IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received December 5, 2019, accepted December 24, 2019, date of publication January 7, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964571

A Novel Model for Predicting Essential Proteins Based on Heterogeneous Protein-Domain Network

ZHIPING CHEN[1], ZIXUAN MENG[2], CHAOPING LIU[1], XIANGYI WANG[1], LINAI KUANG[2], TINGRUI PEI[2], 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

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

The 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 Essential proteins play significant roles in cell survive. In current years, some Protein-Protein Interaction (PPI) data have been discovered in saccharomyces cerevisiae. Due to the high costs of biological experiments, a growing number of computational models are adopted to predict essential proteins. However, the identification accuracy of these computational models still has broad space for improvement. In this paper, a novel prediction model called NPRI is proposed to infer potential essential proteins based on the PageRank algorithm. In NPRI, a new heterogeneous Protein-Domain network will be constructed by integrating three kinds of networks such as the weighted PPI network, the Domain-Domain network and the initial Protein-Domain network first. Here, these three kinds of networks are established in accordance with gene expression data, original PPI network and known Protein-Domain network respectively. Next, based on the newly constructed heterogeneous Protein-Domain network, we will extract functional features and topological characteristics for each protein to further construct a novel distribution rate network. And then, an improved iteration method based on the PageRank algorithm will be implemented on the novel distribution rate network to infer essential proteins. Finally, in order to evaluate the performance of NPRI, we will compare NPRI with other state-of-the-art prediction models, and simulation results show that NPRI can achieve reliable identification accuracies of 90%, 84.5% and 79% in top 100, 200 and 300 predicted candidate essential proteins separately, which outperform these competitive models remarkably, and means that NPRI is a promising framework for identifying essential proteins as well.

INDEX TERMS Essential proteins, protein-protein interaction network, domain-domain network, heterogeneous protein-domain network, pagerank algorithm.

I. INTRODUCTION

Increasing evidences indicate that proteins are involved in almost all life activities, while the functions and importance of different proteins in life activity are different. As an important group of proteins, essential proteins play a vitally important role in the development and survival of organisms, which can provide fundamental requirements for sustaining life and have practical value in synthetic biology. Lack of these proteins will result in the losing of biological function of the protein complex and even death of the organism. Thus, predicting essential proteins has gradually become a hot issue, which is advantageous to the disease treatment and medicine development. In biology, essential proteins are identified mainly through biological experiments, such as single gene knockout RNA interference, conditional knockout, etc. However, biological experiments are very time-consuming and expensive. Hence, in recent years, a large number of computational methods have been proposed successively. Among them, the accuracy for detecting key proteins is still a critical and challenging task [1]–[4].

Up to now, existing models for predicting essential proteins can be roughly classified into two major categories. Models of the first category focus on adopting topological characteristics of the PPI network to predict key proteins only. For instance, based on the centrality-lethality

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

rule presented by Jeong H et al. [5], a variety of methods including DC (Degree Centrality) [6], IC (Information Centrality) [7], CC (Closeness Centrality) [8], BC (Betweenness Centrality) [9], SC (Subgraph Centrality) [10], NC (Neighbor Centrality) [11] and EC (Eigenvector Centrality) [12] have been proposed to infer potential essential proteins successively. Since all these PPI network topology based methods do not need additional biological information, they can break through the limitation of traditional biological experiments and achieve great progress on identification of essential proteins. However, due to the lack of integrated PPI networks, false positive and negative data involved in PPI networks may bring negative influences to the prediction results. Therefore, these methods based on topological features of PPI network cannot achieve satisfactory prediction results while being applied to identify essential proteins.

The second category of prediction methods aim to improve the prediction accuracy of the first category of prediction methods by combining PPI networks with some biological information, including the sub-cellular localization, the evolutionary conservation and gene expression, to infer key proteins. For example, M Li *et al.* [13] proposed a prediction model named PEC through combining the PPI network with the gene expression to identify basic proteins. Zhang *et al.* [14] developed a calculative model called CoEWC based on both PPI network and the gene expression profile to detect potential key proteins. W Zhang *et al.* [15] put forward an algorithm to discover essential proteins by integrating the PPI Network with gene expression data and gene otology information. Lei *et al.* [16] raised a computational model RSG to recognize key proteins through combining the information of RNA-Seq, Subcellular Localization and GO annotation datasets to build a novel weighted PPI network. J Zhong *et al.* [17] proposed a calculative model by combining various biological data to predict key proteins. Y Fan *et al.* [18] took advantage of the subcellular localization and Person correlation coefficient to construct a new weighted PPI network to identify essential proteins. Shabnam and Izudheen [19] proposed a novel prediction model by integrating both gene expression profile and domain information to infer potential key proteins based on the hypothesis that key proteins are inclined to form dense cluster. W Zhang *et al.* [15] incorporated three kinds of data such as the gene expression data, the Go annotation data, and the topological feature data to calculate the essentiality of proteins. Lei *et al.* [20] introduced a model called GSP to predict essential proteins based on multiply biological data and centralities of proteins. M Li *et al.* [21] presented a computational model for identifying key proteins based on a refine protein interaction network (PIN) and various biological data. Luo and Ling [22] proposed a prediction model called CSC for essential protein prediction by adopting topological characteristics of the PPI network and the complex information of proteins. Zhao *et al.* [23] constructed a reliable weighted network based on gene expression data and topological properties of the weighted network first, and then designed a calculation method called POEM to predict key proteins based on overlapping basic modules. Jun *et al.* [24] adopted the protein complex information and subgraph centrality to develop a new algorithm named ECC to infer essential proteins. Q H Xiao and Wang [4] put forward a framework for predicting key proteins based on active PPI network and dynamic gene expression information. Mistry *et al.* [25] introduced a graph centrality approach to infer potential basic proteins by integrating the PPI network and gene expression data. Qin *et al.* [26] developed a calculation model to discover essential proteins in accordance with topological attributes refined from PPI network and biological information including subcellular localization data and orthologous data.

From above descriptions, we can come to a conclusion that integrating biological data, such as gene expression profile data, subcellular information, and orthologous data etc., with PPI networks can remarkably enhance prediction accuracy for inferring necessary proteins. Different from these prediction methods mentioned above, considering that a sole PPI network cannot completely reflect the diversity of biological characteristics and functional features of proteins, in this manuscript, a new prediction model called NPRI will be proposed to infer key proteins based on a heterogeneous Protein-Domain network. Here, the heterogeneous Protein-Domain network is constructed by combining a novel weighted Protein-Protein network with the initial Protein-Domain association network and Domain-Domain association network. Moreover, for each protein in the heterogeneous Protein-Domain network, different from other model based on the heterogeneous network, for instance, RWHN [1], we will integrate the subcellular localization information, orthologous information and some critical topological properties extracted from the original PPI network to obtain its initial score first. And then, based on the heterogeneous Protein-Domain network, an iteration algorithm based on PageRank will be further constructed to detect potential key proteins. Finally, in order to estimate the prediction performance of NPRI, we will compare it with 13 competitive prediction models including DC [6], IC [7], CC [8], BC [9], SC [10], NC [11], EC [12], PEC [13], CoEWC [14], POEM [23], ION [27], LAC [28] and RWHN [1]. Simulation results illustrate that NPRI can achieve reliable prediction accuracies of 90%, 84%, 79%, 75% and 70.6% in top 100, top 200, top 300 and top 400 candidate inferred essential proteins respectively, which outperform all these state-of-the-art prediction models remarkably.

