frontiers in Genetics

ORIGINAL RESEARCH
published: 17 March 2021
doi: 10.3389/fgene.2021.645932

Method for Essential Protein Prediction Based on a Novel Weighted Protein-Domain Interaction Network

Zixuan Meng[1], Linai Kuang[1], Zhiping Chen[2], Zhen Zhang[2], Yihong Tan[2], Xueyong Li[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

OPEN ACCESS

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

Reviewed by:
Yuhua Yao,
Hainan Normal University, China
Xinguo Lu,
Hunan 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: 24 December 2020
Accepted: 15 February 2021
Published: 17 March 2021

Citation:
Meng Z, Kuang L, Chen Z, Zhang Z, Tan Y, Li X 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

In recent years a number of calculative models based on protein-protein interaction (PPI) networks have been proposed successively. However, due to false positives, false negatives, and the incompleteness of PPI networks, there are still many challenges affecting the design of computational models with satisfactory predictive accuracy when inferring key proteins. This study proposes a prediction model called WPDINM for detecting key proteins based on a novel weighted protein-domain interaction (PDI) network. In WPDINM, a weighted PPI network is constructed first by combining the gene expression data of proteins with topological information extracted from the original PPI network. Simultaneously, a weighted domain-domain interaction (DDI) network is constructed based on the original PDI network. Next, through integrating the newly obtained weighted PPI network and weighted DDI network with the original PDI network, a weighted PDI network is further constructed. Then, based on topological features and biological information, including the subcellular localization and orthologous information of proteins, a novel PageRank-based iterative algorithm is designed and implemented on the newly constructed weighted PDI network to estimate the criticality of proteins. Finally, to assess the prediction performance of WPDINM, we compared it with 12 kinds of competitive measures. Experimental results show that WPDINM can achieve a predictive accuracy rate of 90.19, 81.96, 70.72, 62.04, 55.83, and 51.13% in the top 1%, top 5%, top 10%, top 15%, top 20%, and top 25% separately, which exceeds the prediction accuracy achieved by traditional state-of-the-art competing measures. Owing to the satisfactory identification effect, the WPDINM measure may contribute to the further development of key protein identification.

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

INTRODUCTION

Accumulating evidence indicates that proteins have a tremendous impact on almost all life activities. Essential proteins cannot only maintain normal biological processes but also ensure the integrity of cell functions. With the development of biotechnology (Lu et al., 2019, 2020), more and more essential proteins have been discovered by biological experiments in recent years. However, because biological experiments are quite costly and time-consuming, an increasing number of computational models have been proposed to identify essential proteins based on the topological features of PPI networks. For instance, based on the rule of centrality-lethality (Jeong et al., 2001), researchers have proposed a series of prediction models, which have been designed successively to infer potential critical proteins. These include Information Centrality (IC) (Stephenson and Zelen, 1989), Degree Centrality (DC) (Hahn and Kern, 2004), Subgraph Centrality (SC) (Ernesto and Rodriguez-Velazquez, 2005), Closeness Centrality (CC) (Wuchty and Stadler, 2003), Betweenness Centrality (BC) (Jop et al., 2005), Neighbor Centrality (NC) (Wang et al., 2012), and local average connectivity (LAC) (Li et al., 2015). Wang et al. (2011) designed a predictive model named SoECC by combining the features of edges and nodes and taking advantage of the edge clustering coefficient effectively. Lin et al. (2008) introduced two kinds of prediction models such as the Maximum Neighborhood Component (MNC) and the Density of Maximum Neighborhood Component (DMNC) to infer essential proteins, respectively. However, these prediction models cannot achieve high identification accuracy owing to the incompleteness of current PPI networks (Chen and Yuan, 2006).

