frontiers in Genetics

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/fgene.2021.708162

Method for Identifying Essential Proteins by Key Features of Proteins in a Novel Protein-Domain Network

Xin He[1], Linai Kuang[1], Zhiping Chen[2], Yihong Tan[2] and Lei Wang[1,2]**

[1] College of Computer, Xiangtan University, Xiangtan, China, [2] College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, China

In recent years, due to low accuracy and high costs of traditional biological experiments, more and more computational models have been proposed successively to infer potential essential proteins. In this paper, a novel prediction method called KFPM is proposed, in which, a novel protein-domain heterogeneous network is established first by combining known protein-protein interactions with known associations between proteins and domains. Next, based on key topological characteristics extracted from the newly constructed protein-domain network and functional characteristics extracted from multiple biological information of proteins, a new computational method is designed to effectively integrate multiple biological features to infer potential essential proteins based on an improved PageRank algorithm. Finally, in order to evaluate the performance of KFPM, we compared it with 13 state-of-the-art prediction methods, experimental results show that, among the top 1, 5, and 10% of candidate proteins predicted by KFPM, the prediction accuracy can achieve 96.08, 83.14, and 70.59%, respectively, which significantly outperform all these 13 competitive methods. It means that KFPM may be a meaningful tool for prediction of potential essential proteins in the future.

Keywords: essential proteins, protein-protein network, computational model, domain-domain network, protein-domain network

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Liao Bo,
Hainan Normal University, China
Fan Yang,
Jiangxi Science and Technology Normal University, China

Correspondence:
Linai Kuang
kla@xtu.edu.cn
Lei Wang
wanglei@xtu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *11 May 2021*
Accepted: *31 May 2021*
Published: *29 June 2021*

Citation:
He X, Kuang L, Chen Z, Tan Y and Wang L (2021) Method for Identifying Essential Proteins by Key Features of Proteins in a Novel Protein-Domain Network. Front. Genet. 12:708162. doi: 10.3389/fgene.2021.708162

INTRODUCTION

Essential proteins are indispensable proteins in the reproduction and survival of organisms, and experimental results have shown that removal of essential proteins may lead to inability of organisms to survive and develop (Zhang Z. et al., 2020; Zhao et al., 2020; Meng et al., 2021). In recent years, with the rapid development of high-throughput technologies, more and more interactions between proteins have been found in *Saccharomyces cerevisiae*, and it has become a hot spot of research to identify essential proteins from large amount of known protein-protein interaction (PPI) data by adopting computational methods. Up to now, a lot of computational prediction methods have been proposed successively to infer potential essential proteins, and in general, these methods can be roughly divided into two categories. The first category of methods mainly relies on topological characteristics of PPI networks to predict essential proteins. For instance, based on the centrality-lethality rule (Jeong et al., 2001) that proteins with high degree of interconnectivity are more likely to be essential proteins than those with low degree of interconnectivity in a PPI network, a series of centrality-based methods including DC (Hahn and Kern, 2004), CC (Wuchty and Stadler, 2003), BC (Joy et al., 2005), EC (Bonacich, 1987), SC

(Estrada and Rodriguez-Velazquez, 2005), and IC (Stephenson and Zelen, 1989) have been designed to identify key proteins by the interconnectivities of proteins in PPI networks, and among them, the SC method was proven to be the best (Estrada, 2010). Except for these centrality-based methods, Wang et al. (2012) presented a method named NC for detecting essential proteins based on the edge aggregation coefficients. Li et al. (2011) proposed a method called LAC to predict essential proteins by evaluating the relationship between proteins and their neighbors in the PPI network. Wang et al. (2011) put forward a model called SoECC by the correlation between PPI network proteins. Przulj et al. (2004) designed a prediction model by constructing the shortest path spanning tree for each protein in the PPI network. In the first category of methods, some topological structures of PPI networks such as the node degree of interconnectivities and common neighboring nodes have been adopted to infer key proteins, however, due to the incompleteness of PPI networks, these methods cannot achieve satisfactory prediction accuracy.

In order to overcome the limitations of the first category of methods, the second category of methods focus on predicting essential proteins by combining topological features of PPI networks and functional features of proteins extracted from the gene expression data, orthology information and the subcellular localization of proteins. For example, Lei et al. (2018c) combined topological features of PPI networks and the GO data of proteins to design a novel method called RSG for predicting essential proteins. Zhang et al. (2013) designed a method called COEWC by integrating neighborhood features of the PPI network with the gene expression data of proteins to infer key proteins. Li et al. (2012) developed a prediction model named Pec by combining the PPI network and the gene expression data of proteins. Tang et al. (2014) proposed a novel method named WDC based on the edge clustering coefficients and the Pearson correlation coefficients of proteins. Peng et al. (2012) developed a computational model called ION by integrating the protein orthology information with PPI data to predict essential proteins. Xiao et al. (2013) developed a method for predicting essential proteins by combining the PPI network with the co-expressed gene data of proteins. Ren et al. (2011) invented a method to identify key proteins by integrating PPI networks with the protein complex information. Jiang et al. (2015) integrated topological features of PPI networks with the gene expression data of proteins to design a prediction model called IEW for key protein prediction. Zhao et al. (2014) developed a computational method named POEM by combining the gene expression data of proteins with topological attributes of PPI networks. Zhong et al. (2020) developed a predictive model called JDC by combining topological characteristics of PPI networks and gene expression data of proteins. Keretsu and Sarmah (2016) used the marginal clustering coefficients and the gene expression correlation between interacting proteins to design a method for identifying protein complexes. Zz et al. (2019) proposed a method that refines PPI networks by using gene expression information and subcellular localization information. Ahmed et al. (2021) designed a predictive model called EPD-RW through incorporating PPI networks with four kinds of biological data of proteins including GO data, gene expression profiles, domain information and phylogenetic profile to infer essential proteins. Zhang et al. (2021) proposed an identification model by combining PPI networks with the gene expression profile, GO information, subcellular localization information, and orthology data of proteins to detect essential proteins. Lei et al. (2018a) combined the gene expression data, subcellular location and protein complex information of proteins with the topological characteristics of PPI networks to develop a key protein identification algorithm FPE. Zhao et al. (2019) designed an iterative method called RWHN by integrating the PPI network with domains, subcellular location and homology information of proteins to identify essential proteins. Lei et al. (2018b) designed a novel calculation model named AFSO_EP to identify essential proteins by combining PPI networks with the gene expression, GO annotation and subcellular location information of proteins. Zhang et al. (2019) proposed a predictive model called TEGS by combining multiple functional features including the subcellular location data and gene expression data of proteins with topological features of PPI networks. Li et al. (2020) put forward a prediction model named CVIM by combing gene expressions data and orthologous information of proteins with PPI networks to infer essential proteins.

Experimental results have demonstrated that the second category of methods can achieve better prediction performance than the first category of methods by integrating biological characteristics of proteins and topological characteristics of PPI networks, and it is useful to adopt the biological characteristics of proteins to compensate for the incompleteness of the PPI data. Hence, in order to further improve the accuracy of prediction models, in this paper, we extracted some new topological features from a newly constructed protein-domain network and some new functional features of proteins from the domain data, gene expression data, and orthologous information of proteins etc., based on which, a novel identification model called KFPM was proposed to infer potential essential proteins. Different from existing models, in KFPM, the gene expression data of protein will be processed first by adopting the Pearson Correlation Coefficient (PCC) (Horyu and Hayashi, 2013), and then, an improved Criteria Importance Though Intercrieria Correlation algorithm (CRITIC) (Zhang B. et al., 2020) will be applied to effectively combine multiple biological features of proteins by the contrast strength of features and the conflicts between features, based on which, a novel distribution rate network is constructed and an improved PageRank algorithm will be designed to identify potential essential proteins. Finally, we compared KFPM with 13 advanced methods including DC (Hahn and Kern, 2004), CC (Wuchty and Stadler, 2003), BC (Joy et al., 2005), EC (Bonacich, 1987), SC (Estrada and Rodriguez-Velazquez, 2005), IC (Stephenson and Zelen, 1989), NC (Wang et al., 2012), CoEWC (Zhang et al., 2013), Pec (Li et al., 2012), ION (Peng et al., 2012), POEM (Zhao et al., 2014), TEGS (Zhang et al., 2019), and CVIM (Li et al., 2020). And experimental results showed that KFPM outperformed all these competitive state-of-the-art predictive methods as a whole.