II. METHOD

As illustrated in the following Fig.1, NPRI consists of four major steps:

Step1: First, two original PPI networks will be constructed based on the datasets of known PPIs downloaded from two public databases separately. And then, for each pair of proteins in any given original PPI network N_I, the Gaussian

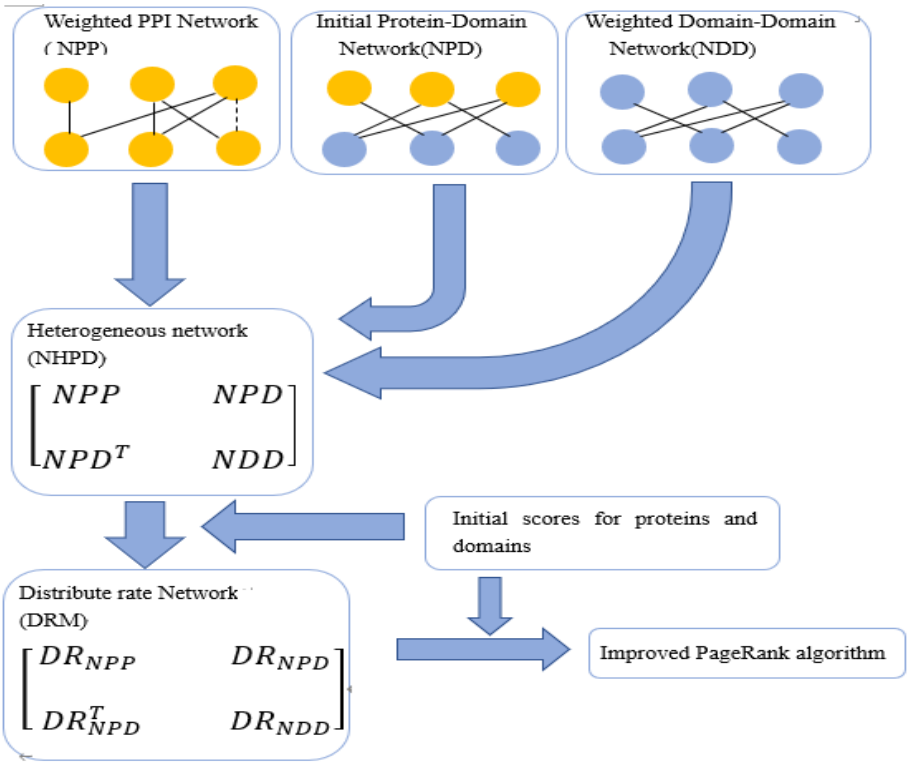


FIGURE 1. Flowchart of NPRI.

interaction profile kernel similarity [29] will be implemented on gene expression data downloaded from public databases to calculate weight between them. Thereafter, based on the original PPI network N_I, a novel weighted PPI network N_{PP} can be obtained.

Step2: Next, based on the domain information and known Protein-Domain associations downloaded from public databases, a weighted Domain-Domain association network N_{DD} and an initial Protein-Domain network N_{PD} will be further constructed respectively. And then, through integrating these three kinds of newly constructed networks such as N_{PP}, N_{DD} and N_{PD}, a novel heterogeneous Protein-Domain network N_{HPD} can be obtained.

Step3: Moreover, based on the original PPI network N_I, we can extract some critical topological features for each protein first, and then, through combining the information of subcellular localization and orthologous downloaded from public databases, an initial score can be calculated for each protein and domain in N_{HPD}.

Step4: Finally, based on the heterogeneous Protein-Domain network N_{HPD}, a novel PageRank based iteration algorithm will be designed to predict potential essential proteins.

A. CONSTRUCTION OF THE WEIGHTED PPI NETWORK

In this section, we first download two datasets of known PPIs from two public databases such as the Gavin [30] database and the DIP [31] database separately. After screening, we finally obtain 1167 essential proteins and 24743 interactions between 5093 proteins from the DIP database, and 617 essential proteins and 24743 interactions between 1855 proteins from the Gavin database. Based on the datasets of known PPIs downloaded from above two databases, we further construct two different original PPI networks respectively. For convenience, we define $N_I = \{P_I, L_I\}$ as the original PPI network based on known PPIs downloaded from the database I, where $P_I = \{p_1, p_2, \ldots, p_K\}$ denotes the set of proteins downloaded from the database I, and L_I represents the set of edges between proteins in P_I. Obviously, for any given proteins p_i and p_j in P_I, there is an edge between p_i and p_j in L_I, if and only if there is a known interaction between them in the database I. Thereafter, based on the newly obtained original PPI network N_I, a $K \times K$ dimensional adjacency matrix $NI = (a_{ij})_{K\times K}$ can be constructed easily, where there is $a_{ij} = 1$, if and only if there is an edge between the proteins p_i and p_j, otherwise, there is $a_{ij} = 0$.

Next, we will download the dataset of gene expressions provided by Tu BP et al. [32] For each protein p P_I, let $Ex(p, i)$ represent its gene expression level at the i-th time, then the gene expression data of the protein p can be represented as $Ex(p) = \{Ex(p,1), Ex(p,2), \ldots, Ex(p, m)\}$. Hence, inspired by the concept of Gaussian interaction profile kernel similarity, for any two given proteins p_i and p_j in P_I, the weight between

them can be assigned as follows:

$$NPP\left(p_i, p_j\right) = \begin{cases} \alpha * \exp\left(-\gamma_p \left\|Ex\left(p_i\right) - Ex\left(p_j\right)\right\|^2\right) + (1-\alpha): \\ \qquad \text{if } a_{ij} = 1 \wedge \exists Ex\left(p_i\right) \wedge \exists Ex\left(p_j\right) \\ 1-\alpha: \qquad \text{if } a_{ij} = 1 \wedge \left(Ex\left(p_i\right) \vee Ex\left(p_j\right)\right) \\ 0: \qquad \text{if } a_{ij} = 0 \wedge \left(Ex\left(p_i\right) \vee Ex\left(p_j\right)\right) \\ \alpha * \exp\left(-\gamma_p \left\|Ex\left(p_i\right) - Ex\left(p_j\right)\right\|^2\right): \\ \qquad \text{if } a_{ij} = 0 \wedge \left(\exists Ex\left(p_i\right) \wedge \exists Ex\left(p_j\right)\right). \end{cases} \quad (1)$$

Here,

$$\gamma_p = \gamma_p' / \left(\frac{1}{NE_p} \sum_{t=1}^{NE_p} \left\|Ex\left(p_t\right)\right\|^2\right) \quad (2)$$

where γ_p denotes the normalized kernel bandwidth on the basis of new bandwidth parameter γ_p', NE_p represents the number of proteins having known gene expression, and the $\alpha \in [0, 1]$ is the parameter of distribution proportion.

Obviously, based on above equation (2), we can obtain a $K \times K$ dimensional adjacency matrix NPP, based on which, we can further construct a novel weighted PPI network N_{PP} as well.

