022.3205282

18. W. Zhang, J. Xu, Y. Li, X. Zou, Detecting essential proteins based on network topology, gene expression data, and gene ontology information, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **15** (2016), 109–116. https://doi.org/10.1109/TCBB.2016.2615931

19. B. Chen, W. Fan, J. Liu, F.-X. Wu, Identifying protein complexes and functional modules—from static ppi networks to dynamic ppi networks, *Brief. Bioinform.*, **15** (2014), 177–194. https://doi.org/10.1093/bib/bbt039

20. R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, A. Raval, Identifying hubs in protein interaction networks, *PloS One*, **4** (2009), e5344. https://doi.org/10.1371/journal.pone.0005344

21. M. P. Joy, A. Brock, D. E. Ingber, S. Huang, High-betweenness proteins in the yeast protein interaction network, *J. Biomed. Biotechnol.*, **2005** (2005), 96. https://doi.org/10.1155/JBB.2005.96

22. E. Estrada, J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, *Phys. Rev. E*, **71** (2005), 056103. https://doi.org/10.1103/PhysRevE.71.056103

23. J. Wang, M. Li, H. Wang, Y. Pan, Identification of essential proteins based on edge clustering coefficient, *IEEE/ACM Transact. Comput. Biol. Bioinform.*, **9** (2011), 1070–1080. https://doi.org/10.1109/TCBB.2011.147

24. P. Lu, J. Yu, A mixed clustering coefficient centrality for identifying essential proteins, *Int. J. Modern Phys. B*, **34** (2020), 2050090. https://doi.org/10.1142/S0217979220500897

25. I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, D. Eisenberg, Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.*, **30** (2002), 303–305. https://doi.org/10.1093/nar/28.1.289

26. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, et al., Global landscape of protein complexes in the yeast saccharomyces cerevisiae, *Nature*, **440** (2006), 637–643. https://doi.org/10.1038/nature04670

27. U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, et al., Mpact: The mips protein interaction resource on yeast, *Nucleic Acids Res.*, **34** (2006), D436–D441. https://doi.org/10.1093/nar/gkj003

28. G. O. Consortium, Gene ontology annotations and resources, *Nucleic Acids Res.*, **41** (2012), D530–D535. https://doi.org/10.1093/nar/gks1050

29. J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, et al., Compartments: Unification and visualization of protein subcellular localization evidence, *Database*, **2014**. https://doi.org/10.1093/database/bau012

30. R. Zhang, Y. Lin, Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes, *Nucleic Acids Res.*, **37** (2009), D455–D458. https://doi.org/10.1093/nar/gkn858

31. H.-W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, et al., Mips: Analysis and annotation of proteins from whole genomes, *Nucleic Acids Res.*, **32** (2004), D41–D44. https://doi.org/10.1093/nar/gkh092

32. J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, et al., Sgd: Saccharomyces genome database, *Nucleic Acids Res.*, **26** (1998), 73–79. https://doi.org/10.1093/nar/26.1.73

33. E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, et al., Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis, *Science*, **285** (1999), 901–906. https://doi.org/10.1126/science.285.5429.90

AIMS Press