Hence, to address this problem, some different methods based on both biological information on proteins and the topological properties of PPI networks have been proposed to detect essential proteins. For example, Li et al. (2012) proposed a calculation method called Pec by uniting the gene expression data with the centrality-lethality rule to identify key proteins from PPI networks. Zhang et al. (2013) presented a method based on integrating the topological features of PPI networks with the co-expressions of proteins. Peng et al. (2012) designed a prediction method called ION based on topological features extracted from the PPI network and the orthologous information of proteins. Additionally, inspired by the model of Degree Centrality, Tang et al. (2014) developed an identification model for predicting essential proteins by combining the Person correlation coefficient (PCC) and the edge clustering coefficient (ECC) with the gene expression data of proteins. Kim (2012) proposed a method for predicting key proteins by implementing a machine learning algorithm on both Gene Ontology and topological information of PPI networks. Luo et al. (2015) developed a computational model by integrating the local interaction density with protein complexes to detect key proteins. Li et al. (2016) designed a method for identifying essential proteins by adopting the subcellular localization and orthologous information. Luo and Kuang (2014) proposed a prediction model called CDLC to detect essential proteins by employing the dynamic local average connectivity and in-degree of proteins in complexes. Zhang et al. (2018) introduced a calculative algorithm named TEO for inferring essential proteins by integrating gene ontology annotation information and the gene expression data with PPI networks. Zhong et al. (2013) designed a learning algorithm to predict essential proteins by combining the biological information of proteins with PPI networks. Shang et al. (2016) introduced a strategy to detect essential proteins through integrating the RNA-Seq dataset and biological information of proteins with dynamic PPI networks. Zhang et al. (2016) introduced a prediction measure named PINs for identifying essential proteins based on gene expression profiles and PPI networks through integrating five approaches including the DC, BC, SC, CC, and the eigenvector centrality (EC) (Bonacich, 1987).

This study proposes a novel prediction model called WPDINM that can be used to detect key proteins by combing a weighted protein-domain interaction (PDI) network with the biological information containing the subcellular localization and orthologous information of proteins. WPDINM is based on the original PPI network and the original PDI network, obtained by known protein-protein interactions (PPIs) and known protein-domain associations that have been downloaded from benchmark databases. In this prediction model, a weighted PPI network and a weighted domain-domain interaction (DDI) network are established first, based on the gene expression data of proteins and the topological information of the original networks respectively. Then, a weighted PDI network is constructed by combining these two newly constructed weighted networks. Next, based on the weighted PDI network, initial scores are assigned to proteins based on the biological information of proteins such as the subcellular localization and orthologous information of proteins, and a novel iterative method is implemented to estimate the criticality of proteins.

Different from traditional prediction models, in WPDINM, the Discrete Fourier transform (DFT) is applied to the gene expression profiles of proteins to calculate the weight between proteins, which can translate gene expression profiles from the time domain to frequency domain effectively. A novel weighted PDI network is then constructed by integrating a weighted DDI network and weighted PPI network. Moreover, by taking into account the associations between proteins, a new directed distribution network is designed to calculate the rankings of proteins iteratively, based on the weighted PDI network. Finally, to evaluate the prediction performance of WPDINM, the WPDINM is compared with other competitive measures such as SC (Ernesto and Rodriguez-Velazquez, 2005), DC (Hahn and Kern, 2004), IC (Stephenson and Zelen, 1989), CC (Wuchty and Stadler, 2003), BC (Jop et al., 2005), NC (Wang et al., 2012), EC (Bonacich, 1987), Pec (Li et al., 2012), CoEWC (Zhang et al., 2013), TEGS (Zhang et al., 2019), ION (Peng et al., 2012), and POEM (Zhao et al., 2014). Experimental results indicate that WPDINM can achieve better prediction accuracies than competing prediction models, achieving 90.19, 81.96, 70.72, 62.04, 55.83, and 51.13% in the top 1%, top 5%, top 10%, top 15%, top 20%, and top 25% of predicted proteins separately.