MATERIALS AND METHODS

Experimental Data

In this section, In order to evaluate the prediction accuracy of KFPM, known protein-protein interactions (PPI) would be downloaded first from the saccharomyces cerevisiae related public databases including DIP database (Xenarios et al., 2002), the Krogan database (Krogan et al., 2006), and the Gavin database (Gavin et al., 2006), respectively. As illustrated in **Table 1**, after filtering out repetitive interactions, we finally obtained 5,093 proteins and 24,743 interactions from the DIP database, 3,672 proteins and 14,317 interactions from the Krogan database, and 1,855 proteins and 7,669 interactions from the Gavin database. Next, we downloaded 1,107 domains from the Pfam (Bateman et al., 2004) database as well. Therefore, we constructed a (5,093+1,107) × (5,093+1,107), a (3,672+1,107) × (3,672+1,107) and a (1,855+1,107) × (1,855+1,107) dimensional networks by combining the datasets downloaded from the DIP, the Krogan and the Gavin databases with the dataset downloaded from the Pfam database separately. Moreover, we downloaded the gene expression data for calculating the initial protein scores from the Tu-BP database (Tu et al., 2005). Gene expression data contains 6,776 lines with length of 36, and each line represents the corresponding expression data of a different gene. Through comparison, we found that in datasets downloaded from the DIP and the Gavin databases, the number of proteins containing the gene expression data is more than 95%. Additionally, we downloaded orthologous information of proteins from the InParanoid database (Gabriel et al., 2010) and subcellular localization data of proteins from the COMPART-MENTS databases (Binder et al., 2014) to calculate initial scores for proteins, and as a result, we derived eleven subcellular locations such as the Mitochondrion, Peroxisome, Plasma, Extracellular, Endosome, Vacuole, Endoplasmic, Cytosol, Golgi, and Cytoskeleton Nucleus, that are related to essential proteins. Finally, a benchmark dataset for testing different prediction models was downloaded from the following four databases such as MIPS (Mewes et al., 2006), SGD (Cherry et al., 1998), DEG (Zhang and Lin, 2009), and SGDP (Saccharomyces Genome Deletion Project, 2012), which contains 1,293 key proteins. In this paper, we would provide comparison results based on datasets downloaded from the DIP and the Krogan databases in detail, and introduce briefly the experimental results based on the dataset downloaded from the Gavin database instead.

As shown in **Figure 1**, the flowchart of KFPM consists of the following four major steps:

Step 1: Based on known PPI dataset downloaded from a given public database, an original PPI network will be constructed first. And then, based on key topological characteristics of the original PPI network, weights between protein nodes will be calculated and adopted to transform the original PPI network to a weighted PPI network.

Step 2: Next, based on known relationships between proteins and domains, a weighted domain-domain network and an original protein-domain network will be constructed sequentially. And then, a novel heterogeneous protein-domain network will be established by integrating these three newly constructed networks such as the weighted PPI network, the weighted domain-domain network and the original protein-domain network.

Step 3: Moreover, an improved CRITIC algorithm will be applied to effectively integrate multiple biological features of proteins with key topological features extracted from the heterogeneous protein-domain network to calculate initial scores for proteins and domains.

Step 4: Finally, a novel transition probability matrix will be obtained, and then, through combining initial scores of proteins and domains with the transition probability matrix, a new iterative algorithm will be designed to identify potential essential proteins based on the PageRank algorithm.

Construction of the Weighted PPI Network

For convenience, let $P = \{p_1, p_2, \cdots, p_N,\}$ denote the set of different proteins downloaded from a given public database, and for a pair of proteins p_u and p_v in P, if there is a known interaction between them, we define that there is an edge $e(p_u, p_v) = 1$. Hence, let E represent the set of edges between proteins in P, Then it is obvious that we can obtain an original PPI network $PPIN = (P, E)$.

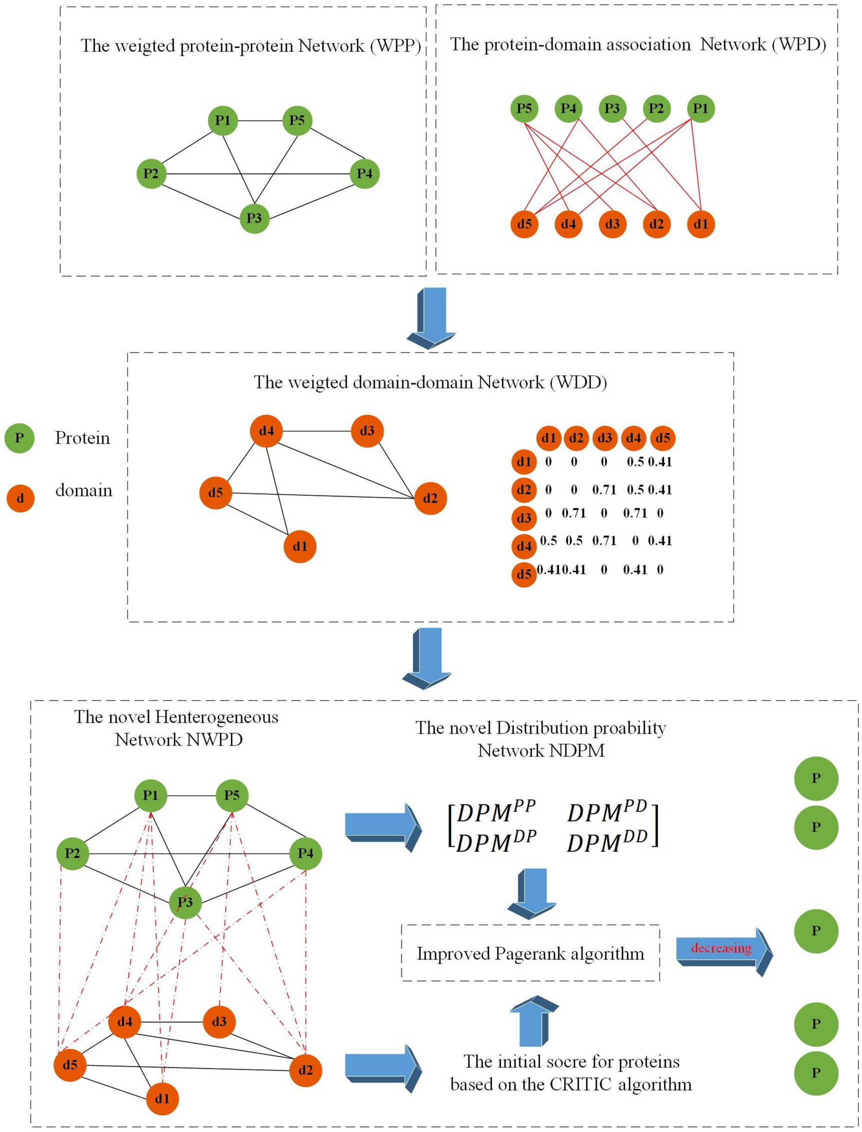
Additionally, inspired by the assumption that degrees of connections between essential proteins are mostly higher than degrees of connections between non-essential proteins (Zhang et al., 2016), for any two given protein nodes p_u and p_v in $PPIN$, it is obvious that we can estimate the degree of relationship between them according to the following equation (1):

$$WPP\left(p_u, p_v\right) = \begin{cases} \dfrac{\left|NG\left(p_u\right) \cap NG\left(p_v\right)\right|^2}{\left(\left|NG\left(p_u\right)\right| + 1\right) \times \left(\left|NG\left(p_v\right)\right| + 1\right)} & if\ e\left(p_u, p_v\right) = 1 \\ 0 & otherwise \end{cases} \tag{1}$$

Here, $NG\left(p_u\right)$ represents the set of neighboring nodes of p_u in $PPIN$, $\left|NG\left(p_u\right)\right|$ denotes the total number of neighboring nodes of p_u in $PPIN$, and $NG\left(p_u\right) \cap NG\left(p_v\right)$ means the set of common neighboring nodes of both p_u and p_v in $PPIN$. Obviously, according to above Eq. 1, it is easy to obtain a $N \times N$ dimensional adjacency matrix WPP, based on which, we can obtain a weighted PPI network easily as well.

TABLE 1 | The information of the DIP, Krogan and Gavin database.

Database	Proteins	Interactions	Essential proteins
DIP	5,093	24,743	1,167
Krogan	3,672	14,317	929
Gavin	1,855	7,669	714


The weigted protein-protein Network (WPP)
P1
P5
P2
P4
P3
The protein-domain association Network (WPD)
P5
P4
P3
P2
P1
d5
d4
d3
d2
d1
P
Protein
d
domain
The weigted domain-domain Network (WDD)
d4
d3
d5
d2
d1
The novel Henterogeneous Network NWPD
The novel Distribution proability Network NDPM
$\begin{bmatrix} DPM^{PP} & DPM^{PD} \\ DPM^{DP} & DPM^{DD} \end{bmatrix}$
Improved Pagerank algorithm
decreasing
The initial socre for proteins based on the CRITIC algorithm
P

	d1	d2	d3	d4	d5
d1	0	0	0	0.5	0.41
d2	0	0	0.71	0.5	0.41
d3	0	0.71	0	0.71	0
d4	0.5	0.5	0.71	0	0.41
d5	0.41	0.41	0	0.41	0

FIGURE 1 | Flowchart of KFPM.

Construction of the Heterogeneous Protein-Domain Network

In this section, we will download the domain set of proteins $D = \{d_1, d_2, \cdots, d_M\}$ from the Pfam database (Bateman et al., 2004), based on which, an initial protein-domain interaction network will be constructed as follows: for any given $p_u \in P$ and domain $d_v \in D$, if there is $p_u \in d_v$, we define that there is an edge existing between them. Thereafter, it is easy to see that we can obtain an initial protein-domain interaction network and a $N \times M$ dimensional adjacency matrix WPD as follows: for any given $p_u \in P$ and domain $d_v \in D$, if there is an edge between them, then there is $WPD(p_u, d_v) = 1$, otherwise there is $WPD(p_u, d_v) = 0$.

