IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2020, accepted May 7, 2020, date of publication May 11, 2020, date of current version May 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993860

An Iteration Method for Identifying Yeast Essential Proteins From Weighted PPI Network Based on Topological and Functional Features of Proteins

SHIYUAN LI[1], ZHIPING CHEN[1], XIN HE[2], ZHEN ZHANG[3], TINGRUI PEI[2], YIHONG TAN[1], AND LEI WANG[1,2]

[1]College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China
[2]Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China
[3]College of Electronic Information and Electrical Engineering, Changsha University, Changsha 410022, China

Corresponding authors: Tingrui Pei (peitingrui@xtu.edu.cn) and Lei Wang (wanglei@xtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873221 and Grant 61672447, in part by the Natural Science Foundation of Hunan Province under Grant 2018JJ4058 and Grant 2019JJ70010, and in part by the National Natural Science Foundation of China under Grant 61873221.

ABSTRACT Accumulating studies have indicated that essential proteins play critical roles in numerous biological processes. With the rapid development of high-throughput technologies, a large number of Protein-Protein Interaction (PPI) data have been found in Saccharomyces cerevisiae, which facilitate the formation of PPI networks. Up to now, a series of computational methods for predicting essential proteins from PPI networks have been proposed successively. However, the prediction accuracy of these computational methods is still not quite satisfactory. In this paper, a novel prediction method called CVIM is proposed to infer potential essential proteins. In CVIM, original PPI networks will be first transferred into weighted PPI networks by implementing PCC (Pearson Correlation Coefficient) on protein gene expression data. And then, based on weighted PPI networks and information of orthologous proteins, some critical network topological features and protein functional features will be extracted for each protein in the weighted PPI network. Finally, based on these newly extracted topological and functional features of proteins, an iterative algorithm will be designed to predict essential proteins. In order to evaluate the identification performance of CVIM, we have compared CVIM with 13 kinds of state-of-the-art prediction methods. Experimental results show that CVIM can achieve prediction accuracies of 92%, 80% and 71% out of the top 1%, 5% and 10% candidate proteins separately, which significantly outperform the prediction accuracies achieved by those state-of-the-art prediction methods. We have demonstrated that the prediction accuracy of essential proteins can be effectively improved by integrating the functional and network topological characteristics of proteins, which means that the novel method CVIM may be an excellent addition to the protein researches in the future.

INDEX TERMS Characteristic vector, orthologous proteins, essential proteins, weighted protein-protein interaction network, iteration method.

I. INTRODUCTION

More and more evidences have shown that essential proteins are critical to the development and survival of organisms, and absence of these proteins will lead to loss of biological functions of protein complexes and death of organisms. Prediction of essential proteins plays a crucial role in research of bioinformatics, which is not only of great significance to the study of life sciences, but also of great application value in drug design and treatment of diseases. In recent years, a number of computational methods for essential protein prediction have been proposed successively. However, the identification accuracy of essential proteins is still not quite high. Hence, it is an important and challenging task to design

The associate editor coordinating the review of this manuscript and approving it for publication was Haipeng Yao.

efficient prediction methods to identify potential essential proteins [1]–[3].

Up to now, existing prediction methods for essential protein can be roughly divided into two major categories. Methods of the first category mainly rely on the topological features of PPI networks. For instance, Li M *et al.* proposed a topology potential based calculative method to infer essential proteins from PPI Networks [4], and a calculative method called LAC (Local Average Connectivity-based) to infer essential proteins through evaluating the relationship between proteins and their neighborhoods [5] separately. Xu, Bin and Guan, Jihong et al developed a model to detect key proteins by weighting random walks on protein-protein interaction networks [6]. Y. Jiang and y. Wang *et al.* established a method for the identification of key proteins based on the prediction of key protein-protein interactions based on comprehensive edge weights [7]. Especially, based on the centrality-lethality rule proposed by Jeong *et al.* [9], researchers have developed various centrality-based methods, such as DC (Degree Centrality) [10], SC (Subgraph Centrality) [11], BC (Betweenness Centrality) [12], EC (Eigenvector Centrality) [13], IC (Information Centrality) [14], CC (Closeness Centrality) [15] and NC (Neighbor Centrality) [16]. These methods identify important proteins based on the topology of the PPI network, such as the number of protein connections, the number of common neighbors, and so on. Although methods of the first category have made great progress compared to traditional bio-experiments, however, due to the incomplete PPI data, which are obtained through biological experiments and often contain noise such as false positive data and false negative data, the first category of methods cannot achieve satisfactory identification accuracy of essential proteins on most occasions.

Hence, different from methods of the first category, the second method is to combine the topology of PPI network with biological information (gene expression data, subcellular location data, orthology data, gene ontology) to construct a prediction model and improve the prediction accuracy. For example, Chen Lei *et al.* used the rich gene ontology and KEGG pathway to predict and analyze essential genes [8]. Zhao and Wang designed an iteration method called RWHN for identifying yeast essential proteins from heterogeneous network by combining PPI networks with protein domains, the subcellular localization information and orthologous information [1]. M Li *et al.* proposed a prediction method called PEC to identify essential proteins by combining PPI network topology and gene expression [17]. Zhang *et al.* developed a computational method called CoEWC through combining the characteristics of PPI network topology and protein co-expression characteristics based on gene expression profiles [18]. Seketoulie Keretsu *et al.* proposed a calculative method based on the weight of the edge between two interacting proteins to identify protein complexes, in which, the weight was defined by the edge clustering coefficient and the gene expression correlation between the interacting proteins [19]. Bingjing Cai *et al.* presented a biased random walk based method to identify protein complexes by integrating Tandem Affinity Purification/Mass Spectrometry Data with PPI networks [20]. Jiawei L *et al.* put forward a computational method for detecting essential proteins by integrating local interaction density and protein complexes [21]. B.H. Zhao *et al.* adopted the gene expression data and network topology attributes to construct a reliable weighted network, based on which, a novel computational method called POEM was further designed to forecast essential protein based on overlapping essential modules [22]. Yijia Zhang *et al.* constructed a dynamic PPI network by integrating dynamic active information into high-throughput PPI data, based on which, a novel method for predicting protein complexes from the dynamic PPI networks is proposed based on core-attachment structural feature [23]. Ma CY *et al.* presented a novel algorithm called NEOComplex to infer protein complexes by integrating functional orthology information obtained from different types of multiple network alignment approaches with PPI networks [24]. Lei X *et al.* proposed a method called IFPA for protein complex detection in multi-relation reconstructed dynamic protein networks by adopting the flower pollination mechanism [25]. Peng W *et al.* put forward an iterative method called ION to reveal essential proteins through integrating homologous information and PPI networks [26]. Luo J *et al.* designed a new algorithm to discover essential proteins by combining protein complex co-expression information with edge clustering coefficient [27]. Xu B *et al.* developed a machine learning based method to identify protein complex through integrating protein-protein interaction evidence from 6 different sources [28], and a calculative method called GANE to predict protein complexes based on go attributed network embedding [29] separately. Lei X *et al.* proposed a computational method called NABCAM to discover protein complexes from dynamic PPI networks [30]. Ou-Yang L *et al.* presented a multi-network clustering method to infer protein complexes from multiple heterogeneous networks [31]. Srihari S *et al.* proposed a refinement of MCL by incorporating core-attachment structure to predict yeast complexes from weighted PPI networks [32].