MATERIALS AND METHODS

Experimental Data

To construct original PPI networks, we first download known PPIs from three different databases including the DIP database (Xenarios et al., 2002), the Gavin database (Gavin et al., 2006), and the Krogan database (Krogan et al., 2006), respectively. After removing duplicated interactions, we finally obtain three different datasets such as the DIP-based dataset, consisting of 24,743 known PPIs between 5,093 proteins, the Krogan-based dataset, consisting of 14,317 known PPIs between 3,672 proteins, and the Gavin-based dataset consisting of 7,669 known PPIs between 1,855 proteins. Next, we further download known domains from the Pfam database (Bateman et al., 2004), and after preprocessing, obtain a dataset consisting of 1,107 different domains. Based on these three kinds of datasets obtained from the DIP database, the Gavin database, and the Krogan database, we finally construct three kinds of original PPI networks and corresponding matrices with dimensions of $5,093 = 1,107$, $1,855 = 1,107$, and $3,672 = 1,107$ separately. The gene expression data is provided by Tu et al. (2005), which consists of 6,776 gene expression sequences with a length of 36.

In order to obtain the initial scores of proteins, we download the subcellular localization data from the COMPART-MENTS databases (Binder et al., 2014). As a result, we obtain a dataset consisting of 11 kinds of subcellular localizations, including the Extracellular, Peroxisome, Nucleus, Plasma, Endosome, Mitochondrion, Vacuole, Cytosol, Golgi, Cytoskeleton Endoplasmic, that are intimately linked with downloaded known key proteins. We also download the orthologous information of proteins from the InParanoid database (Gabriel et al., 2010). Furthermore, the set of essential proteins existing in Saccharomyces cerevisiae is was downloaded from four different databases including DEG (Zhang and Lin, 2009), MIPS (Mewes et al., 2006), SGD (Cherry et al., 1998), and SGDP (Saccharomyces Genome Deletion Project, 2012).

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

Step 1: Firstly, based on known PPIs downloaded from any given benchmark database, an original PPI network is obtained. Then, a weighted PPI network is further constructed by implementing the DFT method on the gene expression data of proteins and extracting topological features from the original PPI network.

Step 2: Based on known PPIs and known protein-domain interactions (PDIs) downloaded from given benchmark databases, a weighted DDI network is then constructed. Thereafter, a weighted PDI network is further established by integrating the weighted DDI network with the weighted PPI network.

Step 3: Then, by combining the weighted PDI network with biological information, including the orthologous information and subcellular information of proteins, each protein in the weighted PDI network is assigned an initial score.

Step 4: Finally, a novel prediction method based on the Page Rank algorithm is designed and applied on the weighted PDI network to compute the final scores of criticality for all proteins iteratively.

Construction of the Weighted PPI Network

In this section, based on the datasets consisting of known PPIs downloaded from three different databases, including the DIP database (Xenarios et al., 2002), the Gavin database (Gavin et al., 2006), and the Krogan database (Krogan et al., 2006), respectively, we construct three original PPI networks simultaneously. For convenience, let $OppiN = \{N_P, E_P\}$ represent a newly constructed original PPI network, where $N_P = \{p_1, p_2 \ldots p_O\}$ is the set of protein nodes in $OppiN$ and E_P is the set of edges between protein nodes in $OppiN$. Here, for two given proteins p_i and $p_j$in N_P, there is an edge $ed(p_i, p_j)$ between them in E_P, if and only if there is a known interaction between these two proteins. Based on the original PPI network $OppiN$, we can further obtain an $O \times O$ dimensional adjacency matrix$OppiM$ as follows: for any two protein p_i and p_j in N_P, there is $OppiM(i, j) = 1$, only if there is a known interaction between $p_i$and p_j, otherwise there is $OppiM(i, j) = 0$.