Moreover, for any two given domains d_u and d_v, let $N(d_u)$ and $N(d_v)$ represent the number of proteins belonging to d_u and d_v separately, $N(d_u) \cap N(d_v)$ denote the number of proteins belonging to both d_u and d_v simultaneously, then we can calculate the weight between d_u and d_v according to the following Eq. 2:

$$WDD(d_u, d_v) = \begin{cases} \frac{N(d_u) \cap N(d_v)}{\sqrt{|N(d_u)| \times |N(d_v)|}} & if\ |N(d_u)| > 0\ and\ |N(d_v)| > 0 \\ 0 & otherwise \end{cases} \quad (2)$$

Based on above Eq. 2, it is obvious that we can further obtain a $M \times M$ dimensional adjacency matrix WDD. And then, through combining above obtained $N \times N$ dimensional adjacency matrix WPP, $N \times M$ dimensional adjacency matrix WPD and $M \times M$ dimensional adjacency matrix WDD, wen can obtain a new $(M+N) \times (M+N)$ dimensional adjacency matrix *NWPD* as follows:

$$NWPD = \begin{bmatrix} WPP & WPD \\ WPD^T & WDD \end{bmatrix} \tag{3}$$

Calculation of Initial Scores for Proteins and Domains

In order to reduce the negative impact of false positives, in this section, we will adopt topological and functional characteristics of proteins to calculate initial scores for proteins. For any given protein p_u, let $I\left(p_u\right)$ denote the orthologous information of p_u, then we can obtain the orthologous score $BIO\_I\left(p_u\right)$ of p_u as follows:

$$BIO\_I\left(p_u\right) = \frac{I\left(p_u\right)}{max_{\,p_v \in P}\left(I\left(p_v\right)\right)} \tag{4}$$

Moreover, considering that gene expression refers to the process of synthesizing protein under the guidance of genes, and Pearson correlation coefficient (PCC) is suitable for measuring the degree of linear correlation between two vectors, hence, for any two given proteins p_u and p_v, it is obvious that we can implement PCC on gene expressions of these two proteins to calculate the similarity between them as follows:

$$PCC\left(p_u, p_v\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Exp\left(p_u, i\right) - \overline{Exp\left(p_u\right)}}{\sigma\left(p_u\right)}\right)\left(\frac{Exp\left(p_v, i\right) - \overline{Exp\left(p_v\right)}}{\sigma\left(p_v\right)}\right) \tag{5}$$

Here, $Exp\left(p_u, i\right)$represents the expression level of p_u at the i^{th} time node. $\overline{Exp(p_u)}$ denotes the average gene expression value of p_u, and $\sigma(p_u)$ is the standard deviation of gene expressions of p_u. Therefore, we can obtain a gene expression based functional characteristic of p_u as follows:

$$BIO\_Exp\left(p_u\right) = \sum_{p_v \epsilon NG(p_u)} PCC\left(p_u,\ p_v\right) \tag{6}$$

Next, based on subcellular localizations of proteins, for any given protein p_u, let $Sub_{(p_u)}$ represent the set of subcellular localizations associated with p_u, we can as well obtain a subcellular localization based functional characteristic of p_u as follows:

$$BIO\_sub\left(p_u\right) = \sum_{i \epsilon Sub_{(p_u)}} Eve_{sub}(i) \tag{7}$$

Where,

$$Eve_{sub}(i) = \frac{N_{sub}\left(i\right)}{Ave_{sub}} \tag{8}$$

$$Ave_{sub} = \frac{\sum_{i=1}^{N_{sub}} N_{sub}\left(i\right)}{N_{sub}} \tag{9}$$

Here, N_{sub} means the number of all subcellular localizations of proteins and $N_{sub}\left(i\right)$ represents the number of proteins associated with the i^{th} subcellular localization.

In KFPM, We apply an improved CRITIC method, which can be used to measure weights of different characteristics based on the contrast strengths of characteristics and the conflicts between characteristics, to integrate three kinds of biological characteristics obtained above to calculate final biological feature scores for proteins as follows:

First, let C_j denote the amount of information contained in the j^{th} biological feature of protein, where C_j can be expressed as follows:

$$C_j = \sigma_j \sum_{i=1}^{n}\left(1 - \left|r_{ij}\right|\right) \tag{10}$$

Here, r_{ij} represents the correlation coefficient between biological characteristics i and j. σ_j represents the standard deviation of the j^{th} biological feature. Obviously, the greater the value of C_j, the greater the amount of information contained in the j^{th} biological feature. Therefore, the objective weight w_j of the j^{th} biological feature can be defined as follows:

$$W_j = \frac{C_j}{\sum_{j=1}^{n} C_j} \tag{11}$$

Hence, based on three kinds of biological characteristics obtained above, the final biological feature score of protein p_u can be calculated as follows:

$$BIO\left(p_u\right) = \sum_{j=1}^{n} w_j BF\left(p_u\right) \tag{12}$$

$$\text{Where } BF\left(p_u\right) = \left(BIO\_I\left(p_u\right), BIO\_Exp\left(p_u\right), BIO\_sub\left(p_u\right)\right) \tag{13}$$

Based on above formula (12), we have obtained biological feature scores for proteins, next, for any given protein p_u, we will further calculate its topological feature score based on the topological structure of the newly constructed heterogeneous protein-domain network as follows :

$$TOP\left(p_u\right) = \frac{\sum_{v \epsilon NG(p_u)}\left|NG\left(p_u\right) \cap NG\left(p_v\right)\right|}{\left|NG\left(p_u\right)\right|} \tag{14}$$

Where $\left|NG\left(p_u\right) \cap NG\left(p_v\right)\right|$ denotes the number of elements in the set of $NG\left(p_u\right) \cap NG\left(p_v\right)$ and $\left|NG\left(p_u\right)\right|$denotes the number of nodes in $NG\left(p_u\right)$.

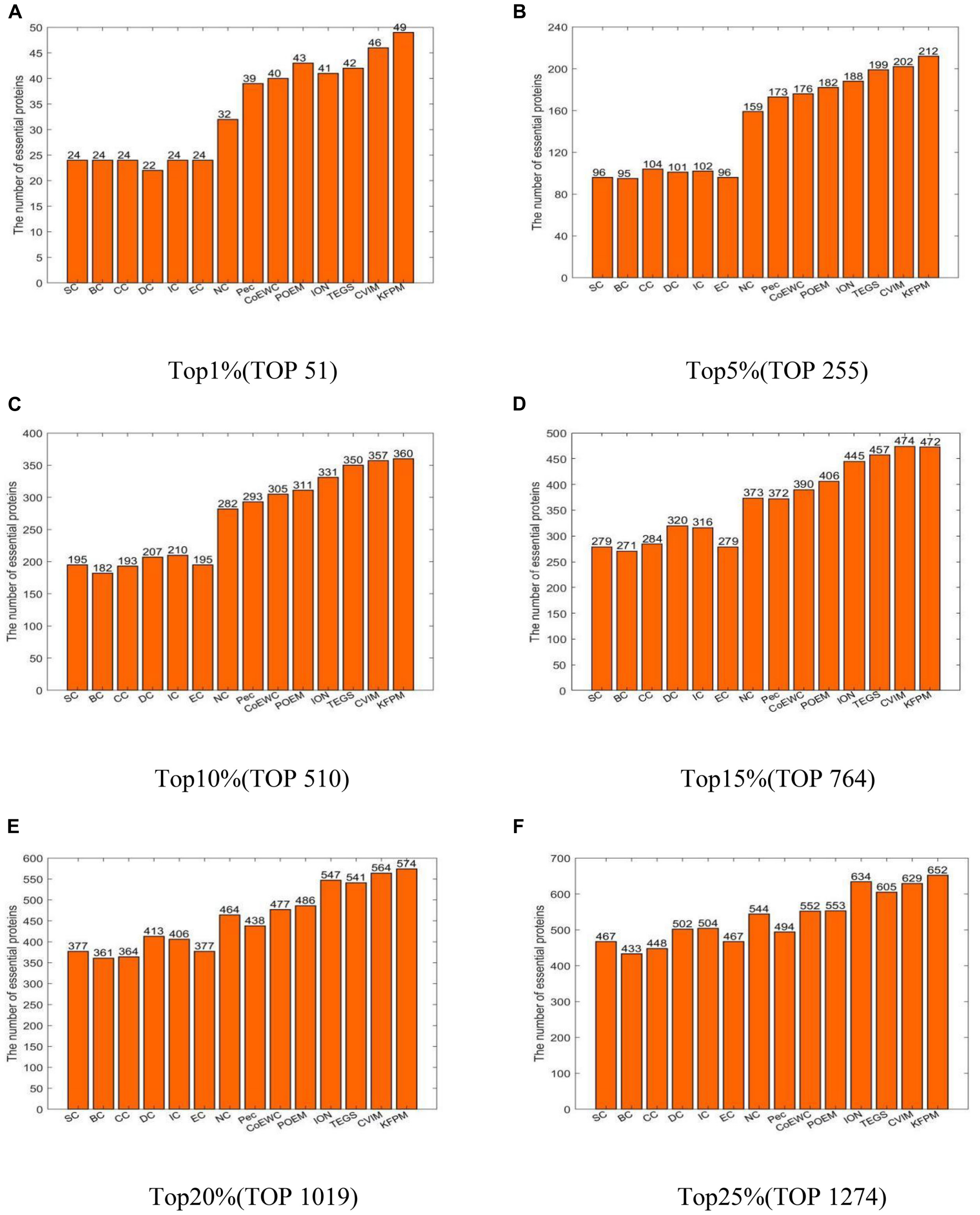
Therefore, through combining the topological feature and biological feature of p_u, we can define an unique final score for p_u as follows:

$$S_0\left(p_u\right) = BIO\left(p_u\right) \times \theta + (1-\theta) \times TOP\left(p_u\right) \tag{15}$$

Here, $\theta \in (0, 1)$ is a parameter of weight factor.