In different to the first category central approach, to reduce the negative impact of incomplete protein interaction data and inherent PPI network topological characteristics on essential protein prediction, we combined multi-source biological data: gene expression, direct homologous information. Although gene expression is mentioned in the second category method above, most methods simply combine gene expression with network topological data, but ignore the essential differences in the meanings of biological data and network topological data. For example, in Pec, PCC $*$ ECC is used directly to get the final result. Therefore, this paper proposes a new iterative method, called CVIM. The method detects essential proteins by combining protein function and network topology. In CVIM, considering the current incomplete PPI data set, we first used PCC (Pearson correlation coefficient) [33] for protein gene expression data to convert the original PPI network into a weighted PPI network. Then, based on the weighted PPI network and the information of the direct homologous proteins, we will further extract some key network topological characteristics and protein functional characteristics of each protein in the weighted PPI network. Generate new protein interaction matrix (network) from network topological data. Finally, based on these newly acquired protein interaction networks and functional properties, we will construct an iterative method called CVIM to predict the required protein. In order to estimate the identification performance of CVIM, intensive experiments will be implemented. Experimental results show that CVIM can achieve the prediction accuracies of 92%, 80% and 71% in the top 1%, 5% and 10% proteins respectively, which are much better than that achieved by 13 state-of-the-art competitive methods including DC [10], SC [11], BC [12], EC [13], IC [14], CC [15], NC [16], LAC [5], RWHN [1], PEC [17], CoEWC [18], POEM [22] and ION [26].

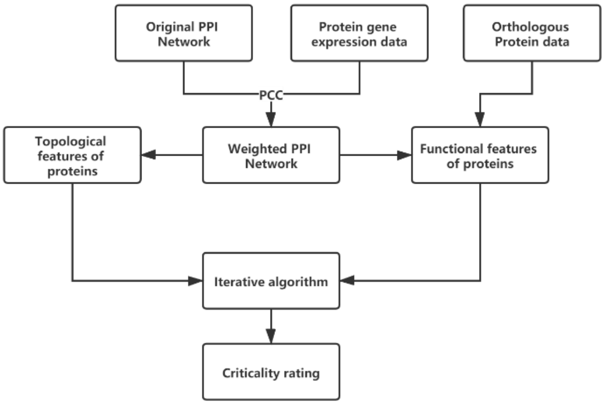


FIGURE 1. **Procedure of CVIM.**

II. METHOD

As illustrated in Fig.1, the procedure of CVIM consists of the following three major steps:

Step1: First, we adopt PPC on gene expression data to establish weights between protein nodes in the original PPI network, and then the original PPI network will be transferred into a weighted PPI network.

Step2: Next, based on the weighted PPI network and information of orthologous proteins, some critical network topological features and protein functional features will be extracted for each protein in the weighted PPI network separately.

Step3: Finally, based on the topological and functional features of proteins, a novel iteration method called CVIM will be designed to identify essential proteins by using an iterative algorithm.

A. CONSTRUCTION OF THE WEIGHTED PPI NETWORK

Let $G = (V, E)$ denote an original PPI network constructed by the dataset of known PPIs downloaded from a public database D. Here, $V = \{p_1, p_2, \ldots, p_N\}$ represents the set of different proteins in D, and E represents the set of edges between proteins in V. Additionally, for a pair of proteins p and q in V, there is an edge $e(p, q)$ between them, if and only if there is a known interaction between p and q in D. Based on the original PPI network G, it is clear that we can obtain an adjacency matrix $A = (a_{ij})_{N \times N}$, where there is $a_{ij} = 1$, if and only if there is an edge $e(p_i, p_j)$ between p_i and p_j, otherwise there is $a_{ij} = 0$.

PCC measures the linear correlation between two vectors. Gene expression is the process of using gene information to synthesize functional gene products. These gene products are usually proteins. We believe that the gene expression of key proteins at different times may have similar performance, that is, the gene expression vectors of the two key proteins may have a large linear correlation. Moreover Horyu*et al.* [33] found that the Pearson correlation coefficient is more suitable as a similarity measures for gene expression profiles. Therefore, we use PCC as the measurement factor of the new method calculated the co-expression intensity of the two genes, and transformed the two original PPI networks into two weighted PPI networks, as follows:

For a given protein p, its gene expression at different times can be expressed by a vector: $Exp(p) = \{Exp(p,1), Exp(p,2), \ldots, Exp(p, n)\}$, where $Exp(p, i)$ is the expression level of the protein p at the ith time. Evidently, based on the Pearson Correlation Coefficient, in an original PPI network, the weight between two proteins p and q can be calculated as follows:

$$\begin{aligned} weight\,(p, q) &= PCC\,(p, q) \\ &= \frac{1}{n-1} \sum\nolimits_{i=1}^{n} \left(\frac{Exp\,(p, i) - \overline{Exp(p)}}{\sigma\,(p)} \right) \\ &\quad \times \left(\frac{Exp\,(q, i) - \overline{Exp(q)}}{\sigma\,(q)} \right) \end{aligned} \tag{1}$$

Here, $\overline{Exp(p)}$ denotes the average expression of protein p at all times, $\sigma\,(p)$ is the standard variance of expression for protein p at all times. If $PCC(p, q)$ has a positive value, then it means a positive correlation between these two proteins p and q, otherwise, if $PCC(p, q)$ has a negative value, then it means a negative correlation between these two proteins p and q.

Evidently, based on above formula (1), an original PPI network can be transferred into a weighted PPI networks easily.

B. EXTRACTION OF TOPOLOGICAL AND FUNCTIONAL FEATURES FOR PROTEINS

For a given protein p in an original PPI network $G = (V, E)$, let $NG(p)$ denote the set of neighboring nodes that have known interactions with p in G, then there is

$$NG\,(p) = \{q | \exists e\,(p, q) \in E, q \in V\} \tag{2}$$

Through the analysis of the network structure formed by protein interactions, a lot of research has been conducted on the identification of key proteins, and some good results have been achieved, such as the LAC [4] method. In the studies of Hart *et al.* [43] and Dezso *et al.* [44], it was found that in many cases, the necessities are not the functional products of individual proteins, but the products of complex functions. Considering that triangles have the most stable properties in the geometric structure, the triangle structure of the PPI network to happen to be a local measurement feature that determines the protein necessity according to the modular nature of the protein necessity. Therefore, the number of triangles formed by the connections between proteins constitutes a feature of our algorithm. In this section, according to the weighted PPI network newly constructed above, we will first calculate the number of triangles for each protein p in PPI network $G = (V, E)$ as follows:

$$Tris\,(p,q) = \begin{cases} |NG\,(p) \cap NG\,(q)|+1; & if\ weight\ (p,q) > 0 \\ 1; & otherwise \end{cases} \tag{3}$$

$$Tris(p) = \sum\nolimits_{q \epsilon NG(p)} Tris(p,q) \tag{4}$$

Here, $|NG(p) \cap NG(q)|$ denotes the number of elements in the set of $NG(p) \cap NG(q)$.

Based on above formula (3) and (4), we can extract the first network topological feature TF_1 for the protein p as follows:

$$TF_1(p) = avgTris(p) = \frac{Tris(p)}{|NG(p)|} \tag{5}$$

Here, $|NG(p)|$ denotes the number of elements in the set of $NG(p)$.