B. CONSTRUCTION OF THE INITIAL PROTEIN-DOMAIN ASSOCIATION NETWORK

In this section, we first download the dataset of known domain information from the Pfam Database [33]. After screening, we obtain 4936 protein-domain associations including 3630 proteins and 1107 domains. For convenience, let $D = \{d_1, d_2, \ldots, d_N\}$ denote all these newly obtained domains and L_{PD} represent the set of edges between domains and proteins, then we can construct an initial protein-domain association network $N_{PD} = \{P_I, L_{PD}\}$ as follows: for any given protein $p_i \in P_I$ and domain $d_j \in D$, there is an edge between p_i and d_j in L_{PD}, if and only if there is a known association between them in these newly downloaded 4936 protein-domain associations.

Obviously, based on the initial protein-domain association network N_{PD}, we can further construct a $K \times N$ dimensional adjacency matrix $NPD = (b_{ij})_{K \times N}$, where there is $b_{ij} = 1$, if and only if there is an edge between the protein $p_i \in P_I$ and domain $d_j \in D$, otherwise, there is $b_{ij} = 0$.

C. CONSTRUCTION OF THE DOMAIN-DOMAIN ASSOCIATION NETWORK

For any two given domains d_i and d_j in D, let D_i and D_j represent the sets of proteins in d_i and d_j respectively, and $D_i \cap D_j$ denote the proteins in both d_i and d_j, then we can calculate the weight between the domains d_i and d_j as follows:

$$NDD\left(d_i, d_j\right) = \begin{cases} max_{p_x, p_y \in (D_i \cap D_j)} \left(NPP\left(p_x, p_y\right)\right): \\ \qquad if\ D_i \bigcap D_j \neq \emptyset \wedge (d_i \neq d_j) \\ 0: \qquad Otherwise \end{cases} \quad (3)$$

Obviously, based on above equation (3), we can further construct a novel Domain-Domain association network N_{DD} and obtain a corresponding $N \times N$ dimensional adjacency matrix NDD simultaneously.

D. CONSTRUCTION OF THE HETEROGENEOUS PROTEIN-DOMAIN NETWORK

Based on these newly constructed $K \times K$ dimensional adjacency matrix NPP, $K \times N$ dimensional adjacency matrix NPD and $N \times N$ dimensional adjacency matrix NDD, we can easily obtain a $(K+N) \times (K+N)$ dimensional heterogeneous matrix $NHPD$ as follows:

$$NHPD = \begin{bmatrix} NPP & NPD \\ NPD^T & NDD \end{bmatrix} \quad (4)$$

Obviously, based on above equation (4), we can obtain a heterogeneous Protein-Domain network N_{HPD}.

E. CALCULATING INITIAL SCORES FOR PROTEINS AND DOMAINS

In order to assign initial scores for protein and domain nodes in N_{HPD}, in this section, we first download subcellular localization information and orthology information from the COMPART-MENTS database [34] and the InParanoid database (Version 7) [37] separately. And then, let $S = \{s_1, s_2, \ldots, s_n\}$ denote the set of downloaded subcellular localizations, $pro(s_i)$ represent the set of proteins related to the subcellular localization s_i, and $|pro(s_i)|$ denote the number of proteins in $pro(s_i)$, it is easy to know that we can obtain the average number of proteins associated with each subcellular localization as follows:

$$Avg\_sub = \left[\sum_{i=1}^{n} pro(s_i)\right] / n \quad (5)$$

Next, based on above equation (5), for each subcellular localization $s_i \in S$, we define its rank as follows:

$$Rsub(s_i) = \frac{rsub(s_i)}{max_{1 \le j \le n}(rsub(s_j))} \quad (6)$$

where,

$$rsub(s_i) = \frac{|pro\left(s_i\right)|}{avg\_sub} \quad (7)$$

So far, based on above equation (6), for any given protein $p_i \in P_I$, we can define its feature of the subcellular localization as follows:

$$pro\_sub(p_i) = \sum_{s_j \in S(p_i)} Rsub(s_j) \quad (8)$$

where $S(p_i)$ is the set of subcellular localizations related to p_i.

Similar to above description, for each protein $p_i \in P_I$, let $ort(p_i)$ denote its conservative score downloaded from the InParanoid database, then we can as well define its feature of orthology information as follows:

$$pro\_ort(p_i) = \frac{ort(p_i)}{max_{p_j \in P_I}(ort(p_j))} \quad (9)$$

Furthermore, for any given protein p_i in $N_I = \{P_I, L_I\}$, we define the set of its neighboring nodes as follows:

$$NS(p_i) = \{p_j | \exists l(p_i, p_j) \in L_I, p_j \in P_I \quad (10)$$

Based on above equation (10), considering that triangle has the characteristics of both simplicity and stability, for each protein p_i in $N_I = \{P_I, L_I\}$, then we can obtain the number of triangles related to p_i in N_I as follows:

$$num\_tri(p_i) = \sum\nolimits_{p_j \in NS(p_i)} |NS(p_i) \cap NS(p_j)| \quad (11)$$

where $|NS(p_i) \cap NS(p_j)|$ is the number of nodes in $NS(p_i) \cap NS(p_j)$.

So far, based on above equation (11), for each protein p_i in $N_I = \{P_I, L_I\}$, we can define its topological feature of average triangles as follows:

$$pro\_tri(p_i) = \frac{avg\_tri(p_i)}{max_{p_j \in P_I}(avg\_tri(p_j))} \quad (12)$$

Here,

$$avg\_tri(p_i) = \frac{num\_tri(p_i)}{|NS(p_i)|} \quad (13)$$

where $|NS(p_i)|$ is the number of nodes in $NS(p_i)$.

Finally, based on above equations (8), (9) and (12), for each protein p_i in $N_I = \{P_I, L_I\}$, we define its initial score as follows:

$$NPR^0_{p_i} = \beta * pro\_sub(p_i) + \gamma * pro\_ort(p_i) + \delta * pro\_tri(p_i) \quad (14)$$

where $\beta \in [0, 1]$, $\gamma \in [0, 1]$ and $\delta \in [0, 1]$ are parameters satisfying $\beta + \gamma + \delta = 1$.

Based on above equations (14), for any given domain d_i in N_{HPD}, we define its initial score as follows:

$$NPR^0_{d_i} = \frac{\sum_{p \in D_i} NPR^0_p}{\max_{1 \le j \le N} \sum_{q \in D_j} NPR^0_q} \quad (15)$$

F. CONSTRUCTION OF (K+N)×(K+N) DIMENSIONAL DISTRIBUTION RATE MATRIX DRM

In this section, based on the weighted PPI network N_{PP}, for any given protein p_i in P_I, we first define a new set of proteins related to p_i as follows:

$$N\_NP(p_i) = \{p_j | NPP(p_i, p_j) \neq 0, p_j \in P_I\} \quad (16)$$

Next, let $MNPP = \max_{p_e, p_f \in P_I} NPP(p_e, p_f)$ and $BN(p_i, p_j) = NPP(p_i, p_j)/(1 + MNPP)^2$, then for any two given proteins p_i and p_j in P_I, we can further define the distribution rate between them in N_{PP} as follows:

$$DR_{NPP}(p_i, p_j) = \begin{cases} BN(p_i, p_j) * \frac{NPR^0_{p_j}}{\sum_{p_t \in N\_NP(p_i)} NPR^0_{p_t}} : & If\ NPP(p_i, p_j) \neq 0 \\ 0 : & Otherwise \end{cases} \quad (17)$$