Next, based on $OppiN$, for any given protein p with a known gene expression sequence in N_P, let $Gep(p) = < Gep(p, 1), Gep(p, 2), \ldots, Gep(p, M) >$ represent the gene expression sequence of p, where $Gep(p, t)$ is the degree of gene expression at t^{th} time. As $Gep(p)$ is a time sequence with the length of M, then we can adopt the DFT method to convert it from time domains to frequency domains, since while $N \geq M$, the N-point Discrete Fourier can transform a $1 \times M$ dimensional time series vector $Gep(p)$to a $1 \times N$ dimensional spectrum vector $DF(p) = < |DGep(0)|, |DGep(1)|, \ldots |DGep(N-1)| >$as follows:

$$DGep(t) = \sum_{y=1}^{M} Gep(p, y) * W_M^{ty}, \text{where } t = 0, 1, \ldots..N\text{-}1 \quad (1)$$

$$W_M = e^{-i\frac{2\pi}{M}} \quad (2)$$

Thereafter, through combining the above formulas with the Gaussian kernel interaction profiles, for any two given proteins p_i and $p_j$with gene expression sequences in N_P, we can estimate the probability of association between them by calculating the spectra similarity as follows:

$$GK(p_i, p_j) = \exp\left(-\alpha_p \left|\left|DF(p_i) - DF(p_j)\right|\right|^2\right) \quad (3)$$

Here, α_p is the adjustment coefficient for the kernel bandwidth, which is defined as follows:

$$\alpha_p = \frac{\alpha_p'}{\frac{1}{|P_{Gep}|}\sum_{k=1}^{|P_{Gep}|} \left|\left|DF(p_k)\right|\right|^2} \quad (4)$$

In the above formula (4), P_{Gep} is the set of proteins with gene expression sequences in $OppiN$.