Next, in the study of Li *et al.* [17], it was mentioned that key proteins tended to form tightly connected clusters. The neighbors of key proteins are also in a closely related cluster. Based on this view, we believe that if protein p is an essential protein, then its neighbor may also be an essential protein, for each protein p, we will extract another network topological feature TF_2 for it as follows:

$$TF_2(p) = avgTris(p) = \frac{NG_{Tris}(p)}{NG_e(p)} \tag{6}$$

where $NG_e(p)$ denotes the number of edges of all nodes in $NG(p)$, and $NG_{Tris}(p)$ means the number of triangles of all nodes in $NG(p)$, which can be calculated according to the following formulas:

$$NG_e\,(p) = \sum\nolimits_{q \in NG(p)} |NG(q)| \tag{7}$$

$$NG_{Tris}\,(p) = \sum\nolimits_{q \in NG(p)} |Tris(q)| \tag{8}$$

Moreover, in the study of Peng *et al.* [26], the key proteins proved to be relatively conservative. By studying 99 reference organisms from Homo sapiens to modern humans. Whether each protein has homology, get the homology score of each protein, which indicates the degree of conservation of each protein. For each protein p in an original PPI network $G = (V, E)$, supposing that its orthologous score is $I(p)$, then, we can extract its first functional feature $FF_1(p)$ from the information of orthologous proteins as follows:

$$FF_1(p) = \frac{I(p)}{\max\limits_{q \in V} \{I\,(q)\}} \tag{9}$$

Finally, based on the weighted PPI network, for each protein p, we can further extract its another functional feature $FF_2(p)$ as follows:

$$FF_2\,(p) = (\sum\nolimits_{q \in NG(p)} weight(p,q))/|NG(p)| \tag{10}$$

where $\sum_{q \in NG(p)} weight(p,q)$ represents the sum of the co-expression degree of protein p and all its neighbor nodes, and the ratio of $\sum_{q \in NG(p)} weight(p,q)$ to the number of neighbor nodes represents the average level of co-expression degree of protein p in the whole PPI network.

C. CONSTRUCTION OF CVIM

Based on above descriptions, let $\{TF_{i1}, TF_{i2}, \ldots, TF_{iM}\}$ denote all these topological features (such as TF_1 and TF_2) extracted for the protein p_i from the PPI network, then it is obvious that we can obtain a $N \times M$ dimensional characteristic matrix TF for all these N different proteins in the PPI network as follows:

$$\boldsymbol{TF} = \begin{bmatrix} \boldsymbol{TF_{11}} & \cdots & \boldsymbol{TF_{1M}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{TF_{N1}} & \cdots & \boldsymbol{TF_{NM}} \end{bmatrix} \tag{11}$$

After normalizing above matrix TF, we can obtain a transformation matrix B as follows:

$$\boldsymbol{B} = [\boldsymbol{b_{ij}}]_{\boldsymbol{N \times M}}, \quad \text{with } \boldsymbol{b_{ij}} = \boldsymbol{TF_{ij}} \Big/ \sum\nolimits_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{TF_{kj}} \tag{12}$$

Based on above formula (12), for the jth network topological feature of proteins, we can obtain its entropy e_j, which represents the stability of the jth feature, as follows:

$$\boldsymbol{e_j} = -\sum\nolimits_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{b_{ij} lnb_{ij} / lnN} \tag{13}$$

Based on above formula (13), for the jth network topological feature of proteins, we can calculate its weight in all M different network topological features according to the following formula (14):

$$\boldsymbol{w_j} = (1 - \boldsymbol{e_j}) / \sum\nolimits_{\boldsymbol{i=1}}^{\boldsymbol{M}} (\boldsymbol{1 - e_i}) \tag{14}$$

Thereafter, based on above formula (14), for a given protein p_i, we can calculate its score of network topological features as follows:

$$\boldsymbol{TFscore(i)} = \sum\nolimits_{\boldsymbol{j=1}}^{\boldsymbol{M}} \boldsymbol{w_j TF_{ij}} \tag{15}$$

Based on above formula (15), for all these N proteins in the PPI network, we can construct a protein interaction matrix

TABLE 1. A rough introduction to other algorithms.

algorithm	Network topology	Biological information
DC[10]	Degree Centrality	No
SC[11]	Subgraph Centrality	No
BC[12]	Betweenness Centrality	No
EC[13]	Eigenvector Centrality	No
IC[14]	Information Centrality	No
CC[15]	Closeness Centrality	No
NC[16]	Neighbor Centrality	No
LAC[5]	Degree Centrality, Common neighbor node	No
PEC[17]	Edge clustering coefficient	Gene expression data
CoEWC[18]	Clustering coefficient,	Gene expression data
ION[26]	Edge clustering coefficient	Orthologous data
POEM[22]	Degree Centrality, Subgraph, Edge clustering coefficient, Closeness Centrality	Gene expression data
RWHN[1]	Degree Centrality, protein- domain	Orthologous data, subcellular localization

$\boldsymbol{H} = [\boldsymbol{h_{ij}}]_{N\times N}$ as follows:

$$h_{ij} = \begin{cases} \dfrac{TFscore(i)}{\sum_{k=1}^{N} TFscore\ (k)}; & if\ i = j \\ \dfrac{min\{TFscore\ (i)\ ,\ TFscore\ (j)\}}{\sum_{k=1}^{N} TFscore\ (k)}; & Otherwise \end{cases} \tag{16}$$

Based on above formula (9) and formula (10), for a given protein p_i, we define its total score of protein functional features as follows:

$$FFscore(i) = (FF_1(p_i) + FF_2(p_i))/2 \tag{17}$$

For all N proteins $\{p_1, p_2, \ldots, p_N\}$ in the weighted PPI network, then we can obtain their initial scores as follows:

$$T(0) = (FFscore(1), FFscore(2), \ldots, FFscore(N)) \tag{18}$$

Finally, we adopt formula (19) to compute all the proteins' criticality score iteratively

$$T(t+1) = \alpha * H * T(t) + (1-\alpha) * T(0) \tag{19}$$

Here, the parameter $\alpha(0 \le \alpha \le 1)$ is utilized to adjust the proportion of initial score T(0) and last iteration score T(t). Thereafter, based on above descriptions, we can present our CVIM algorithm as follows:

```
Algorithm CVIM
    Input: Original PPI network $G = (V, E)$, orthologous and gene expression data, the parameters $\varepsilon$ and $K$
    Output: Top $K$ percent of proteins sorted by the vector T in descending order
    Step1: Generate the weighted network according to formula (1);
    Step2: For each protein $p$, extract its network topological features $TF_1$ and $TF_2$ from the novel weighted PPI network according to formulas (5) and (6) separately;
    Step3: For each protein $p$, extract its functional features $FF_1$ and $FF_2$ from the novel weighted PPI network, orthologous data and gene expression data according to formulas (9) and (10) respectively;
    Step4: Obtain the protein interaction matrix $H$ according to formula (16);
    Step5: Let $t = 0$, Compute $T_{(t)}$ according to (18);
    Step6: Let $t = t + 1$; Compute $T_{(t)}$ according to formula (19);
    Step7: Repeat Step6 until $\|T\ (t) - T(t-1)\| \Big/ \sqrt{N} < \varepsilon$;
    Step8: Sort proteins by the value of $T$ in the descending order;
    Step9: Output top $K$ percent of sorted proteins.
```

III. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA

In order to evaluate the performance of CVIM, we will compare it with 13 representative methods in Table 1 based on the datasets downloaded from two databases DIP [34] and GAVIN [35] separately. During experimental, after filtering out self-interactions and repeated interactions, we finally obtained 5093 different proteins and 24743 interactions including 1167 essential proteins from the DIP database, and 1855 different proteins and 7669 interactions including 714 essential proteins from the GAVIN database. Obviously, based on these two datasets downloaded from the DIP and GAVIN databases, two kinds of original PPI networks, such as a DIP-based PPI network and a GAVIN-based PPI network, can be constructed.

Moreover, information of orthologous proteins was downloaded from the InParanoid database (Version 7) [36], which consists of a collection of pair wise comparisons between 100 whole genomes. And additionally, the gene expression data of yeast was downloaded from the dataset provided by

TABLE 2. Effects of the parameter α to CVIM based on the DIP-based PPI network.