Similarly, based on the Protein-Domain association network N_{PD}, for any given protein $p_i \in P_I$ and domain $d_j \in D$, we can define the weight between them as follows:

$$WNPD(p_i, d_j) = \begin{cases} NPR^0_{p_i} / \sum_{p_t \in D_j} NPR^0_{p_t} : & If\ b_{ij} = 1 \\ 0 : & Otherwise \end{cases} \quad (18)$$

Based on above equation (18), let $MNPD = \max_{p_i \in P_I, d_j \in D} WNPD(p_i, d_j)$, then for any given protein $p_i \in P_I$ and domain $d_j \in D$, we can define the distribution rate between them in N_{PD} as follows:

$$DR_{NPD}(p_i, d_j) = \begin{cases} WNPD(p_i, d_j)/(1+MNPD)^2 : & If\ WNPD(p_i, d_j) \neq 0 \\ 0 : & Otherwise \end{cases} \quad (19)$$

Moreover, based on domain-domain association network N_{DD}, for any given domains d_i and d_j in D, let $MNDD = \max_{d_i, d_j \in D} NDD(d_i, d_j)$, then we can define the weight between them as follows:

$$DBN(d_i, d_j) = \begin{cases} NDD(d_i, d_j)/(1+MNDD)^2 : & If\ NDD(d_i, d_j) \neq 0 \\ 0 : & Otherwise \end{cases} \quad (20)$$

Next, for any given domain d_i in D, we define a new set of domains related to d_i as follows:

$$N\_D(d_i) = \{d_j | NDD(d_i, d_j) \neq 0, d_j \in D\} \quad (21)$$

Thereafter, for any given domains d_i and d_j in D, we can define the distribution rate between them in N_{DD} as follows:

$$DR_{NDD}(d_i, d_j) = \begin{cases} DBN(d_i, d_j) * \frac{NPR^0_{d_j}}{\sum_{d_t \in N\_D(d_i)} NPR^0_{d_t}} : & IfDBN(d_i, d_j) \neq 0 \\ 0 : & Otherwise \end{cases} \quad (22)$$

So far, based on above equations (17), (19) and (22), we can obtain a new distribution rate matrix DRM as follows:

$$DRM = \begin{bmatrix} DR_{NPP} & DR_{NPD} \\ DR^T_{NPD} & DR_{NDD} \end{bmatrix} \quad (23)$$

Based on the PageRank algorithm, let a denote any given protein node or domain node in the heterogeneous Protein-Domain network N_{HPD}, then we can calculate its rank iteratively according to the following equation (24):

$$NPR_a(t+1) = \varphi * DRM * NPR_a(t) + (1-\varphi) * NPR_a(t) \quad (24)$$

where $NPR_a(t)$ is the score vector of node a at the t-th time, and $\varphi \in [0, 1]$ is the proportional adjustment parameter.

Based on above equations, our prediction model NPRI can be described in detail as follows:

Algorithm 1 NPRI

Input: Downloaded dataset of known PPIs, downloaded orthologous dataset, downloaded subcellular localization dataset, downloaded domain dataset, downloaded gene expression dataset, the iteration termination condition ε, and the proportional adjustment parameter $\alpha, \beta, \gamma, \delta, \varphi$.

Output: Top K percent of proteins sorted by the vector NPR_a in descending order

Step1: Establish the heterogeneous Protein-Domain network N_{HPD} according to equations (1)-(4);

Step2: For each protein in N_{HPD}, calculate its initial score according to equations (5)∼(14);

Step3: For each domain in N_{HPD}, calculate its initial score according to equation (15);

Step4: Construct the distribution rate matrix according to equations (16)-(23);

Step5: Let t=t+1, calculate $NPR_a(t+1)$ according to equation (25) iteratively;

Step6: Repeat step 5 until $\|NPR_a(t+1) - NPR_a(t)\|_2 < \varepsilon$;

Step7: Sort proteins by the value of NPR_a in descending order;

Step8: Output top K percent of sorted proteins.

TABLE 1. Influence of α to the prediction accuracy of NPRI based on the DIP database.

α / Top ranks	0.1	0.2	0.3	0.4	0.5
1% (51)	0.86	0.86	0.86	0.82	0.78
5% (255)	0.82	0.81	0.81	0.8	0.79
10% (510)	0.7	0.7	0.69	0.69	0.69
15% (764)	0.62	0.61	0.61	0.61	0.61
20% (1019)	0.55	0.55	0.55	0.55	0.55
25% (1274)	0.5	0.5	0.5	0.5	0.5

TABLE 2. Influence of α to the prediction accuracy of NPRI based on the Gavin database.

α / Top ranks	0.1	0.2	0.3	0.4	0.5
1% (19)	0.89	0.84	0.84	0.84	0.73
5% (93)	0.8	0.8	0.8	0.79	0.79
10% (186)	0.82	0.82	0.81	0.8	0.77
15% (278)	0.8	0.79	0.78	0.78	0.78
20% (371)	0.75	0.74	0.74	0.74	0.73
25% (464)	0.69	0.69	0.69	0.69	0.7

III. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA

In order to estimate the performance of NPRI, in this section, we first compare it with 13 competitive prediction methods such as RWHN [1], DC [6], IC [7], CC [8], BC [9], SC [10], NC [11], EC [12], PEC [13], CoEWC [14], POEM [23], ION [27] and LAC [28] respectively. Based on these 1855 proteins downloaded from the Gavin database [30], 5093 proteins downloaded from the DIP database [31], and 1107 domains downloaded from the Pfam database [33], two kinds of heterogeneous matrixes can be constructed, one is a (5093+1107)×(5093+1107) dimensional matrix based on the DIP database and the other is a (1855+1107)×(1855+1107) dimensional matrix based on the Gavin database.

Next, we further download the gene expression data from the dataset provided by Tu BP [32] and the subcellular localization data from COMPART-MENTS database [34]. In the newly downloaded gene expression data, proteins with gene expression data account for 95% of total number of proteins in both Gavin and DIP databases. However, as for the newly downloaded subcellular localization data, we only take advantage of 11 different subcellular localizations that are closely related to essential proteins, including Endoplasmic, Cytoskeleton, Golgi, Cytosol, Vacuole, Mitochondrion, Endossome, Plasma, Nucleus, Peroxisome and Extracellular, to define initial scores for proteins. Additionally, we also download the information of orthologous proteins including a collection of pairwise comparisons between 100 whole genomes from the InParanoid database [35], which will be used for computing initial scores for proteins. Finally, a dataset containing 1293 essential genes of Saccharomyces cerevisiae will be downloaded from four databases such as MIPS [36], SGD [37], DEG [38] and SGDP [39] as the benchmark set. Moreover, we will present the simulation results of NPRI based on DIP in detail, while present the experimental results of NPRI based on GAVIN briefly.