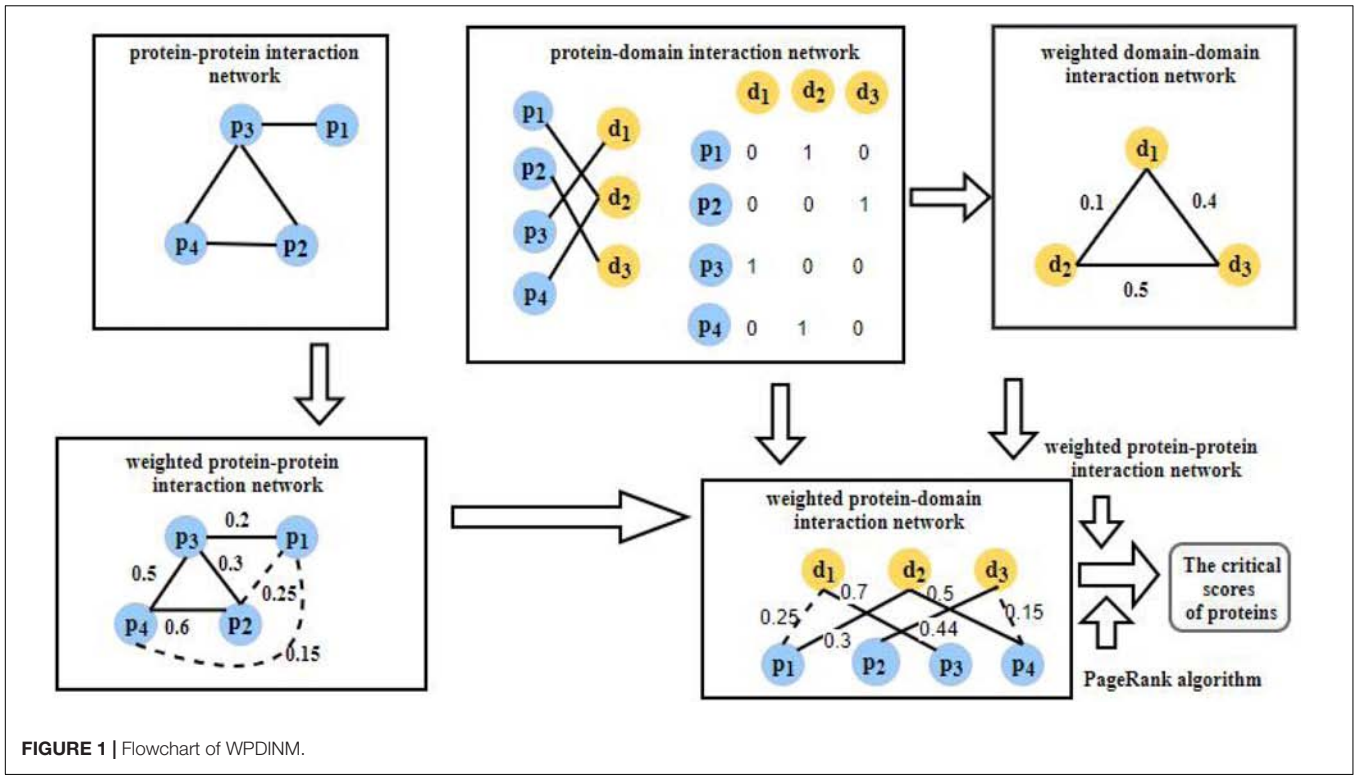

FIGURE 1 | Flowchart of WPDINM.

Additionally, for any two given proteins p_i and p_j without gene expression sequences in *OppiN*, we adopt the topological features extracted from the original PPI network *OppiN* to calculate the possibility of an association between them. Thus, the weight between p_i and p_j can be calculated as follows:

$$TFP\left(p_i, p_j\right) = \frac{\left|Com\left(p_i, p_j\right)\right| + 1}{\left(\left|Np\left(p_i\right)\right| + 1\right) * \left(\left|Np\left(p_j\right)\right| + 1\right)} \quad (5)$$

Here, $Np\left(p_i\right)$ and $Np\left(p_j\right)$ denote the set of neighboring protein nodes of p_i and p_j in *OppiN* separately, $Com\left(p_i, p_j\right)$ represents the set of common neighbors between p_i and p_j in *OppiN*, and $|X|$ means the number of different elements in the set X.

Integrating formula (3) and formula (5) for any two given proteins p_i and p_j in *OppiN*, we can calculate the possibility of an association between them, b, as follows:

$$PA(p_i, p_j) = \begin{cases} GK\left(p_i, p_j\right) : if\ p_i \text{ and } p_j \text{ have gene expression sequences} \\ TFP\left(p_i, p_j\right) : \qquad \text{Otherwise} \end{cases} \quad (6)$$

Based on the above formulas (6), a weighted PPI network *WppiN* can be constructed according to the following $O \times O$ dimensional matrix *WppiM*:

$$WppiM\left(p_i, p_j\right) =$$

$$\begin{cases} \beta * PA\left(p_i, p_j\right) + (1-\beta) * OppiM\left(p_i, p_j\right) \text{ if } OppiM\left(p_i, p_j\right) = 1 \\ TFP\left(p_i, p_j\right) * GK\left(p_i, p_j\right) \qquad \text{if } OppiM\left(p_i, p_j\right) = 0 \end{cases} \quad (7)$$

In the above formula (7), β is the scaling parameter with a value from 0 to 1.

Construction of the Original PDI Network

In this section, based on the dataset consisting of known PDIs downloaded from the Pfam database (Bateman et al., 2004), we construct an original PDI network $OpdiN = \{N_{PD}, E_{PD}\}$, where $N_{PD} = N_P \cup N_D, N_D = \{d_1, d_2 \ldots d_Q\}$ is the set of domain nodes in *OpdiN*, and E_{PD} is the set of edges between protein nodes in N_P and domain nodes in N_D. Here, for a given protein p_i and a given domain d_j in N_{PD}, there is an edge between them in E_{PD}, only if there is p_i belonging to d_j. Based on the original PDI network *OpdiN*, we can further obtain an $O \times Q$ dimensional adjacency matrix *OpdiM* as follows: for a given protein node p_i and a given domain node d_j in N_{PD}, there is $OpdiM\left(i, j\right) = 1$, if and only if there is p_i belonging to d_j, otherwise, there is $OpdiM\left(i, j\right) = 0$.