Additionally, in a similar way, for any given domain d_u, we can as well calculate an initial topological feature score for it as follows:

$$S\left(d_u\right) = \sum_{p \in d_u} S_0\left(p\right) \tag{16}$$



Top1%(TOP 51)

Top5%(TOP 255)

Top10%(TOP 510)

Top15%(TOP 764)

Top20%(TOP 1019)

Top25%(TOP 1274)

FIGURE 2 | Performance comparison between KFPM and 13 competitive methods based on the DIP database **(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.

Since the numbers of proteins in different domains are quite different, which lead to big difference between scores of domains obtained by above formula (16), therefore, after normalization, we can obtain the final topological feature score of d_u as follows:

$$S_0\left(d_u\right)=\frac{S\left(d_u\right)}{max_{1\leq j\leq N}S\left(d_j\right)} \tag{17}$$

Design of KFPM

First, for any two given proteins p_u and p_v in the heterogeneous protein-domain network, let $PPN\left(p_u,p_v\right)=\frac{WPP\left(p_u,p_v\right)}{\left(1+max\left(WPP\left(p_u,p_v\right)\right)\right)^2}$, it is obvious that *we* can obtain a distribution probability of p_u to p_v as follows:

$$DPM^{PP}\left(p_u,p_v\right)=\begin{cases}\frac{PPN\left(p_u,p_v\right)}{\sum_j PPN\left(p_u,p_j\right)}\times S_0\left(p_v\right), if\ PPN\left(p_u,p_v\right)\neq 0\\ 0 \qquad otherwise\end{cases} \tag{18}$$

Next, for any given protein p_u and domain d_v, let $PDN\left(p_u,d_v\right)=\frac{WPD\left(p_u,d_v\right)}{\left(1+\max\left(WPD\left(p_u,d_v\right)\right)\right)^2}$, it is obvious that

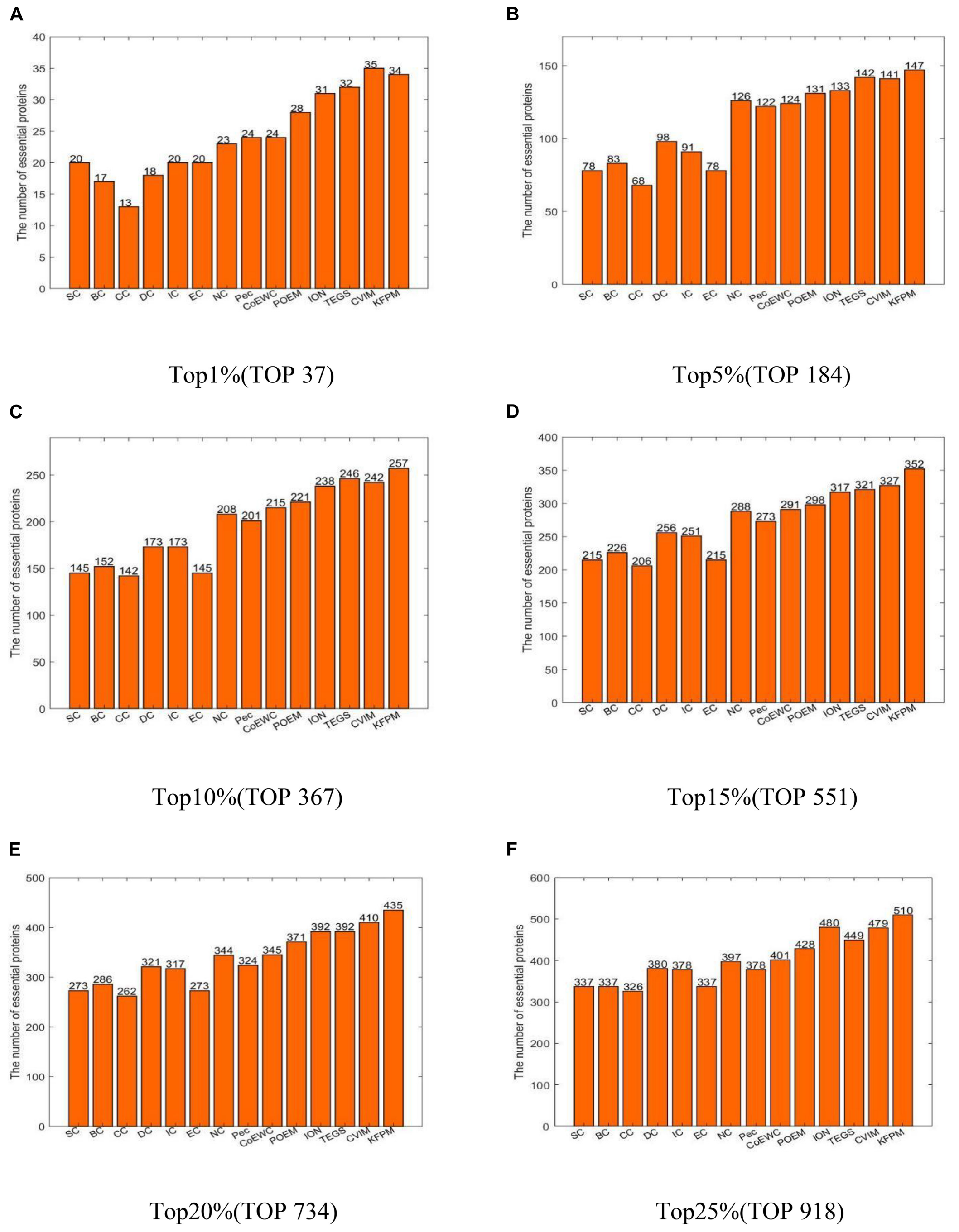


FIGURE 3 | Performance comparison between KFPM and 13 competitive methods based on the Krogan database **(A)** Top 1% ranked proteins. **(B)** Top 5% ranked proteins. **(C)** Top 10% ranked proteins. **(D)** Top 15% ranked proteins. **(E)** Top 20% ranked proteins. **(F)** Top 25% ranked proteins.

we can obtain a distribution probability of p_u to d_v as follows:

$$DPM^{PD}\left(p_u, d_v\right) = \begin{cases} \frac{PDN\left(p_u, d_v\right)}{\sum_j PDN\left(p_u, d_j\right)} \times S_0\left(d_v\right), \textit{if } PDN\left(p_u, d_v\right) \neq 0 \\ 0 \qquad \textit{otherwise} \end{cases} \tag{19}$$

Similarly, for any given domain d_u and protein p_v, *we* can obtain a distribution probability of d_u to p_v as follows:

$$DPM^{DP}\left(d_u, p_v\right) = \begin{cases} \frac{PDN^T\left(d_u, p_v\right)}{\sum_j PDN^T\left(d_u, p_j\right)} \times S_0\left(p_v\right), \textit{if } PDN^T\left(d_u, p_v\right) \neq 0 \\ 0 \qquad \textit{otherwise} \end{cases} \tag{20}$$