α Rank	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Top1%	0.80	0.84	0.86	0.88	0.94	0.94	0.94	0.92	0.92
Top5%	0.69	0.74	0.76	0.77	0.79	0.80	0.80	0.80	0.80
Top10%	0.62	0.63	0.64	0.66	0.68	0.69	0.69	0.71	0.70
Top15%	0.57	0.57	0.58	0.59	0.59	0.60	0.60	0.61	0.61
Top20%	0.51	0.52	0.52	0.52	0.53	0.54	0.54	0.55	0.55
Top25%	0.47	0.47	0.47	0.47	0.48	0.48	0.48	0.49	0.49

Table 2: This table shows the effects of the parameter α to CVIM based on the DIP-based PPI network. While α is set to different values from 0.1 to 0.9, the top 1 to 25 percent of identified proteins are selected, and the table records the proportion of true key proteins in the set of selected proteins.

TABLE 3. Effects of the parameter α to CVIM based on the GAVIN-based PPI network.

α Rank	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Top1%	0.89	0.89	0.95	1	1	0.95	0.95	0.84	0.69
Top5%	0.83	0.85	0.86	0.87	0.86	0.85	0.88	0.86	0.78
Top10%	0.79	0.81	0.82	0.85	0.86	0.85	0.84	0.82	0.71
Top15%	0.72	0.75	0.78	0.78	0.79	0.79	0.79	0.79	0.68
Top20%	0.68	0.70	0.70	0.71	0.73	0.74	0.74	0.73	0.68
Top25%	0.65	0.66	0.67	0.67	0.68	0.68	0.70	0.69	0.66

Table 3: This table shows the effects of the parameter α to CVIM based on the GAVIN-based PPI network. While α is set to different values from 0.1 to 0.9, the top 1 to 25 percent of identified proteins are selected, and the table records the proportion of true key proteins in the set of selected proteins.

Tu *et al.* [37]. In experiment, the coverage of the DIP-based PPI network and the GAVIN-based PPI network in the gene expression data reached over 95%. For proteins that do not have corresponding gene expression data, we would set their values of gene expression to zero.

Finally, we would further download a dataset consisting of 1285 essential genes of Saccharomyces cerevisiae from four databases such as MIPS [39], SGDP [42], DEG [40] and SGD [41] as the benchmark set. By comparing the key proteins screened by CVIM with these 1285 real key proteins, the recognition rate of CVIM method in DIP database and GAVIN database was obtained. We will present the experimental results of PPI network based on DIP in detail, and briefly present the experimental results of PPI network based on GAVIN.

B. EFFECTS OF THE PARAMETER α

In CVIM, we introduced a user-defined parameter α with value between 0 and 1. By setting different values to α, we illustrated the prediction results based on the DIP-based PPI network and the GAVIN-based PPI network in the following Table 2 and Table 3 respectively.

As shown in Table 2, We sort the final score of the protein in descending order, and selected the top 1%, 5%, 10%, 15%, 20%, and 25% of the potentially essential proteins identified by CVIM, while α was set to 0.1, 0.2,..., 0.8, and 0.9. It is not difficult to see that the prediction accuracy of CVIM will change with different α values. Overall, as the value of α increases, the accuracy of CVIM prediction will steadily increase. Although the recognition rate of the top 20% and 25% dropped to 0.47-0.55, this is because in the data set, key protein data only accounts for about 20% of all data, the data distribution is extremely uneven, and our research is mainly for the identification of key proteins, so we mainly consider the key protein within 20% The recognition situation, and the proportion of more than 20% of the data corresponding to the data of non-critical proteins has increased significantly, resulting in a rapid decline in accuracy. Therefore, we think that when performing comparative experimental on a DIP-based PPI network, setting the value of α to 0.8 is the most appropriate.

As shown in Table 3, it is easy to see that when α is increased to 0.7, the top 5%, 15%, 20%, and 25% of the potential essential proteins identified by CVIM all reach the best prediction accuracy. However, when α is set to 0.5, the first 1% and 10% identified by CVIM can obtain the best prediction accuracy. Therefore, considering both the experimental results, when comparing the analog network PPI GAVIN performed based on the value of α is set to 0.7 is the most suitable.

Although we get the best effect at $\alpha = 0.8$ for DIP dataset and at $\alpha = 0.7$ for GAVIN dataset, it will cause over-fit for different dataset with different parameter values. Therefore, combining the two databases, we chose 0.8 as value of the parameter α in the following experiment. And we also tested in Krogan [45], BioGRID database, the Krogan dataset consists of 3672 proteins and 14317 interactions. The BioGRID yeast data set used in [4] contains 5616 proteins and 52833 distinct interactions, which are denser than the other three data sets. we found that the alpha parameter does not change much in different data sets, and has little effect on the experimental results.

C. COMPARISONS BETWEEN CVIM AND 13 REPRESENTATIVE METHODS

First, we adopt the dataset downloaded from the DIP database to compare CVIM with 13 representative methods in Table 1 simultaneously. And the experimental results are illustrated in the following Fig.2.

From observing the Fig.2, it is easy to see that in the top 1% (51), 5% (255) and 10% (510) potential essential proteins