Construction of the Weighted DDI Network

For any two given domains d_i and d_j in *OpdiN*, in this section, we further obtain a $Q \times Q$ dimensional matrix *WddiM* by adopting the Gaussian kernel interaction profiles to estimate the association between d_i and d_j as follows:

$$WddiM\left(d_i, d_j\right) = \exp\left(-\delta_d \left|\left|IP_d\left(d_i\right) - IP_d\left(d_j\right)\right|\right|^2\right) \quad (8)$$

Here, $IP_d\left(d_l\right)$ denotes the vector at the l^{th} column of the matrix *OpdiM*, and δ_d is an adjustment coefficient for the kernel

bandwidth based on the new bandwidth parameter $\delta_d^{'}$, which is defined as follows:

$$\delta_d = \frac{\delta_d^{'}}{\frac{1}{Q}\sum_{k=1}^{Q}\left|\left|IP_d\left(d_k\right)\right|\right|^2} \quad (9)$$

Based on the above formula (8), it is easy to construct a weighted DDI network *WddiN*.

Construction of the Weighted PDI Network

In this section, through combining the weighted PPI network *WppiN* and original PDI network *OpdiN* with the weighted DDI network *WddiN*, we calculate two $O \times Q$ dimensional matrices *WpdiM* and *WdpiM* as follows:

$$WpdM\left(t_i, t_j\right) =$$

$$\begin{cases} WppiM\left(t_i, t_j\right): & if\ t_i \in N_P \text{ and } t_j \in N_P \\ WddiM\left(t_i, t_j\right): & else\ if\ t_i \in N_D \text{ and } t_j \in N_D \\ \frac{\sum_{k=1}^{O} WppiM(t_i,t_k)OpdiM\left(t_k,t_j\right)}{\sum_{k=1}^{O} WppiM(t_i,t_k)}: & else\ if\ t_i \in N_P \text{ and } t_j \in N_D \end{cases} \quad (10)$$

$$WdpM\left(t_i, t_j\right) =$$

$$\begin{cases} WppiM\left(t_i, t_j\right): & if\ t_i \in N_P \text{ and } t_j \in N_P \\ WddiM\left(t_i, t_j\right): & else\ if\ t_i \in N_D \text{ and } t_j \in N_D \\ \frac{\sum_{k=1}^{Q} OpdiM(t_i,t_k)WddiM\ (t_k,t_j)}{\sum_{k=1}^{Q} WddiM\left(t_k,t_j\right)}: & else\ if\ t_i \in N_P \text{ and } t_j \in N_D \end{cases} \quad (11)$$

Thereafter, for any two given nodes t_i and t_j in *OpdiN*, we can obtain a new $O \times Q$ dimensional matrix WPDIM as follows:

$$WpdiM\left(t_i, t_j\right) = \frac{WpdM\left(t_i, t_j\right) + WdpM\left(t_i, t_j\right)}{2} \quad (12)$$

According to the above formula (12), it is easy to construct a weighted PDI network *WpdiN*.

Calculation of the Initial Scores of Proteins

First, based on the weighted PDI network *WpdiN*, for a given protein p_i and a given domain d_j in N_{PD}, we can obtain a $Q \times O$ dimensional allocation probability matrix *APM* as follows:

$$APM\left(d_j, p_i\right) = \frac{WpdiM\left(d_j, p_i\right)}{\sum_{p_k \in d_j} WpdiM\left(d_j, p_k\right)} \quad (13)$$

Next, for simplicity, let the initial score vector for all domains in *WpdiN* be $S_d = <1, 1, 1>^T$, we assign an initial score of 1 to each domain in *WpdiN*, then based on the allocation matrix *APM*, we can distribute the initial scores of domains to all proteins in *WpdiN* in the following way:

$$PSD = APM^T * S_d \quad (14)$$

PSD is an *O* dimensional vector, and *PSD*(*i*) denotes the score, which is the i^{th} protein node p_i obtained from all domain nodes in *WpdiN*.