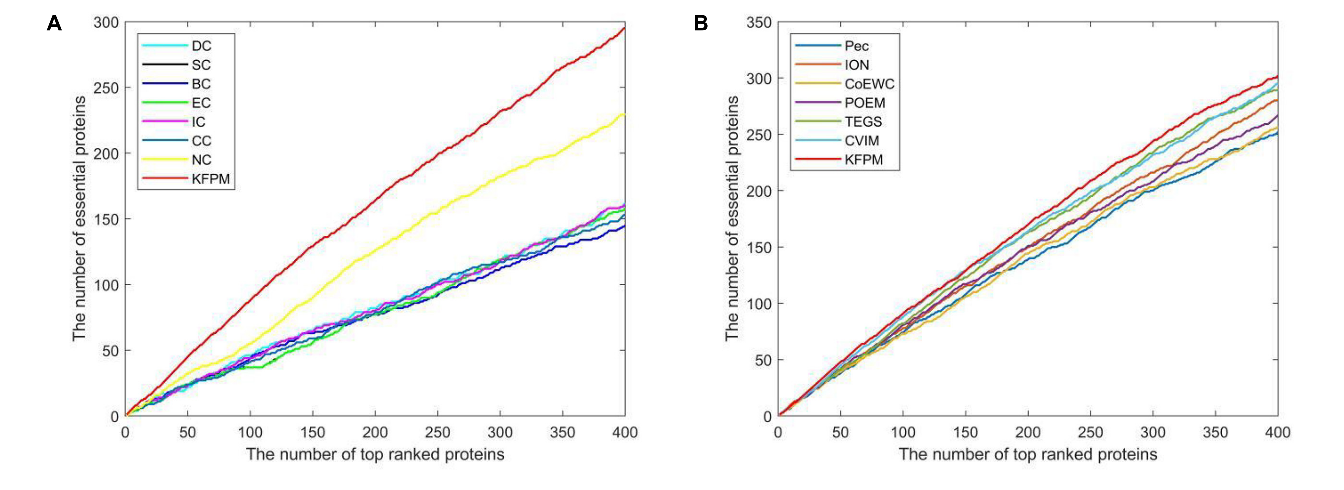


FIGURE 4 | Performance comparison between KFPM and 13 state-of-the-art methods based on the method of Jackknife under the DIP database. **(A)** Comparison between KFPM and DC, SC, BC, EC, IC, CC, NC. **(B)** Comparison between KFPM and Pec, CoEWC, POEM, ION, TEGS, CVIM.

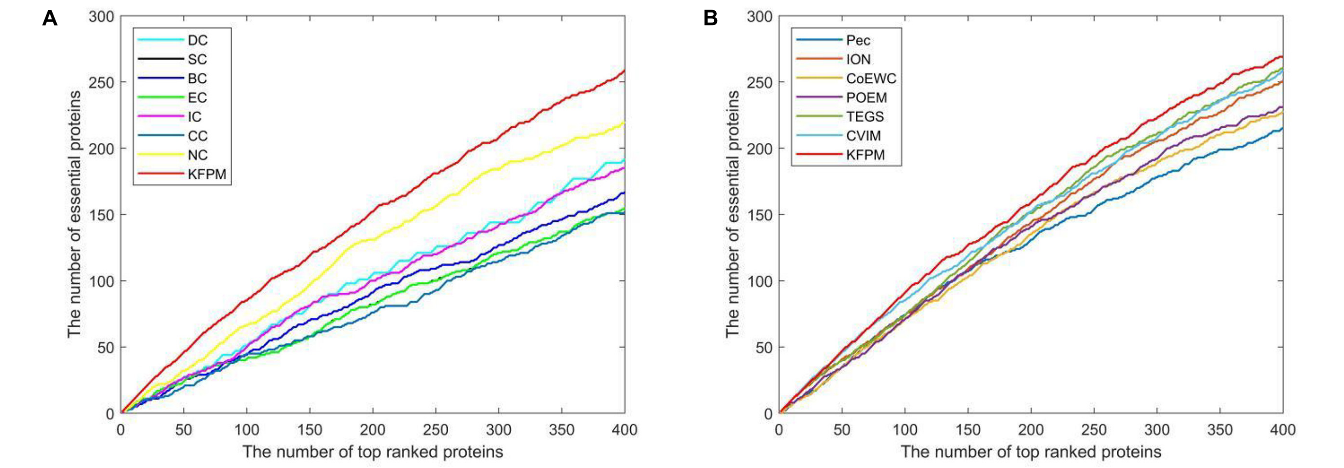


FIGURE 5 | Performance comparison between KFPM and 13 state-of-the-art methods based on the method of Jackknife under the Krogan database. **(A)** Comparison between KFPM and DC, SC, BC, EC, IC, CC, NC. **(B)** Comparison between KFPM and Pec, CoEWC, POEM, ION, TEGS, CVIM.

For any given domain d_u and domain d_v, let $DDN\left(d_u, d_v\right) = \frac{WDD\left(d_u,d_v\right)}{\left(1+\max\left(WDD\left(d_u,d_v\right)\right)\right)^2}$, *we* can obtain a distribution probability of d_u to d_v as follows:

$$DPM^{DD}\left(d_u, d_v\right) = \begin{cases} \frac{PDD\left(d_u,d_v\right)}{\sum_j PDD\left(d_u,d_j\right)} \times S_0\left(d_v\right), \, if \; DDN\left(d_u, d_v\right) \neq 0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{21}$$

Hence, based on above description, we can obtain a novel distribution probability matrix NDPM as follows:

$$NDPM = \begin{bmatrix} DPM^{PP} & DPM^{PD} \\ DPM^{DP} & DPM^{DD} \end{bmatrix} \tag{22}$$

Based on above formula (22), let $S_{(t)}$ denote critical scores of proteins obtained at the t^{th} round of iteration, then we can calculate the final critical scores of proteins by an improved PageRank algorithm according to the following Eq. 23:

$$S_{(t+1)} = \alpha \times NDPM \times S_{(t)} \left(1 - \alpha\right) \times S_0 \tag{23}$$

Here,$\alpha \in (0, 1)$ is a parameter used to adjust the iterative ratio.

Based on the above descriptions, the process of KFPM can be described in detail as follows:

Algorithm: KFPM

Input: Original PPI network, orthologous data, subcellular data, gene expression data and domain data, iteration termination condition ε, parameter α and θ.

Output: Final critical scores of proteins.

TABLE 2 | Commonalities and differences between KFPM and 13 competitive methods based on top 200 ranked proteins under the DIP database.

Different methods (Mi)	| KFPM∩Mi|	| KFPM-Mi|	Percentage of key proteins in {KFPM-Mi}	Percentage of key proteins in {Mi-KFPM}
DC	37	163	84.66%	30.67%
IC	35	165	84.85%	30.30%
EC	31	169	84.62%	29.59%
SC	31	169	84.62%	29.59%
BC	30	170	84.71%	30.00%
CC	27	173	84.39%	31.79%
NC	82	118	83.90%	46.61%
Pec	95	105	80.95%	51.43%
CoEWC	94	106	79.25%	54.72%
POEM	100	100	79.00%	60.00%
ION	100	100	78.00%	58.00%
TEGS	110	90	74.44%	66.67%
CVIM	120	80	72.50%	65.00%

TABLE 3 | Commonalities and differences between KFPM and 13 competitive methods based on top 200 ranked proteins under the Krogan database.

Different methods (Mi)	| KFPM∩Mi|	| KFPM-Mi|	Percentage of key proteins in {KFPM-Mi}	Percentage of key proteins in {Mi-KFPM}
DC	68	132	75.76%	34.85%
IC	70	130	76.15%	30.77%
EC	54	146	78.77%	26.03%
SC	54	146	78.77%	26.03%
BC	52	148	78.38%	32.43%
CC	44	156	80.13%	26.92%
NC	102	98	72.45%	43.88%
Pec	92	108	70.37%	44.44%
CoEWC	92	108	70.37%	48.15%
POEM	100	100	69.00%	51.00%
ION	88	112	72.32%	58.93%
TEGS	106	94	64.89%	56.38%
CVIM	144	56	66.07%	53.57%

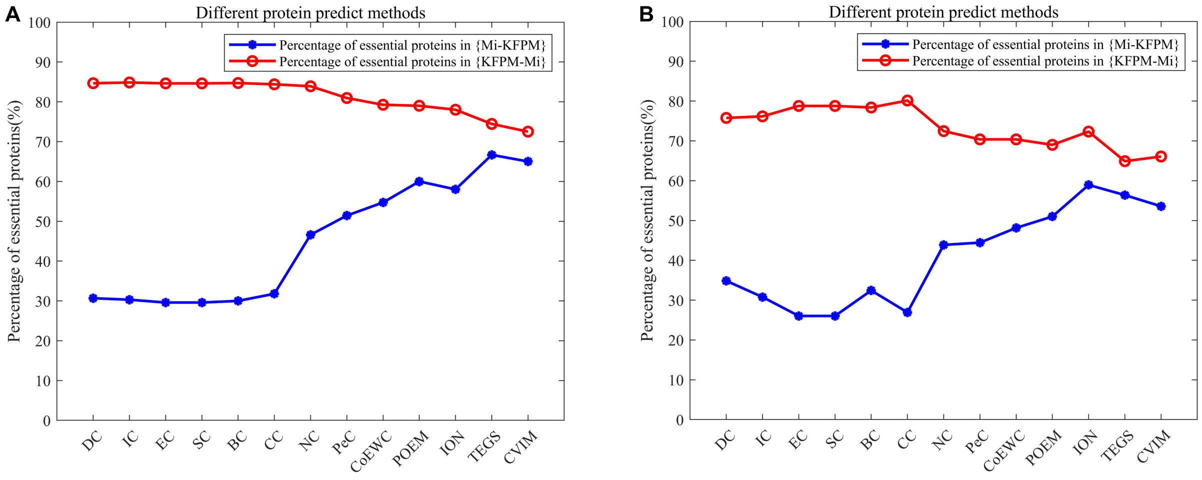


FIGURE 6 | Intuitive comparison of predictive performance between KFPM and 13 competitive methods. The *X*-axis represents 13 methods. The *Y*-axis denotes the percentage of essential proteins in {KFPM−Mi} or {Mi−KFPM}.