B. EFFECTS OF THE PARAMETER α

In NPRI, we define a parameter α with value between 0 and 1 to adjust the allocated proportion during iteration. Through assigning different values to α, the prediction results based on the DIP database and the Gavin database are illustrated in the following Table 1 and Table 2 respectively. And as shown in Table 1 and Table 2, we pick out the top 1%, 5%, 10%, 15%, 20% and 25% real essential proteins detected by NPRI when α is set to 0.1, 0.2, 0.3, 0.4 and 0.5 respectively. Obviously, the identification rate of NPRI will vary with different values of α. And with the increasing of the value of α, the recognition rate of NPRI will decrease gradually. Hence, it is easy to see that NPRI can reach the best prediction performance while $\alpha = 0.1$.

C. COMPARISON WITH STATE-OF-THE-ART PREDICTION METHODS

In this section, we make use of the dataset obtained from the DIP database to compare NPRI with 13 state-of-the-art prediction methods including CC, IC, SC, EC, BC, NC, DC,

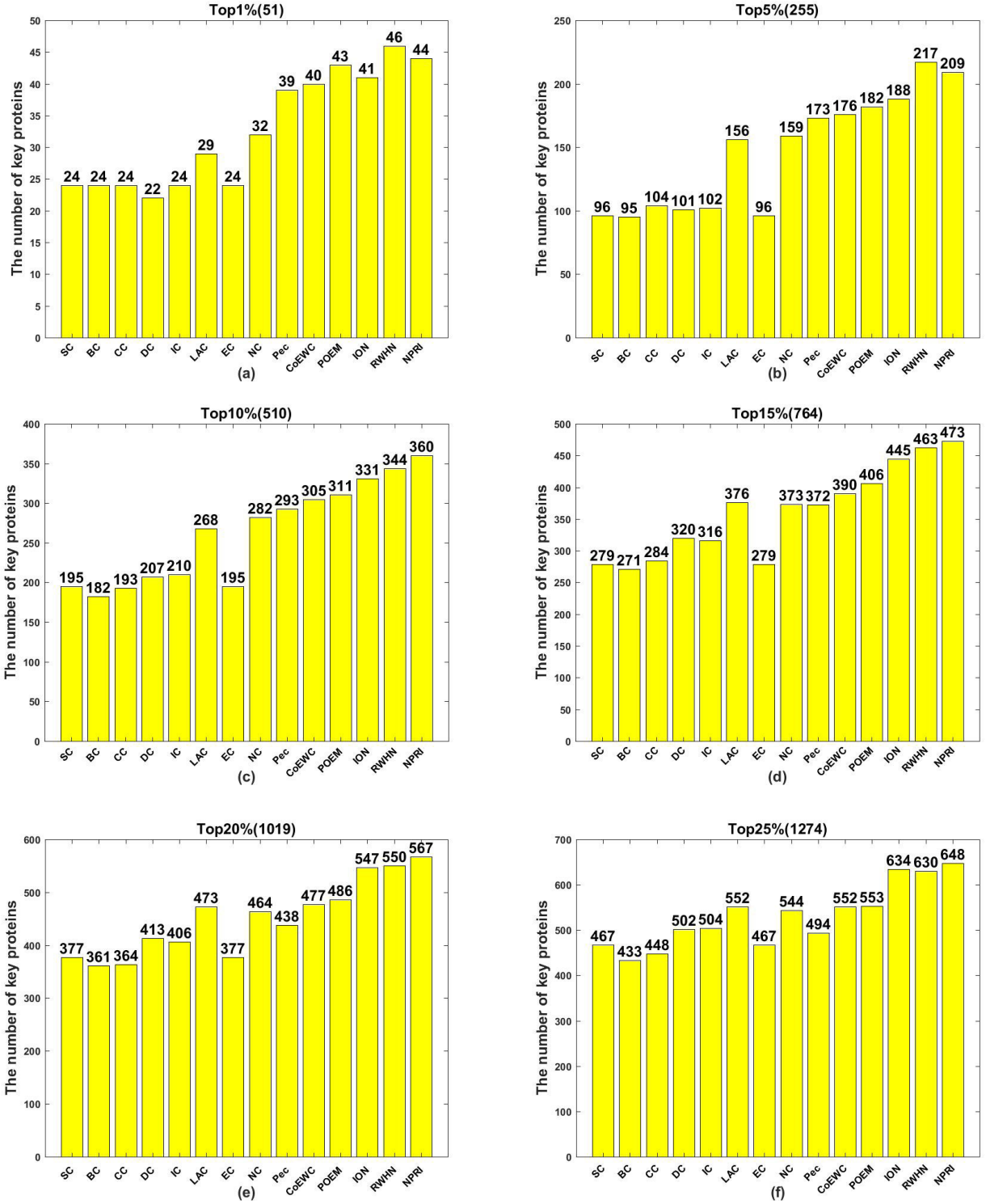


FIGURE 2. (*a*) Top 1% ranked proteins (*b*) Top 5% ranked proteins (*c*) Top 10% ranked proteins (*d*) Top 15% ranked proteins (*e*) Top 20% ranked proteins (*f*) Top 25% ranked proteins. This figure shows the comparison of the number of essential proteins identified by NPRI and 13 competitive prediction models. During simulation, proteins are ranked in descending order based on their ranking scores calculated by NPRI, CC, IC, SC, EC, BC, NC, DC, LAC, PEC, CoEWC, POEM, ION and RWHN respectively.And then, the top 1%, top 5%, top 10%, top 15%, top 20% and top 25% ranked proteins will be chosen as candidate essential proteins. Thereafter, through comparing with known essential proteins, the number of true essential proteins detected by each method will be used as the judgment criteria of prediction ability. This figure shows the number of true key proteins identified by each method.

LAC, PEC, CoEWC, POEM, ION and RWHN respectively. And simulation results are shown in the following Fig.2.

The receiver operating characteristic(ROC) curve was introduced to evaluate the performance of NPRI method.

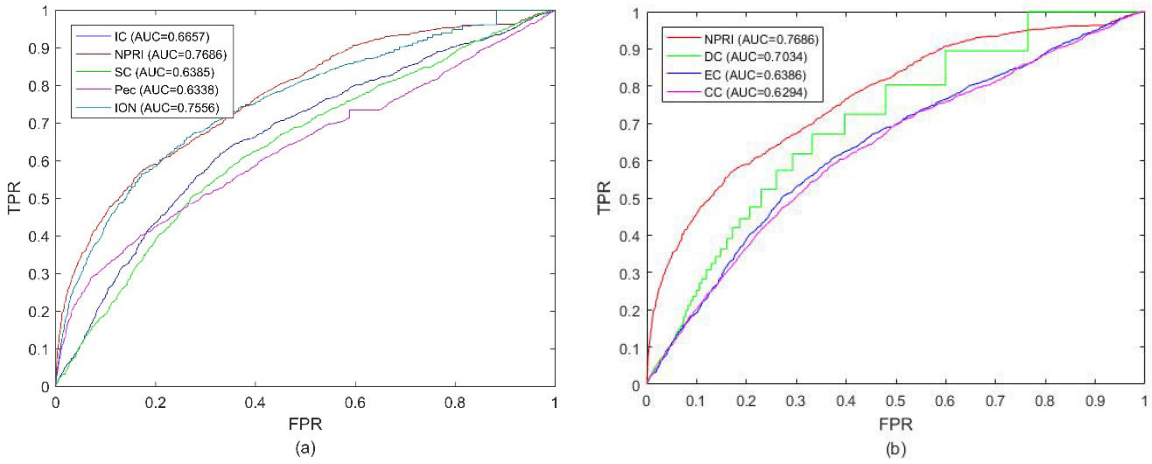


FIGURE 3. The ROC curves of IC NPRI,SC,Pec,ION (b) The ROC curves of NPRI,DC,EC,CC.