To calculate the score of the subcellular localization feature, let N_{SL} represent the number of all subcellular localizations, and $N_{SL}(j)$ denote the number of proteins related to the j^{th} subcellular localization. The *s_avg* means the average sum of the protein associated with subcellular localization. Then, the score of j^{th} subcellular localization can be computed as follows:

$$S_{SL}\left(j\right) = \frac{N_{SL}(j)}{S\_avg} \quad (15)$$

Where:

$$S\_avg = \frac{\sum_{j=1}^{N_{SL}} N_{SL}(j)}{N_{SL}} \quad (16)$$

Hence, for any given protein p_i, its subcellular localization feature score can be calculated as follows:

$$FS_{SL}\left(p_i\right) = \sum\nolimits_{j \in SL(p_i)} S_{SL}\left(j\right) \quad (17)$$

Where $SL(p_i)$ is the set of subcellular localization related to the protein p_i.

In addition, because triangles have the characteristic of stability, we further adopt the topological feature of triangles extracted from the *OppiN* to calculate at biological feature score for each protein p_i. Here, for a given protein p_i, its set of neighbor nodes is represented as $Np(p_i)$, then there is:

$$Np\left(p_i\right) = \left\{q \middle| ed(p_i,q) \in E_d\right\} \quad (18)$$

Therefore, the triangles for protein p_i is computed as follows:

$$TRI\left(p_i,q\right) = \begin{cases} \left|Np\left(p_i\right) \cap Np\left(q\right)\right| + 1 & \text{if } ed\left(p_i, q\right) \in E_P \\ 1 & \text{if } ed\left(p_i, q\right) \notin E_p \end{cases} \quad (19)$$

$$TRI\left(\mathrm{p_i}\right) = \sum\nolimits_{q \in Np(p_i)} TRI\left(p_i,q\right) \quad (20)$$

$$Avg_{TRI}\left(p_i\right) = \frac{TRI(p_i)}{|\mathrm{Np}(p_i)|} \quad (21)$$

Where the $TRI(p_i)$ is the set of triangles related to the protein p_i and $|\mathrm{Np}(p_i)|$ represents the degree of the protein p_i. According to the above calculated triangle numbers for each protein, we compute the triangle feature score for p_i;

$$FS_{TRI}\left(p_i\right) = \frac{Avg_{TRI}\left(p_i\right)}{\max_{1 \le j \le O} Avg_{TRI}\left(p_j\right)} \quad (22)$$

Based on the orthologous information obtained from the InPaianoid database (Gabriel et al., 2010), for any given protein p_i, let $f_{oth}(p_i)$ be its score of orthologous information, then we can calculate an orthologous feature score for p_i as follows:

$$FS_{ORT}\left(p_i\right) = \frac{f_{oth}\left(p_i\right)}{\max_{1 \le j \le O} f_{oth}\left(p_j\right)} \quad (23)$$

Based on the above formulas (14)∼(18), for any given protein p_i, we can obtain its feature score as follows:

$$FS\left(p_i\right) = \varphi * FS_{SL}\left(p_i\right) + \theta * FS_{TRI}\left(p_i\right) + \tau * FS_{ORT}\left(p_i\right) \quad (24)$$