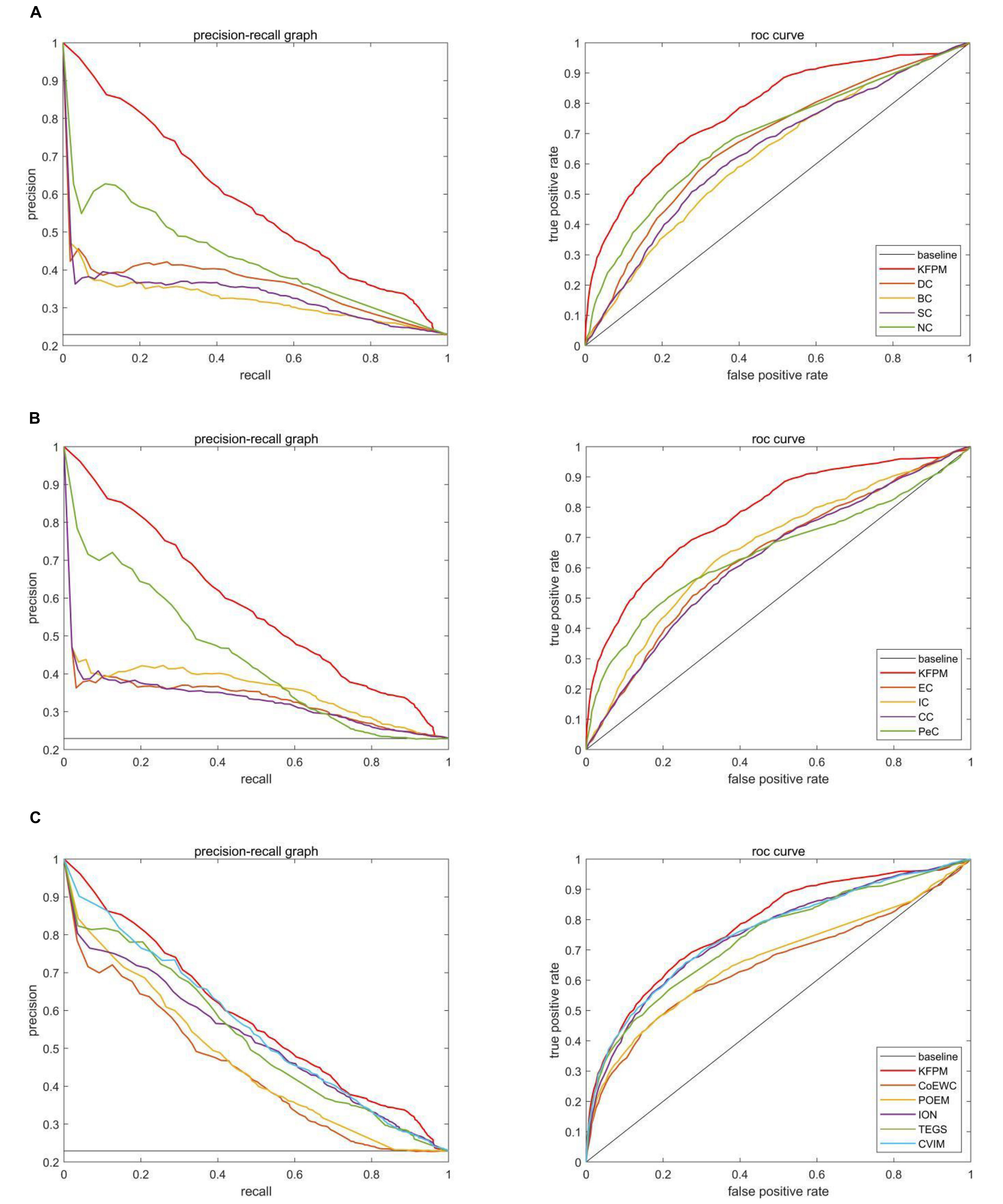


FIGURE 7 | Comparison of PR curves and ROC curves between KFPM and 13 competing methods based on the DIP database. **(A)** PR curves and the ROC curves of DC, BC, SC, and NC. **(B)** PR curves and ROC curves of EC, IC, CC and Pec. **(C)** PR curves and ROC curves of CoEWC, POEM, ION, TEGS, and CVIM.

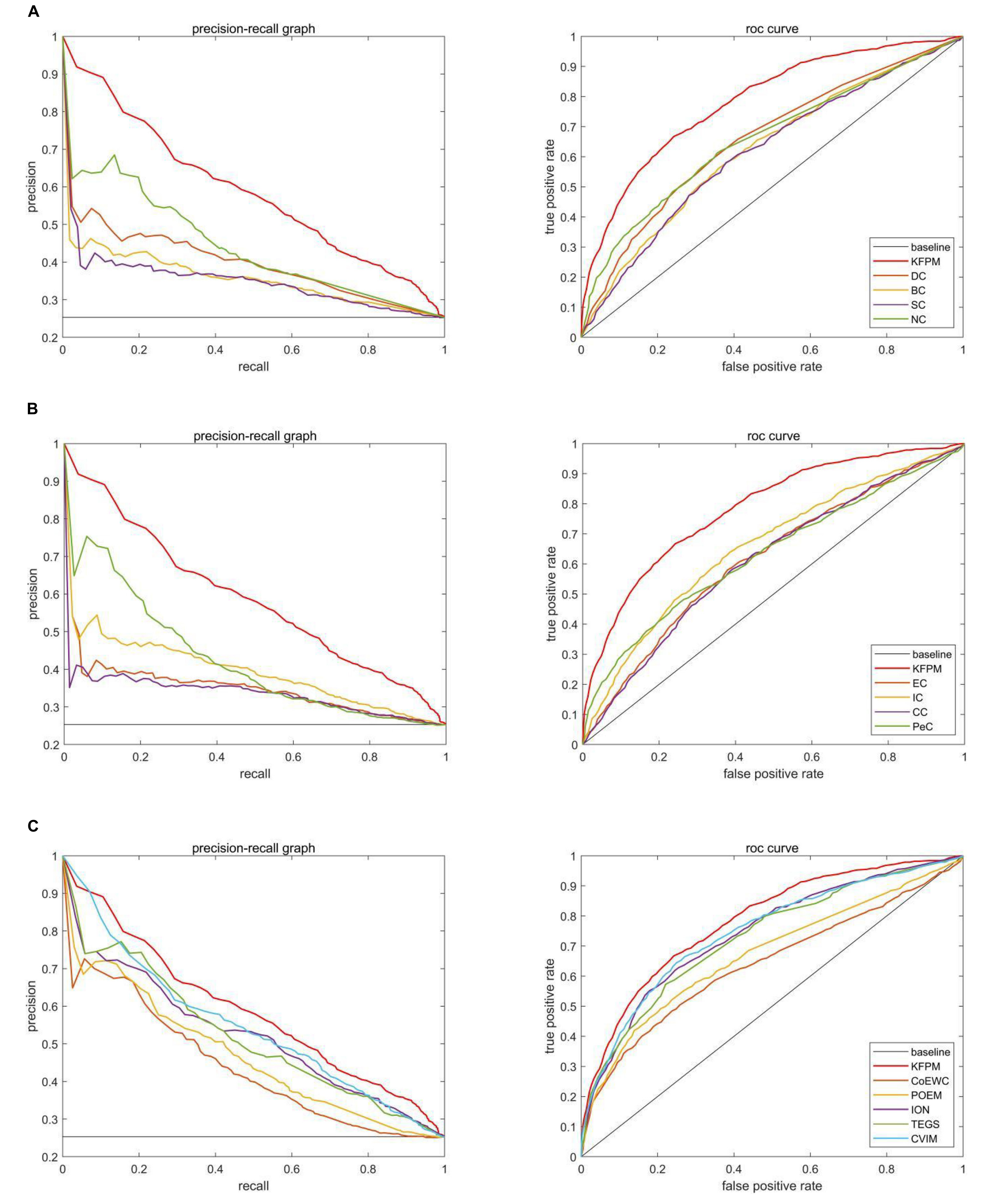


FIGURE 8 | Comparison of PR curves and ROC curves between KFPM and 13 competing methods based on the Krogan database. **(A)** PR curves and the ROC curves of DC, BC, SC, and NC. **(B)** PR curves and ROC curves of EC, IC, CC, and Pec. **(C)** PR curves and ROC curves of CoEWC, POEM, ION, TEGS, and CVIM.

- **Step 1:** Establishing the heterogeneous protein-domain network according to formulas (1)–(3);
- **Step 2:** Calculating initial scores of proteins and domains in the heterogeneous protein-domain network according to formulas (4)–(17);
- **Step 3:** Establishing the transition probability matrix *NDPM* according to formulas (18)–(22);
- **Step 4:** Computing $S_{(t+1)}$ by equation (23), let $t = t+1$;
- **Step 5:** Repeating step4 until $\left\| S_{(t+1)} - S_{(t)} \right\|^2 < \varepsilon$;
- **Step 6:** Outputting the top k% predicted proteins in the descending order.