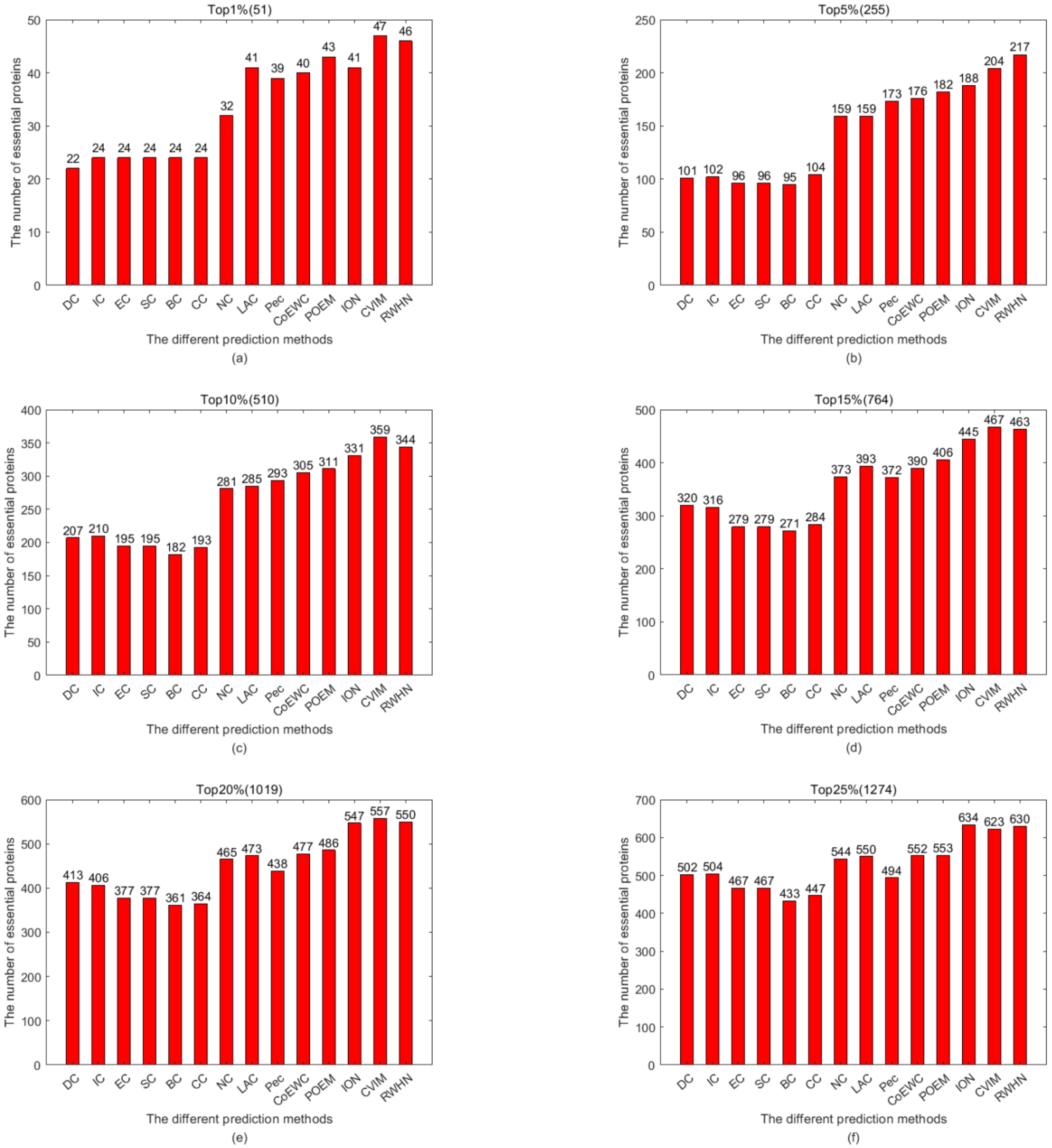


FIGURE 2. **(*a*) Top 1% ranked proteins. (*b*) Top 5% ranked proteins. (*c*) Top 10% ranked proteins. (*d*) Top 15% ranked proteins. (*e*) Top 20% ranked proteins. (*f*) Top 25% ranked proteins. This figure illustrates the comparison of the number of essential proteins predicted by CVIM and 13 competing methods. During experimental, the proteins calculated by 13 methods in CVIM and table 1 were sorted in PPI network in order from high to low. Then, the top 1%, 5%, 10%, 15%, 20% and 25% ranked proteins will be selected as candidate essential proteins. Thereafter, by comparing with known key protein libraries, the performance is judged by the number of true essential proteins identified by each method. This figure shows the number of true essential proteins discovered by each method. Because the total number of ranked proteins is 5093. The digits in brackets indicate the number of proteins ranked in each top percentage.**

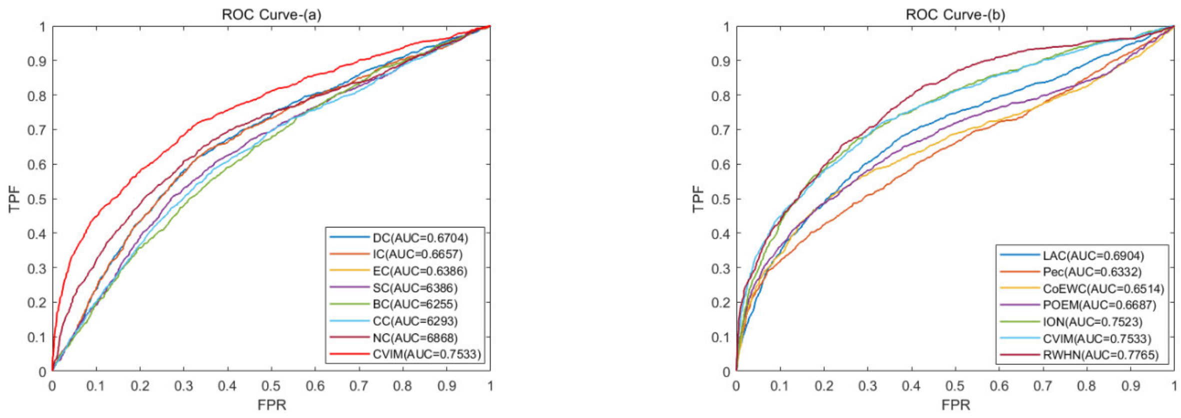


FIGURE 3. ROC curve of dip - based PPI network and AUC of various methods. (*a*) comparison of CVIM with IC, EC, DC, SC, BC, CC and NC. (*b*) comparison of CVIM with LAC, Pec, CoEWC, POEM, ION and RWHN.

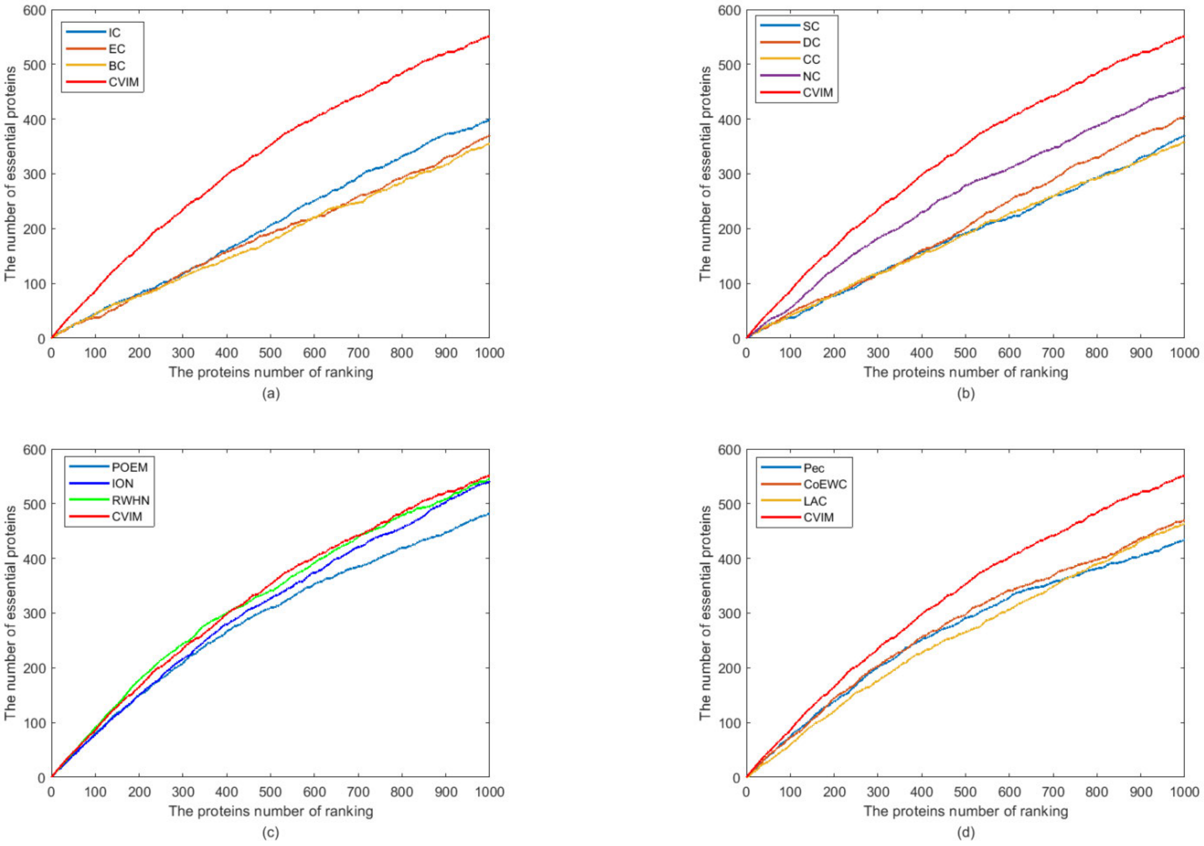


FIGURE 4. Results of comparisons between CVIM and thirteen competing methods based on the top 600 ranked key proteins by implementing the Jackknife methodology on the DIP-based PPI network. The X-axis of this figure denotes the number of ranked proteins, while the Y-axis represents the number of true key proteins identified by prediction models. (*a*) comparison between CVIM and IC, EC and BC. (*b*) comparison between CVIM and NC, DC, CC and SC. (c) comparison between CVIM and ION, RWHN and POEM. (*d*) comparison between CVIM and PEC, LAC and CoEWC.

detected by CVIM, there are 47, 204 and 359 true essential proteins respectively, which mean that the recognition rates of CVIM can reach 92%, 80% and 71% in the top 1%, 5% and 10% newly identified potential proteins separately. Particularly, while compared with the 8 representative prediction methods based on PPI network topology in table 1, our method CVIM can achieve the highest predictive accuracy in all top percentages. Moreover, compared with the five representative prediction methods based on network topology and related biological data in table 1, our method CVIM outperforms PEC, POEM, CoEWC and ION in any interval of top percentages. And in the top 1%, 10%, 15% and 20%

TABLE 4. Commonalities and differences between CVIM and 13 competing methods based on the top 200 ranked proteins and the DIP-based PPI network.