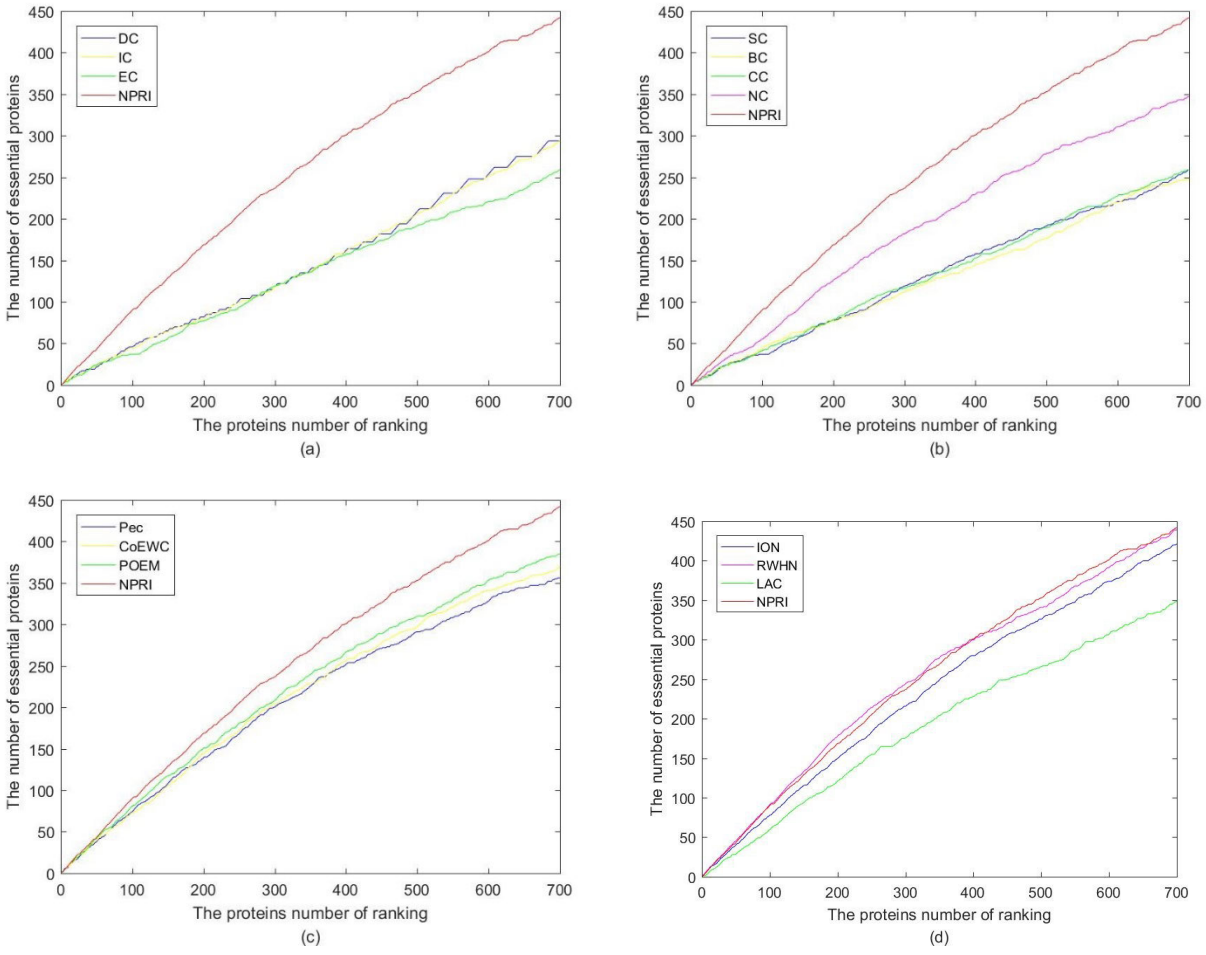


FIGURE 4. Results of comparisons between NPRI and 13 state-of-the-art competitive prediction models based on the top 700 ranked key proteins by implementing the Jackknife methodology on the DIP database. The X-axis of this figure denotes the number of ranked proteins, while the Y-axis represents the number of true key proteins identified by prediction models. (*a*) comparison between NPRI and DC, IC and EC. (*b*) comparison between NPRI and SC, BC, CC and NC. (*c*) comparison between NPRI and Pec, CoEWC and POEM. (*d*) comparison between NPRI and ION and RWHN.

The larger the area under the ROC curve(AUC),the better effect of method. If the AUC = 0.5, that indicates random performance. From the FIGURE 3 we can see that the AUC of NPRI is 0.7686, which is the best effect with the comparison.

TABLE 3. Commonalities and differences between NPRI and 13 competitive methods based on the top 500 ranked proteins and the DIP database.

Centrality measure (M_i)	$\|NPRI \cap M_i\|$	$\|NPRI - M_i\|$	Percentage of essential proteins in $\{NPRI - M_i\}$	$\|M_i - NPRI\|$	Percentage of essential proteins in $\{M_i - NPRI\}$
DC	174	326	69.63%	326	23.31%
IC	176	324	69.44%	324	24.38%
EC	141	359	69.63%	359	24.79%
SC	141	359	69.63%	359	24.79%
BC	138	362	69.61%	362	20.99%
CC	174	326	69.63%	326	24.79%
NC	241	259	64.09%	259	35.52%
Pec	222	278	60.07%	278	37.76%
CoEWC	236	264	58.71%	264	37.50%
POEM	256	244	57.37%	244	39.75%
ION	361	139	56.35%	139	41.43%
LAC	233	267	64.41%	267	31.83%
RWHN	361	139	53.95%	139	45.32%

Table 3: 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.

TABLE 4. Number of essential proteins predicted by NPRI and 13 methods based on the GAVIN database.

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
NPRI	16	75	153	221	278	323

D. VALIDATION BY JACKKNIFE METHOLOGY

Jackknife Methodology [40] is an effective method adopted to assess the advantage and disadvantage of models for identifying essential proteins. In order to evaluate NPRI more comprehensively and concretely, in this section, we will implement the Jackknife Methodology on top 700 candidate essential proteins predicted by NPRI and 13 state-of-the-art competitive prediction models to test their superiority and disadvantages, and the comparison results are shown in the following Fig.4.

If the AUC = 0.5, that indicates random performance. From the FIGURE 3 we can see that the AUC of NPRI is 0.7686, which is the best effect with the comparison.

From observing Fig.4(a), Fig.4(b) and Fig.4(c), it is obvious that the prediction performance of NPRI is significantly better than all these competitive methods. From observing Fig.4(d), it is easy to see that the prediction performance of NPRI is better than LAC and ION, meanwhile, the curves of NPRI and RWHN are intersected with each other. However, through careful observation, we will find that when the number of candidate key proteins increases to 400, the curve of RWHN will turn lower than that of NPRI. That is to say, with the increasing of predicted scale of proteins, the predictive performance of NPRI will gradually exceed that of RWHN. Hence, we can declare that the prediction performance of NPRI is better than that of these 13 representative methods on the whole.