RESULTS

Comparison Between KFPM and Representative Methods

In this section, we will compare KFPM with 13 state-of-the-art predictive methods based on the DIP and Krogan databases separately. **Figure 2** illustrates experimental results based on the DIP database, from which, it can be seen that KFPM can achieve predictive accuracy of 96.08, 83.14, 70.59, 61.78, 56.33, and 51.18% in top 1, 5, 10, 15, 20, and 25% predicted proteins, respectively, which are better than all 13 competitive methods, except in the top 15% predicted proteins, is a little lower than CVIM. **Figure 3** shows experimental results based on the Krogan database, from which, it can be seen that KFPM can achieve predictive accuracy of 91.89, 79.89, 70.03, 63.88, 59.26, and 55.56% in top 1, 5, 10, 15, 20, and 25% predicted proteins separately, which are better than all 13 competitive methods as well, except in the top 1% predicted proteins, is a little lower than CVIM. Hence, from above two kinds of experimental results, as a whole, we can conclude that the prediction performance of KRPM is better than all these 13 state-of-the-art methods.

Validation With Jackknife Methodology

The method of Jackknife (Holman et al., 2009) can effectively estimate the advantages and disadvantages of essential protein prediction models. Therefore, in this section, we will further utilize the method of Jackknife to compare KFPM with 13 competitive methods. **Figure 4** shows the comparison result based on top 400 predicted proteins under the DIP dataset. From observing **Figures 4A,B**, it is easy to see that the prediction performance of KFPM is not only better than the first category of methods that are based on topological features of PPI networks only, such as DC, SC, BC, EC, IC, CC, and NC, but also better than the second category of methods that are based on the combination of biological data of proteins and PPI networks, such as Pec, CoEWC, POEM, ION, TEGS, and CVIM, simultaneously. Especially, comparing with CVIM that can achieve the best predictive performance in all these competitive methods, although the performance curves of KFPM and CVIM overlap at some times, but with the number of candidate proteins increasing, the prediction performance of KFPM will become higher and higher than CVIM. **Figure 5** illustrates the comparison result based on top 600 predicted proteins under the Krogan dataset. From observing **Figures 5A,B**, it is obvious that KFPM can achieve better performance than both the first category of methods such as DC, SC, BC, EC, IC, CC, and NC, and the second category of methods such as Pec, CoEWC, POEM, ION, TEGS, and CVIM, as well. Hence, based on above description, we can conclude that the detective ability of KFPM is superior to all these 13 existing advanced methods.

Difference Analysis of KFPM and Competitive Methods

In order to better analyze the difference and uniqueness of KFPM and state-of-the-art predictive methods, in this section, we will compare KFPM with 13 competitive methods based on top 200 predicted proteins under the DIP and Krogan databases, respectively. Comparison results are shown in **Tables 2**, **3**, where Mi represents one of these 13 predictive methods, |KFPM ∩ Mi| represents the number of common essential proteins recognized by both KFPM and Mi. |KFPM−Mi| denotes the number of essential proteins that were detected by KFPM but not by Mi. {KFPM−Mi} is the set of essential proteins predicted by KFPM but ignored by Mi. {Mi−KFPM} is the set of essential proteins predict by Mi but ignored by KFPM. From observing **Tables 2**, **3**, we can see that the proportion of key proteins in {KFPM−Mi} is higher than the percentage of key proteins in {Mi−KFPM}, which means that KFPM can screen out more essential proteins that are not found by competing methods. **Figure 6** shows the superiority of KFPM more intuitively.

Validation by Receiver Operating Characteristic Curve

In this section, we will further utilize the ROC (Receiver Operating Characteristic) curve to evaluate the detection

TABLE 4 | AUCs achieved by KFPM and 13 competitive methods based on the DIP and Krogan databases.

Method	AUCs (based on DIP)	AUCs (based on Krogan)
DC	0.6704	0.6583
IC	0.6657	0.6573
EC	0.6384	0.6167
BC	0.625	0.6248
SC	0.6384	0.6167
CC	0.6291	0.6114
NC	0.6879	0.6584
Pec	0.6329	0.6316
CoEWC	0.6513	0.6404
POEM	0.6662	0.6726
TEGS	0.7386	0.7287
ION	0.7522	0.7413
CVIM	0.7559	0.7458
KFPM	0.7802	0.7833

TABLE 5 | The Number of essential proteins recognized by KFPM and 13 competing methods based on the Gavin database.

Methods	Top1%(19)	Top5%(93)	Top10%(196)	Top15%(279)	Top20%(371)	Top25%(464)
DC	7	36	101	158	222	264
IC	16	55	119	163	213	254
CC	11	45	93	135	180	221
BC	9	40	85	122	162	201
SC	0	17	87	130	190	240
EC	0	38	94	134	166	209
NC	11	51	123	170	213	259
CoEWC	16	69	136	190	237	275
Pec	15	69	142	193	238	285
ION	17	73	150	207	263	312
POEM	17	74	148	199	249	296
CVIM	16	80	160	219	271	322
KFPM	19	86	169	216	279	332

TABLE 6 | Influence of the parameter α on the prediction accuracy of KFPM based on the DIP database.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rank									
Top1% (51)	47	47	47	48	49	48	47	46	45
Top5% (255)	209	210	211	214	212	210	210	210	210
Top10% (510)	358	363	360	363	360	361	360	363	358
Top15% (764)	466	473	474	477	472	470	465	466	461
Top20% (1019)	572	574	575	570	574	570	566	564	568
Top25% (1274)	647	648	648	648	652	654	653	648	643

TABLE 7 | Influence of the parameter α on the prediction accuracy of KFPM based on the Krogan database.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rank									
Top1% (37)	36	36	35	35	34	34	34	33	31
Top5% (184)	147	147	149	148	147	149	150	149	147
Top10% (367)	264	264	260	257	257	255	254	252	255
Top15% (551)	368	365	361	357	353	348	345	346	339
Top20% (734)	441	441	443	440	435	433	426	422	42
Top25% (918)	499	503	503	504	510	504	505	502	482

performance of KFPM. The closer the ROC curve is to the upper left corner, the higher the recall rate of the model (Hanley and Mcneil, 1982). **Figures 7**, **8** show ROC curves and PR (Precision Recall) curves of KFPM and 13 competing methods under the DIP and Krogan databases, respectively. As shown in **Figure 7**, it is obvious that KFPM can achieve better predictive performance than all these 13 state-of-the-art methods based on the DIP database, although the ROC curves of KFPM and CVIM overlap partially in **Figure 7C**. As shown in **Figure 8**, it is easy to see that the predictive performance of KFPM is better than all these 13 state-of-the-art methods based on the Krogan database as well. **Table 4** shows the superiority of KFPM more intuitively based on the performance indicator of AUCs (Area Under roc Curves).

Additionally, in order to verify the applicability of KFPM, we further compared KFPM with 13 competitive methods based on the Gavin database. As shown in **Table 5**, it is easy to see that the prediction performance of KFPM is better than all competing methods, especially, in the top 1% candidate proteins, the number of true essential proteins recognized by KFPM is 19, which means that the recognition rate of KFPM can reach 100%. Hence, we can draw a conclusion as well that KFPM has satisfactory applicability.

Analysis of Parameters

In KFPM, we have introduced a parameter $\alpha \in (0, 1)$ to adjust the iterative ratio. Therefore, we will estimate the effect of α on the prediction accuracy of KFPM in this section. Experimental results based on the DIP and Krogan databases are shown in **Tables 6**, **7** separately. From observing these two tables, it is easy to see that, as a whole, KFEM can achieve the best predictive performance when the value of α is set to 0.5. Moreover, the KFPM can obtain the best performance when the value of θ in formula (15) is set to 0.7.