Different prediction methods (M_i)	$\lvert CVIM \cap M_i \rvert$	$\lvert$ CVIM- Mi $\rvert$	Percentage of key proteins in {CVIM−M$_i$}	Percentage of key proteins in {M$_i$−CVIM}
SC	34	166	87.35%	28.31%
BC	34	166	86.75%	28.31%
EC	34	166	87.35%	27.71%
DC	40	160	80.63%	28.75%
IC	39	161	80.75%	27.95%
CC	31	169	81.07%	30.17%
LAC	76	124	80.65%	45.16%
NC	79	79	80.17%	47.93%
PEC	107	93	72.04%	44.08%
CoEWC	104	96	71.88%	50.00%
POEM	107	93	68.82%	53.76%
ION	99	101	72.28%	57.42%
RWHN	110	90	93.33%	83.33%

Table 4: This table shows the commonalities and differences between CVIM and 13 competing methods, such as DC, IC, EC, SC, BC, CC, NC, Pec, CoEWC, POEM, ION and RWHN, based on the top 200 ranked proteins and the DIP-based PPI network.

candidate proteins, our method CVIM can achieve better performance than RWHN as well. However, in the top 5% and 25% candidate proteins, the predictive performance of CVIM is a little lower than RWHN. This may be because the RWHN method uses different parameter value settings for different data. Thus, we can draw a conclusion that CVIM is superior to these 13 state-of-the-art methods and has a higher recognition rate for key proteins in the overall level.

IV. ROC CURVE VERIFICATION

The receiver operating characteristic (ROC) curve was used to evaluate the performance of the CVIM method. If AUC = 0.5, it means random performance. The larger the area of the model's ROC curve (AUC), the better the model's performance. When FPR = 0.2, TPR = 0.58, CVIM AUC = 0.083, RWHN AUC = 0.081. Therefore, when FPR <= 0.2, the performance of the CVIM algorithm is the best among all algorithms. As FPR grows, the AUC of CVIM is slightly smaller than the RWHN algorithm.

A. VALIDATION BY JACKKNIFE METHODOLOGY

Jackknife Methodology [42] is a common method utilized to evaluate the superiority and disadvantage of algorithms for identifying key proteins. In order to evaluate CVIM more comprehensively and concretely, in this section, we introduced the Jackknife methodology for the top 1000 candidate essential proteins predicted by CVIM and 13 representative methods to test their superiority and disadvantages. The comparison result is shown in the following Fig.4. From observing Fig.4(a), Fig.4(b) and Fig.4(d), it is easy to see that CVIM can achieve better predictive performance than IC, EC, BC, NC, DC, SC, CC, Pec, LAC and CoEWC. Moreover, from observing Fig.4(c), we can find that CVIM outperforms ION and POEM, meanwhile, the curves of CVIM and RWHN are intersected with each other. However, through careful observation, we will find that when the number of candidate key proteins increases to 500, the curve of RWHN will turn lower than that of CVIM. That is to say, with the increasing of predicted scale of proteins, the predictive performance of CVIM will gradually exceed that of RWHN. Hence, we can declare that the prediction performance of CVIM is better than that of these 13 representative methods on the whole.

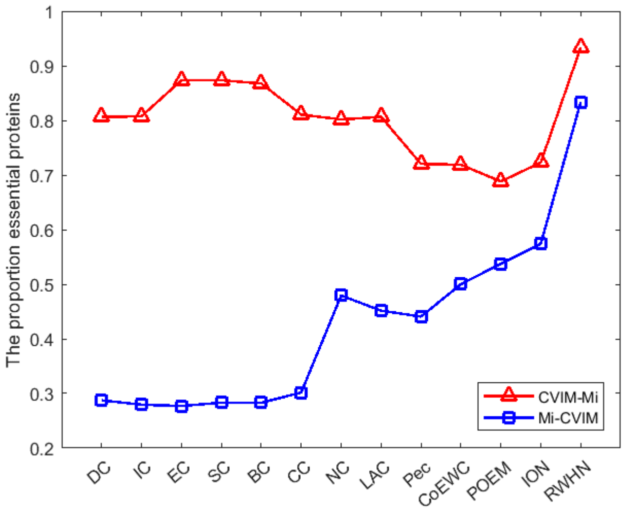


FIGURE 5. The X-axis represents 13 competing methods. The Y-axis represents the proportion of real key proteins in {M_i − CVIM} or {CVIM−M_i}.

B. DIFFERENCES BETWEEN CVIM AND 13 REPRESENTATIVE METHODS

In order to analyze the difference between CVIM and 13 state-of-the-art prediction methods in Table 1, we compared CVIM and 13 methods based on the top 200 ranked proteins. Comparison results are shown in the following table 4 and Fig.4, in which, M_i denotes one of these 13 methods, $\lvert CVIM \cap M_i \rvert$ indicates the number of essential proteins