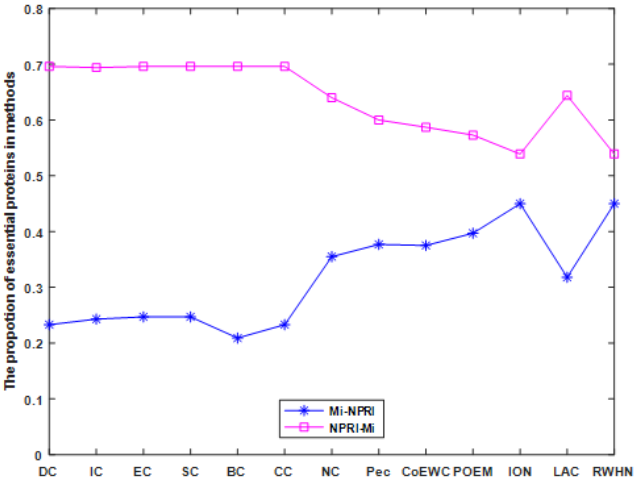


FIGURE 5. Comparison results of essential proteins detected by NPRI method and 13 state-of-the-art competitive methods.

E. DIFFERENCE BETWEEN NPRI AND 13 COMPETITIVE PREDICTION METHODS

In order to analyze the difference between NPRI and these 13 state-of-the-art prediction methods such as CC, IC, SC, EC, BC, NC, DC, LAC, PEC, CoEWC, POEM, ION and RWHN, in this section we further compare NPRI and these 13 state-of-the-art prediction methods based on the top 500 ranked proteins. Comparison results are illustrated in the following table.3 and Fig.5.

In Table.3 and Fig.5, the centrality measures (Mi) denotes one of these 13 competitive methods. |NPRI∩Mi| means the number of common essential proteins detected by both NPRI and Mi. |NPRI-Mi| represents the number of proteins identified by NPRI but not by Mi. |Mi-NPRI| indicates the number of proteins identified by Mi but not by NPRI. {NPRI-Mi} denotes the set of true essential proteins detected by NPRI but not by Mi. {Mi-NPRI} represents the set of true essential proteins detected by Mi but not by NPRI.

As can be seen from the {NPRI-Mi} or {Mi-NPRI}, based on top 700 proteins, the key proteins identified by NPRI method and other prediction method are of discrepancy. We can perceive that the essential proteins identified by Pec method only accounted for 37.78% of the{Mi-NPRI},but in NPRI method, the percentage of key proteins in {NPRI-Mi} is 60.08%.It can be observed from {NPRI - Mi} and {Mi - NPRI} that the essential proteins in {RWHN - NPRI} account for the highest proportion(45%) but the proportion of real key proteins in {NPRI- RWHN} was 53.91%,which was higher than {RWHN - NPRI}. Thus, from what has been analyze above, NPRI method is a special method that can identify more different true essential proteins and effectively eliminate noise data.

F. RECOGNITION PERFORMANCE OF NPRI BASED ON THE GAVIN DATABASE

In order to verify the universal applicability of NPRI, in this section, we adopt the Gavin database to compare the prediction effects between NPRI with 13 representative prediction methods. The comparison results are shown in the following Table.4 and Fig.6. From observing the Table.4, it is easy to see that although the prediction performance of NPRI in top 1% is slightly less than POEN, ION and RWHN, but from top 5% to top 25%, the prediction performances of NPRI are all better than all these 13 competitive methods.

Moreover, from observing the Fig.6(a) and Fig.6(b), it is obvious that the prediction performances of NPRI is better than 8 competitive methods including SC, EC, BC, DC, IC, CC, NC and Pec simultaneously. From observing the Fig.6(c), we can find that although the curve of RWHN at range (100,150) is a little higher than NPRI, but as a whole, the prediction performance of NPRI is much better than RWHN. Therefore, according to the comparison results based on both the DIP database and the Gavin database, we can declare that NPRI is a satisfactory method for predicting potential essential proteins with high-accuracy, high-efficiency and high-practicability.

IV. DISCUSSION

Essential proteins perform a vitally important role in human life. Currently, an increasing number researches aim to predict key proteins by computational models, since it leads to high cost of both money and time to predict essential proteins by using biological experiments. However, it is still an important and challenging work to design stable and accurate prediction algorithms. In recent years, more and more biological data related to proteins have been introduced to identify key proteins based on PPI networks. Inspired by them, in this manuscript, a new prediction model called NPRI is proposed through combining the topological features and relevant biological features of proteins to infer potential essential proteins. Simulation results show that NPRI not only has stability but also can improve the prediction precision quite effectively.