DISCUSSION

Essential proteins are indispensable proteins for the survival and reproduction of organisms. In recent years, identification of essential proteins has become a research hotspot. It takes a lot of time and money to predict the essential proteins through traditional biological experiments. Therefore, many researchers focus on designing effective predictive models by combining PPI networks. With gradual improvement of high-throughput techniques, prediction methods with more accurate predictive performance have been proposed successively based on combination of biological data of proteins and PPI networks. Inspired by this, a novel predictive model called KFPM has been proposed in this paper, which can achieve satisfactory predictive accuracy by combining topological characteristics of a newly constructed protein-domain interaction network and functional characteristics of proteins. Experimental results demonstrate the superiority of KFPM, which may provide a useful tool for future researches on prediction of key proteins.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

XH and LW conceived and designed the study. XH, ZC, and LK obtained and processed datasets. XH and LK wrote this manuscript. YT, LW, and LK provided suggestions and supervised the research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61873221), the Research Foundation of Education Bureau of Hunan Province (No. 20B080), and the Natural Science Foundation of Hunan Province (Nos. 2018JJ4058 and 2019JJ70010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.708162/full#supplementary-material

REFERENCES

Ahmed, N. M., Chen, L., Li, B., Liu, W., and Dai, C. (2021). A random walk-based method for detecting essential proteins by integrating the topological and biological features of ppi network. *Soft Comput.* doi: 10.1007/s00500-021-05780-8

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffithsjones, S., et al. (2004). The pfam protein families database. *Nucleic Acids Res.* 32, D138–D141. doi: 10.1093/nar/gkh121

Binder, J. X., Sune, P. F., Kalliopi, T., Christian, S., O'DonoghueSeán, I., Reinhard, S., et al. (2014). Compartments: unification and visualization of protein subcellular localization evidence. *Database J. Biol. Databases Curation* 2014:bau012. doi: 10.1093/database/bau012

Bonacich, P. (1987). Power and centrality: a family of measures. *Am. J. Soc.* 92, 1170–1182. doi: 10.2307/2780000

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., and Hester, E. T. (1998). SGD: saccharomyces genome database. *Nucleic Acids Res.* 26, 73–79. doi: 10.1093/nar/26.1.73

Estrada, E. (2010). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 6, 35–40. doi: 10.1002/pmic.200500209

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71:056103. doi: 10.1103/PhysRevE.71.056103

Gabriel, O., Thomas, S., Kristoffffer, F., Tina, K., David, N. M., Sanjit, R., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636. doi: 10.1038/nature04532

Hahn, M. W., and Kern, A. D. (2004). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806. doi: 10.1093/molbev/msi072

Hanley, J. A., and Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143:29. doi: 10.1148/radiology.143.1.7063747

Holman, A. G., Davis, P. J., Foster, J. M., Carlow, C. K., and Kumar, S. (2009). Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi. *BMC Microbiol.* 9:243. doi: 10.1186/1471-2180-9-243

Horyu, D., and Hayashi, T. (2013). Comparison between pearson correlation coefficient and mutual information as a similarity measure of gene expression profiles. *Japanese J. Biometr.* 33, 125–143. doi: 10.5691/jjb.33.125

Jeong, H., Mason, S., and Barabási, A. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Jiang, Y., Wang, Y., Pang, W., Chen, L., Sun, H., Liang, Y., et al. (2015). Essential protein identification based on essential protein–protein interaction prediction by integrated edge weights. *Methods* 83, 51–62. doi: 10.1016/j.ymeth.2015.04.013

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 96–103. doi: 10.1155/jbb.2005.96

Keretsu, S., and Sarmah, R. (2016). Weighted edge based clustering to identify protein complexes in protein–protein interaction networks incorporating gene expression profile. *Comput. Biol. Chem.* 65, 69–79. doi: 10.1016/j.compbiolchem.2016.10.001

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

Lei, X., Fang, M., Wu, F. X., and Chen, L. (2018a). Improved flower pollination algorithm for identifying essential proteins. *Bmc Syst. Biol.* 12:46. doi: 10.1186/s12918-018-0573-y

Lei, X., Yang, X., and Wu, F. (2018b). Artificial fish swarm optimization based method to identify essential proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 495–505. doi: 10.1109/TCBB.2018.2865567

Lei, X., Zhao, J., Fujita, H., and Zhang, A. (2018c). Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets. *Knowledge Based Syst.* 151, 136–148. doi: 10.1016/j.knosys.2018.03.027

Li, M., Wang, J., Xiang, C., Wang, H., and Yi, P. (2011). A local average connectivity-based method for identifying essential proteins from the network level. *Comput. Biol. Chem.* 35, 143–150. doi: 10.1016/j.compbiolchem.2011.04.002

Li, M., Zhang, H., Wang, J. X., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *Bmc Syst. Biol.* 6:15. doi: 10.1186/1752-0509-6-15

Li, S., Chen, Z., He, X., Zhang, Z., Pei, T., Tan, Y., et al. (2020). An iteration method for identifying yeast essential proteins from weighted PPI network based on topological and functional features of proteins. *IEEE Access* 8, 90792–90804. doi: 10.1109/access.2020.2993860

Meng, Z., Kuang, L., Chen, Z., Zhang, Z., and Wang, L. (2021). Method for essential protein prediction based on a novel weighted protein-domain interaction network. *Front. Genet.* 12:645932. doi: 10.3389/fgene.2021.645932

Mewes, H. W., Frishman, D., Mayer, K. F. X., Munsterkotter, M., Noubibou, O., Pagel, P., et al. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34, D169–D172. doi: 10.1093/nar/gkj148

Peng, W., Wang, J. X., Wang, W. P., Liu, Q., Wu, F. X., and Pan, Y. (2012). Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *Bmc Syst. Biol.* 6:87. doi: 10.1186/1752-0509-6-87

Przulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics* 20, 340–348. doi: 10.1093/bioinformatics/btg415

Ren, J., Wang, J., Min, L., Wang, H., and Liu, B. (2011). "Prediction of essential proteins by integration of ppi network topology and protein complexes information," in *Bioinformatics Research & Applications-international Symposium*, eds J. Chen, J. Wang, and A. Zelikovsky (Berlin: Springer), 12–24. doi: 10.1186/1477-5956-11-S1-S20

Saccharomyces Genome Deletion Project (2012). Available online at: http://yeastdeletion.stanford.edu/ (accessed June 20, 2012).

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. *Soc. Networks* 11, 1–37. doi: 10.1016/0378-8733(89)90016-6

Tang, X., Wang, J., Zhong, J., and Pan, Y. (2014). Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 407–418. doi: 10.1109/TCBB.2013.2295318

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* 310, 1152–1158. doi: 10.1126/science.1120499

Wang, H., Li, M., Wang, J. X., and Pan, Y. (2011). "A new method for identifying essential proteins based on edge clustering coefficient," in *Bioinformatics Research and ApplicationsISBRA 2011, LNBI*, eds J. Chen, J. Wang, and A. Zelikovsky (Berlin, Heidelberg: Springer), 87–98. doi: 10.1007/978-3-642-21260-4_12

Wang, J. X., Li, M., Wang, H., and Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 1070–1080. doi: 10.1109/TCBB.2011.147

Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. *J. Theor. Biol.* 223, 45–53. doi: 10.1016/S0022-5193(03)00071-7

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Xiao, Q., Wang, J., Peng, X., and Wu, F. X. (2013). Detecting protein complexes from active protein interaction networks constructed with dynamic gene expression profiles. *Proteome Sci.* 11:S20. doi: 10.1186/1477-5956-11-S1-S2

Zhang, B., Wang, Q., Liu, S., Dong, H., Zheng, S., Zhao, L., et al. (2020). Data-Driven abnormity assessment for low-voltage power consumption and supplies based on CRITIC and improved radar chart algorithms. *IEEE Access* 8, 27139–27151. doi: 10.1109/access.2020.2970098

Zhang, R., and Lin, Y. (2009). DEG 5.0.A database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* 37, D455–D458.

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

Zhang, W., Xu, J., and Zou, X. (2019). Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and go annotation data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 2053–2061. doi: 10.1109/TCBB.2019.2916038

Zhang, W., Xue, X., Xie, C., Li, Y., Liu, J., Chen, H., et al. (2021). CEGSO: boosting essential proteins prediction by integrating protein complex, gene expression, gene ontology, subcellular localization and orthology information. *Interdiscip. Sci. Comput. Life Sci.* doi: 10.1007/s12539-021-00426-7 [Epub ahead of print],

Zhang, X., Xu, J., and Wang, X. X. (2013). A new method for the discovery of essential proteins. *PLoS One* 8:e58763. doi: 10.1371/journal.pone.0058763

Zhang, Z., Luo, Y., Hu, S., Li, X., and Zhao, B. (2020). A novel method to predict essential proteins based on tensor and hits algorithm. *Human Genomics* 14:14. doi: 10.1186/s40246-020-00263-7

Zhao, B., Han, X., Liu, X., Luo, Y., Hu, S., Zhang, Z., et al. (2020). A novel method to predict essential proteins based on diffusion distance networks. *IEEE Access* 8, 29385–29394. doi: 10.1109/ACCESS.2020.2972922

Zhao, B., Zhao, Y., Zhang, X., Zhang, Z., Zhang, F., and Wang, L. (2019). An iteration method for identifying yeast essential proteins from heterogeneous network. *BMC Bioinform.* 20:355. doi: 10.1186/s12859-019-2930-2

Zhao, B. H., Wang, J. X., Li, M., Wu, F. X., and Pan, Y. (2014). Prediction of essential proteins based on overlapping essential modules. *IEEE Trans. Nanobioence* 13, 415–424. doi: 10.1109/TNB.2014.2337912

Zhong, J., Tang, C., Peng, W., Xie, M., and Yang, J. (2020). A novel essential protein identification method based on PPI networks and gene expression data. *Res. Square [Preprint]* doi: 10.21203/rs.3.rs-55902/v2

Zz, A. Jr., Jg, A., and Fxw, B. (2019). Predicting essential proteins from protein-protein interactions using order statistics. *J. Theor. Biol.* 480, 274–283. doi: 10.1016/j.jtbi.2019.06.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 He, Kuang, Chen, Tan and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.