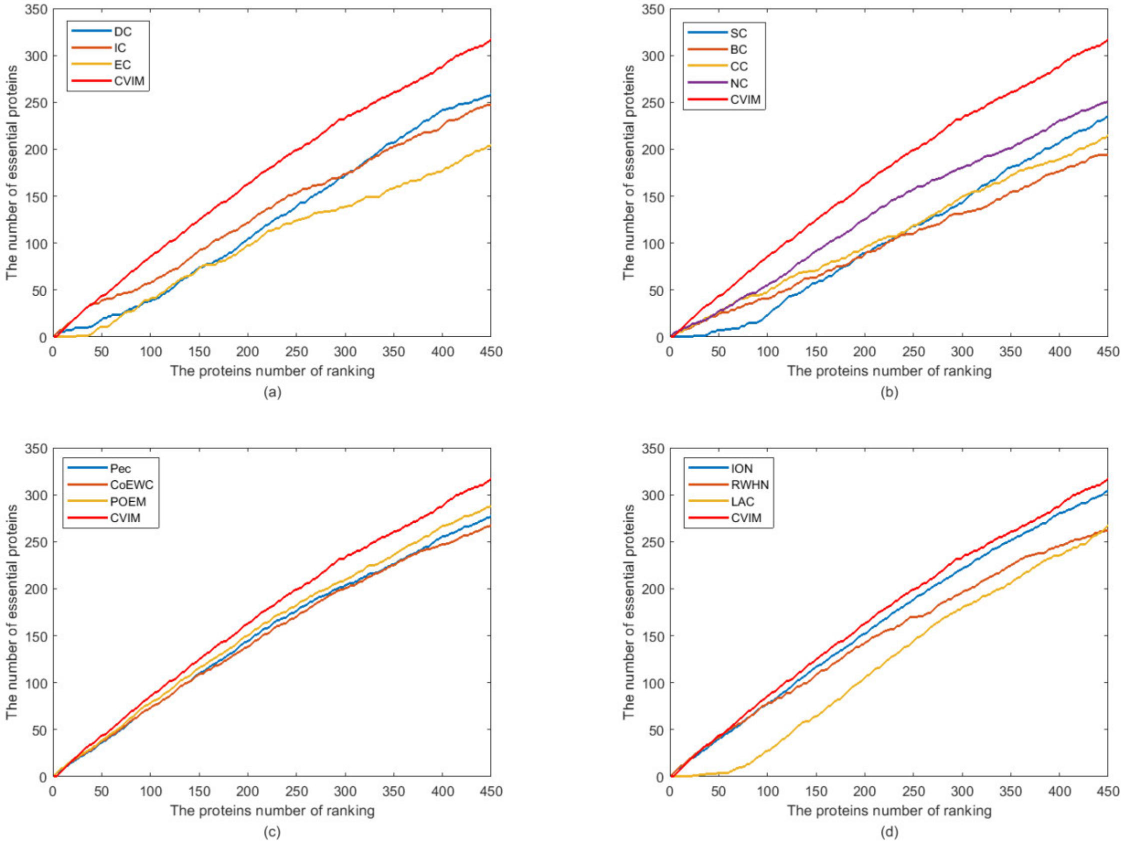


FIGURE 6. Results of comparisons between CVIM and thirteen competing methods based on the top 450 ranked key proteins by implementing the Jackknife methodology on the GAVIN-based PPI network. (*a*) comparison results between CVIM and IC, EC and DC. (*b*) comparison results between CVIM and NC, BC, CC and SC. (*c*) comparison results between CVIM and PEC, CoEWC and POEM. (*d*) comparison results between CVIM and ION, LAC and RWHN.

identified by both CVIM and M_i, $|\mathrm{CVIM}-\mathrm{M}_i|$ represents the number of key proteins detected by CVIM but not detected by M_i, and $|\mathrm{M}_i-\mathrm{CVIM}|$ denotes the number of key proteins identified by M_i but not identified by CVIM. Additionally, $\{\mathrm{CVIM}-\mathrm{M}_i\}$ represents the set of key proteins detected by CVIM but not detected by M_i, while $\{\mathrm{M}_i-\mathrm{CVIM}\}$ denotes the set of key proteins identified by M_i but not identified by CVIM.

From observing table 4 and Fig.5, it is obvious that the percentage of essential proteins in the top 200 ranked proteins discovered by CVIM but not discovered by any given competing method is much higher than the percentage of essential proteins in the top 200 ranked proteins discovered by the given competing method but not discovered by CVIM. That is to say, comparing with state-of-the-art methods, CVIM can detect more true key proteins and has stronger ability to eliminate noise data.

C. PREDICTION PERFORMANCE OF CVIM BASED ON THE GAVIN DATASET

In order to verify the universal applicability of CVIM, in this section, we adopt the GAVIN dataset to compare the predictive performance between CVIM and 13 previous methods. And the comparison results are illustrated in the following table 5 and Fig.6.

As shown in above table 5, in the top 1% (19) ranked proteins, the number of true essential proteins discovered by CVIM is 16, which is higher than that of EC, SC, BC, CC, NC, Pec and LAC, equivalent to that of IC and CoEWC, and a little smaller than that of POEM, RWHN and ION. Although the prediction performance of CVIM in the top 1% ranked proteins is not the best, but the prediction performances of CVIM in the top 5% to 25% ranked proteins are better than all these 13 competing methods.

From observing the Fig.6(a) and Fig.6(b), it is clear that the curves of CVIM are higher than those of DC, EC, SC, BC, IC and NC, which indicate that the performance of CVIM outperforms these methods. From observing the Fig.6(c) and Fig.6(d), we can find as well that the gaps between the curves of CVIM and the curves of PEC, POEM, CoEWC, LAC, RWHN and ION will gradually increase with the increasing of the number of ranked proteins, which demonstrate that with the increasing of ranked proteins, the predictive performance of CVIM will become better and better than that of PEC, POEM, CoEWC, LAC, RWHN and ION. Therefore, we can believe that CVIM is a leading method for predicting potential essential proteins.

TABLE 5. Number of essential proteins predicted by CVIM and 13 methods based on the GAVIN dataset.

Methods	Top1%(19)	Top5%(93)	Top10%(196)	Top15%(279)	Top20%(371)	Top25%(464)
SC	0	17	87	130	190	240
EC	0	38	94	134	166	209
BC	9	40	85	122	162	201
DC	7	36	101	158	222	264
IC	16	55	119	163	213	254
CC	11	45	93	135	180	221
NC	11	51	123	170	213	259
PEC	15	69	142	193	238	285
CoEWC	16	69	136	190	237	275
POEM	17	74	148	199	249	296
ION	17	73	150	207	263	312
RWHN	18	73	140	185	235	269
LAC	0	22	101	167	221	273
CVIM	16	80	160	219	271	322

Table 5: This table shows comparison results between CVIM and 13 competing methods such as DC, EC and SC, BC, CC, IC, NC, Pec, CoEWC, POEM, ION, LAC, RWHN based on the GAVIN dataset. The digits in brackets indicate the number of proteins ranked in each top percentage.

V. DISCUSSION

Essential proteins are indispensable materials to sustain life activities. Up to now, due to the high cost of identifying essential proteins by traditional biological experiments, the recognition of key proteins based on computational techniques has become a hotspot in the research field of proteins. It is an important and challenging work to develop stable and accurate protein identification algorithms by using computational methods instead of biomedical experiments to identify key proteins. More and more researchers are combining PPI networks with biological data to build effective prediction models. Inspired by them, we designed a novel prediction model in this manuscript by integrating the topological features of the weighted PPI network and functional features of the proteins to determine the importance of proteins. Experimental results show that the method can achieve excellent prediction results, which provides a good reference for the future researches.

VI. CONCLUSION

In this manuscript, a novel prediction method called CVIM is proposed to discover potential essential proteins by integrating the PPI network and relevant biological data. In CVIM, a weighted PPI network is constructed first by adopting the PCC scheme on the original PPI network. And then, based on the weighted PPI network and homologous data of proteins and the real-time expression data of genes, for each protein in the weighted PPI network, some network topological features and functional features will be extracted. Finally, based on these different kinds of features, an iterative method is adopted to obtain the final scores of proteins. Based on the DIP2010 and GAVIN yeast PPI networks, intensive experiments have been implemented. Experimental results demonstrate that CVIM outperforms 13 competing representative prediction methods, which shows that CVIM is a unique and effective prediction method as well.

ACKNOWLEDGMENT

(Shiyuan Li and Zhiping Chen are co-first authors.)

REFERENCES

[1] B. Zhao, Y. Zhao, X. Zhang, Z. Zhang, F. Zhang, and L. Wang, "An iteration method for identifying yeast essential proteins from heterogeneous network," *BMC Bioinf.*, vol. 20, no. 1, p. 355, Dec. 2019.

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

[3] W. Zhang, J. Xu, X. Li, and X. Zou, "A new method for identifying essential proteins by measuring co-expression and functional similarity," *IEEE Trans. Nanobiosci.*, vol. 15, no. 8, pp. 939–945, Dec. 2016.

[4] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, "A local average connectivity-based method for identifying essential proteins from the network level," *Comput. Biol. Chem.*, vol. 35, no. 3, pp. 143–150, Jun. 2011.

[5] M. Li, Y. Lu, J. Wang, F.-X. Wu, and Y. Pan, "A topology potential-based method for identifying essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 372–383, Mar. 2015.

[6] B. Xu, J. Guan, Y. Wang, and Z. Wang, "Essential protein detection by random walk on weighted protein-protein interaction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 377–387, Mar. 2019.

[7] Y. Jiang, Y. Wang, and W. Pang, "Essential protein identification based on essential protein–protein interaction prediction by integrated edge weights," *Methods*, vol. 83, pp. 51–62, Jul. 2015.

[8] L. Chen, Y.-H. Zhang, S. Wang, Y. Zhang, T. Huang, and Y.-D. Cai, "Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways," *PLoS ONE*, vol. 12, no. 9, 2017, Art. no. e0184129.

[9] H. Jeong, S. Mason, A. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[10] 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, Apr. 2005.