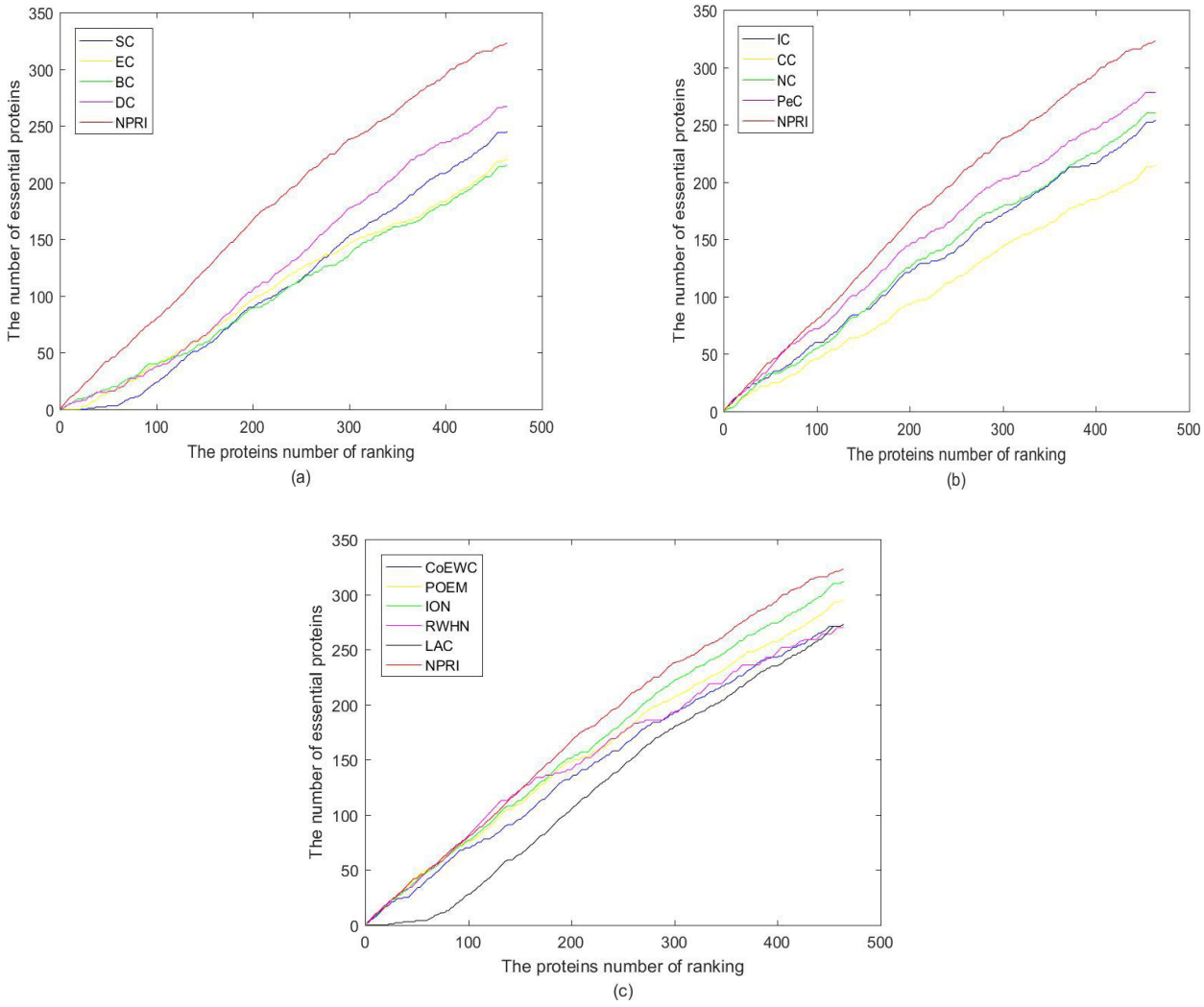


FIGURE 6. Comparison results between NPRI and 13 representative competitive methods.

V. CONCLUSION

In this paper, a novel prediction model called NPRI is proposed to infer potential key proteins by integrating functional features and topological features of proteins. In NPRI, a new heterogeneous Protein-Domain network is established first through combining a weighted PPI network, an initial Protein-Domain network and a weighted Domain-Domain network. And then, based on the newly constructed heterogeneous Protein-Domain network, functional features and topological features will be extracted for each protein, based on which, initial scores can be obtained for each protein and domain. Finally, an improved PageRank algorithm will be implemented on the heterogeneous Protein-Domain network to detect potential essential proteins. Experimental results demonstrate that the identification performance of NPRI is superior to state-of-the-art competitive prediction methods.

ACKNOWLEDGMENT

(Zhiping Chen, Zixuan Meng, and Xiangyi Wang 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, Jun. 2019, Art. no. 355, doi: 10.1186/s12859-019-2930-2.

[2] M. L. Acencio and N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information," *BMC Bioinf.*, vol. 10, pp. 290–307, Dec. 2009.

[3] J. Wang, W. Peng, and F.-X. Wu, "Computational approaches to predicting essential proteins: A survey," *Proteomices, Clin. Appl.*, vol. 7, nos. 1–2, pp. 181–192, Jan. 2013.

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

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

[6] 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.

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

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

[9] 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.

[10] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Phys. Rev. E, Covering Stat., Nonlinear, Biol., Soft Matter Phys.*, vol. 71, no. 5, pp. 122–133,2005.

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

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

[13] 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. BiologyBMC Syst Biol.*, vol. 6, no. 1, p. 15, 2012.

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

[15] W. Zhang, J. Xu, Y. Li, and X. Zou, "Detecting essential proteins based on network topology, gene expression data, and gene ontology information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 109–116, Jan. 2018.

[16] X. Lei, J. Zhao, H. Fujita, and A. Zhang, "Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets," *Knowl.-Based Syst.*, vol. 151, pp. 136–148, Jul. 2018.

[17] J. Zhong, J. Wang, W. Peng, Z. Zhang, and Y. Pan, "Prediction of essential proteins based on gene expression programming," *BMC Genomics*, vol. 14, no. S4, p. S7, 2013.

[18] Y. Fan, X. Hu, X. Tang, Q. Ping, and W. Wu, "A novel algorithm for identifying essential proteins by integrating subcellular localization," presented at the IEEE Int. Conf. Bioinf. Biomed., Dec. 2017, pp. 107-111.

[19] C. F. Shabnam and S. Izudheen, "UDoGeC: Essential protein prediction using domain and gene expression profiles," *Procedia Comput. Sci.*, vol. 93, pp. 1003–1009, 2016.

[20] X. Lei, S. Wang, and L. Pan, "Predicting essential proteins based on gene expression data, subcellular localization and PPI data," presented at the Bio-Inspired Comput., Theories Appl., 2017, pp. 92–105.

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

[22] J. Luo and L. Ma, "A new integration-centric algorithm of identifying essential proteins based on topology structure of protein-protein interaction network and complex information," *CBIO*, vol. 8, no. 3, pp. 380–385, May 2013.

[23] 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.

[24] J. Ren, J. Wang, M. Li, H. Wang, and B. Liu, "Prediction of essential proteins by integration of PPI network topology and protein complexes information," presented at the 7th Int. Symp. Bioinf. Res. Appl. Changsha, China: Central South Univ., May 2011, pp. 12–24.

[25] D. Mistry, R. P. Wise, and J. A. Dickerson, "DiffSLC: A graph centrality method to detect essential proteins of a protein-protein interaction network," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0187091.

[26] C. Qin, Y. Sun, and Y. Dong, "A new computational strategy for identifying essential proteins based on network topological properties and biological information," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0182031.

[27] 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, pp. 1–17, 2012.

[28] 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.

[29] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, 2017.

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

[31] 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.

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

[33] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, and D. J. Studholme, "The Pfam protein families database," *Nucleic Acids Res.* vol. 32, no. 1, pp. D138–D141, 2004.

[34] J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, and L. J. Jensen, "COMPARTMENTS: Unification and visualization of protein subcellular localization evidence," *Database*, vol. 2014, Feb. 2014, Art. no. bau012.

[35] 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, pp. D196–D203, Jan. 2010.

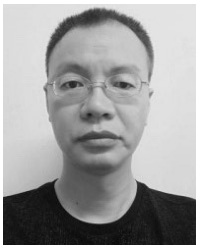
[36] H. W. Mewes, "MIPS: Analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D169–D172, Jan. 2006.

[37] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, and S. Weng, "SGD: *Saccharomyces* genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, 1998.

[38] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Res.*, vol. 37, pp. D455–D458, Jan. 2009.

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

[40] 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 MicrobiologyBMC Microbiol*, vol. 9, no. 1, p. 243, 2009.



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 in Hunan University. He is currently a Professor with Changsha University. His current research area is mainly bioinformatics.

ZIXUAN MENG is currently pursuing the master's degree in computer science and technology with the College of Information and Engineering, Xiangtan University. Her current research interest is bioinformatics.

CHAOPING LIU is currently pursuing the bachelor's degree in Internet of Things with Changsha University. Her current research interest is bioinformatics.

XIANGYI WANG is currently pursuing the bachelor's degree with Changsha University.

LINAI KUANG received the B.S. and M.S. degrees in computer software from Xiangtan University, in 2003, and the Ph.D. degree in technology of computer application from Central South University, in 2011.

He is currently a Professor with the College of Information Engineering with Xiangtan University. He is the author of three books. His main research interests include bioinformatics and wireless sensor networks.

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 the author of 13 invention patents and more than 30 articles. His main research areas are the Internet of Things, wireless sensor networks (WSN), mobile ad-hoc networks, mobile communication networks, and social computing.

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. 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, Changsha University, China. He has published more than 100 peer-reviewed articles. His main research areas include bioinformatics and the Internet of Things.