[11] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 5, pp. 33–122, May 2005.

[12] 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.

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

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

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

[16] J. X. Wang, M. Li, and H. Wang, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1070–1080, Jul./Aug. 2012.

[17] M. Li, H. Zhang, J.-X. Wang, and Y. Pan, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data," *BMC Syst. Biol.*, vol. 6, no. 1, p. 15, 2012.

[18] X. Zhang, J. Xu, and W.-X. Xiao, "A new method for the discovery of essential proteins," *PLoS ONE*, vol. 8, no. 3, 2013, Art. no. e58763.

[19] S. Keretsu and R. Sarmah, "Weighted edge based clustering to identify protein complexes in protein–protein interaction networks incorporating gene expression profile," *Comput. Biol. Chem.*, vol. 65, pp. 69–79, Dec. 2016.

[20] B. Cai, H. Wang, H. Zheng, and H. Wang, "Identification of protein complexes from tandem affinity purification/mass spectrometry data via biased random walk," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 455–466, Mar. 2015.

[21] J. Luo and Y. Qi, "Identification of essential proteins based on a new combination of local interaction density and protein complexes," *PLoS ONE*, vol. 10, no. 6, 2015, Art. no. e0131418.

[22] B. Zhao, J. Wang, M. Li, F.-X. Wu, and Y. Pan, "Prediction of essential proteins based on overlapping essential modules," *IEEE Trans. Nanobiosci.*, vol. 13, no. 4, pp. 415–424, Dec. 2014.

[23] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Liu, and S. Sang, "A method for predicting protein complex in dynamic PPI networks," *BMC Bioinf.*, vol. 17, no. S7, p. 229, Jul. 2016.

[24] C.-Y. Ma, Y.-P. Phoebe Chen, B. Berger, and C.-S. Liao, "Identification of protein complexes by integrating multiple alignment of protein interaction networks," *Bioinformatics*, vol. 33, no. 11, pp. 1681–1688., Jan. 2017.

[25] X. Lei, M. Fang, L. Guo, and F.-X. Wu, "Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks," *BMC Bioinf.*, vol. 20, no. S3, p. 131, Mar. 2019.

[26] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Syst. Biol.*, vol. 6, no. 1, p. 87, 2012.

[27] J. Luo and J. Wu, "A new algorithm for essential proteins identification based on the integration of protein complex co-expression information and edge clustering coefficient," *Int. J. Data Mining Bioinf.*, vol. 12, no. 3, pp. 257–274, 2015.

[28] B. Xu, H. Lin, Y. Chen, Z. Yang, and H. Liu, "Protein complex identification by integrating protein-protein interaction evidence from multiple sources," *PLoS ONE*, vol. 8, no. 12, 2013, Art. no. e83841.

[29] B. Xu, K. Li, W. Zheng, X. Liu, Y. Zhang, Z. Zhao, and Z. He, "Protein complexes identification based on go attributed network embedding," *BMC Bioinf.*, vol. 19, no. 1, p. 535. Dec. 2018.

[30] X. Lei and J. Liang, "Neighbor affinity-based core-attachment method to detect protein complexes in dynamic PPI networks," *Molecules*, vol. 22, no. 7, p. 1223, 2017.

[31] L. Ou-Yang, H. Yan, and X.-F. Zhang, "A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks," *BMC Bioinf.*, vol. 18, no. S13, p. 463, Nov. 2017.

[32] S. Srihari, K. Ning, and H. W. Leong, "MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure," *BMC Bioinf.*, vol. 11, no. 1, p. 504, Dec. 2010.

[33] D. Horyu and T. Hayashi, "Comparison between pearson correlation coefficient and mutual information as a similarity measure of gene expression profiles," *Jpn. J. Biometrics*, vol. 33, no. 2, pp. 125–143, 2013.

[34] I. Xenarios, "DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, Jan. 2002.

[35] A.-C. Gavin *et al.*, "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, Mar. 2006.

[36] G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer, "InParanoid 7: New algorithms and tools for eukaryotic orthology analysis," *Nucleic Acids Res.*, vol. 38, no. Database, pp. D196–D203, Jan. 2010.

[37] B. P. Tu, "Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, Nov. 2005.

[38] H. W. Mewes, "MIPS: Analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D169–D172, Jan. 2006.

[39] *Saccharomyces Genome Deletion Project*. Accessed: Jun. 20, 2012. [Online]. Available: http://yeastdeletion.stanford.edu/

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

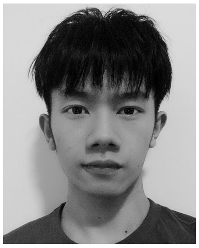
[41] J. Cherry, "SGD: Saccharomyces genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, Jan. 1998.

[42] A. G. Holman, P. J. Davis, J. M. Foster, C. K. Carlow, and S. Kumar, "Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia Malayi," *BMC Microbiol.*, vol. 9, no. 1, p. 243, 2009.

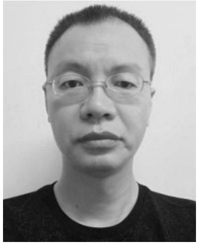
[43] G. T. Hart, I. Lee, and E. R. Marcotte, "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality," *BMC Bioinf.*, vol. 8, no. 1, p. 236, 2007.

[44] Z. Dezso, "Bioinformatics analysis of experimentally determined protein complexes in the yeast saccharomyces cerevisiae," *Genome Res.*, vol. 13, no. 11, pp. 2450–2454, Nov. 2003.

[45] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, and J. Li, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.



SHIYUAN LI is currently pursuing the bachelor's degree major in information and computing science with Changsha University. His current research interest is bioinformatics.



ZHIPING CHEN received the B.S. degree in computer science and technology from Xiangtan University, Xiangtan, Hunan, in 1994, and the M.S. and Ph.D. degrees in computer science and technology from Hunan University, in 1997 and 2003, respectively.

From 1997 to 2009, he has taught at Hunan University. He is currently a Professor with Changsha University. His current research area is mainly bioinformatics.

XIN HE is currently pursuing the master's degree major in computer science and technology with the College of Information and Engineering, Xiangtan University. His current research interest is bioinformatics.

ZHEN ZHANG received the B.S. degree from Anhui Agricultural University, Hefei, Anhui, in 2006, the M.S. degree from Central South University, in 2009, and the Ph.D. degree from Northwestern Polytechnical University, in 2019.

His current research area is mainly big data.

TINGRUI PEI received the B.S. and M.S. degrees from Xiangtan University, in 1992, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, in 2004.

From 2006 to 2007, he was a Visiting Scholar with Waseda University. He is currently a Professor with Xiangtan University. He is also the author of 13 invention patents and more than 30 articles. His main research areas are the Internet of Things, wireless sensor networks (WSNs), mobile *ad hoc* networks, mobile communication networks, and social computing.

YIHONG TAN received the B.S. degree from Nanjing Agricultural University, Nanjing, Jiangsu, in 1994, and the M.S. and Ph.D. degrees in computer science and technology from Hunan University, in 2003 and 2010, respectively.

He is currently a Professor with Changsha University. His current research area is mainly complex networks.

LEI WANG received the Ph.D. degree in computer science from Hunan University, China, in 2005. From 2005 to 2007, he was a Postdoctoral Fellow with Tsinghua University, China. After that, he moved to USA and Canada as a Visiting Scholar with Duke University and Lakehead University successively. From 2005 to 2011, he was an Associate Professor with the College of Software, Hunan University. From 2011 to 2018, he was a Full Professor with the College of Information Engineering, Xiangtan University. He is currently a Full Professor and an Academic Leader of computer engineering with Changsha University, China. He has published more than 100 peer-reviewed articles. His main research areas include bioinformatics and the Internet of Things.