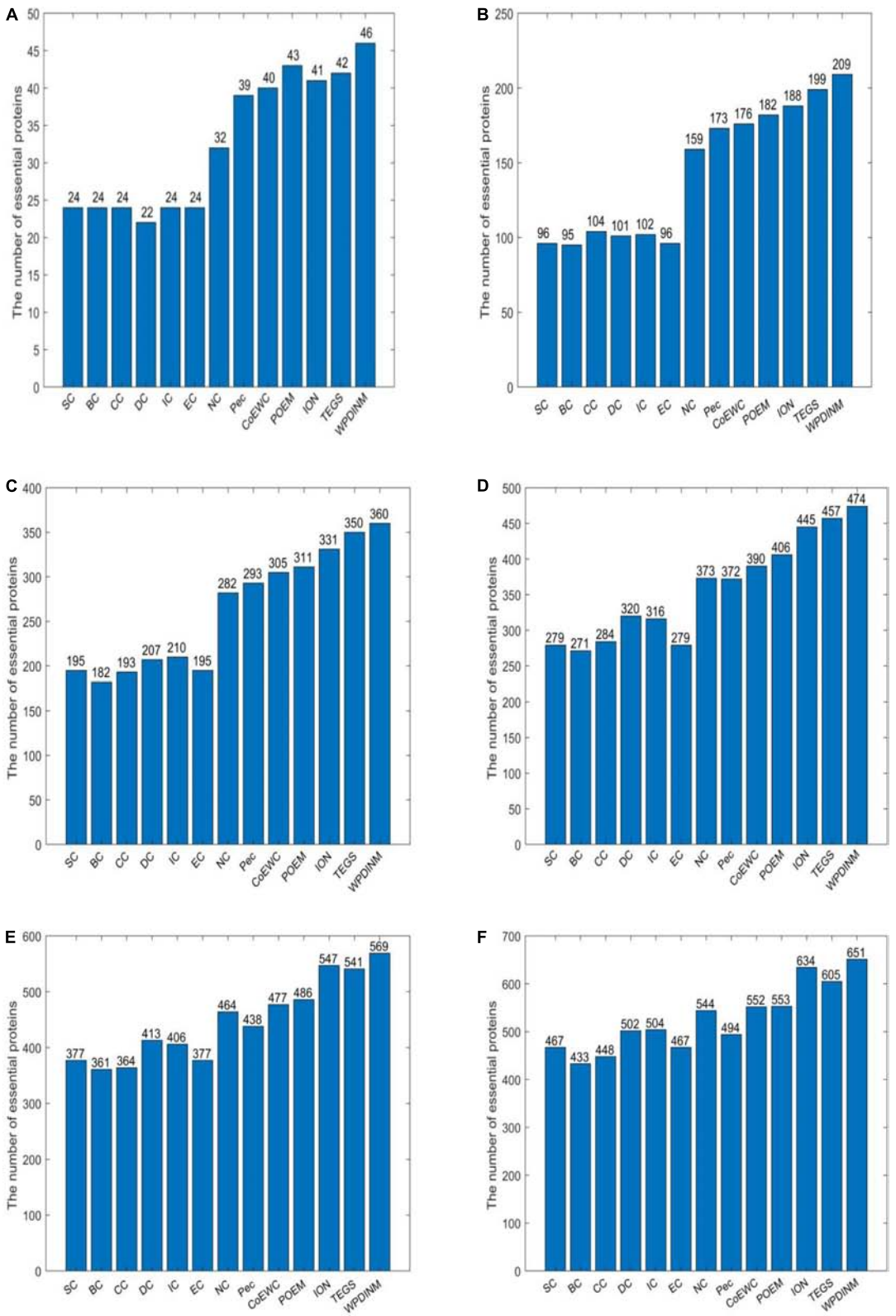

FIGURE 2 | (A) Top 1% ranked proteins, **(B)** Top 5% ranked proteins, **(C)** Top 10% ranked proteins, **(D)** Top 15% ranked proteins, **(E)** Top 20% ranked proteins, **(F)** Top 25% ranked proteins. This bar chart shows the comparison of the number of essential proteins predicted by WPDINM and other models, such as SC, BC, CC, DC, IC, EC, NC, Pec, CoEWC, POEM, ION, TEGS based on the DIP database.

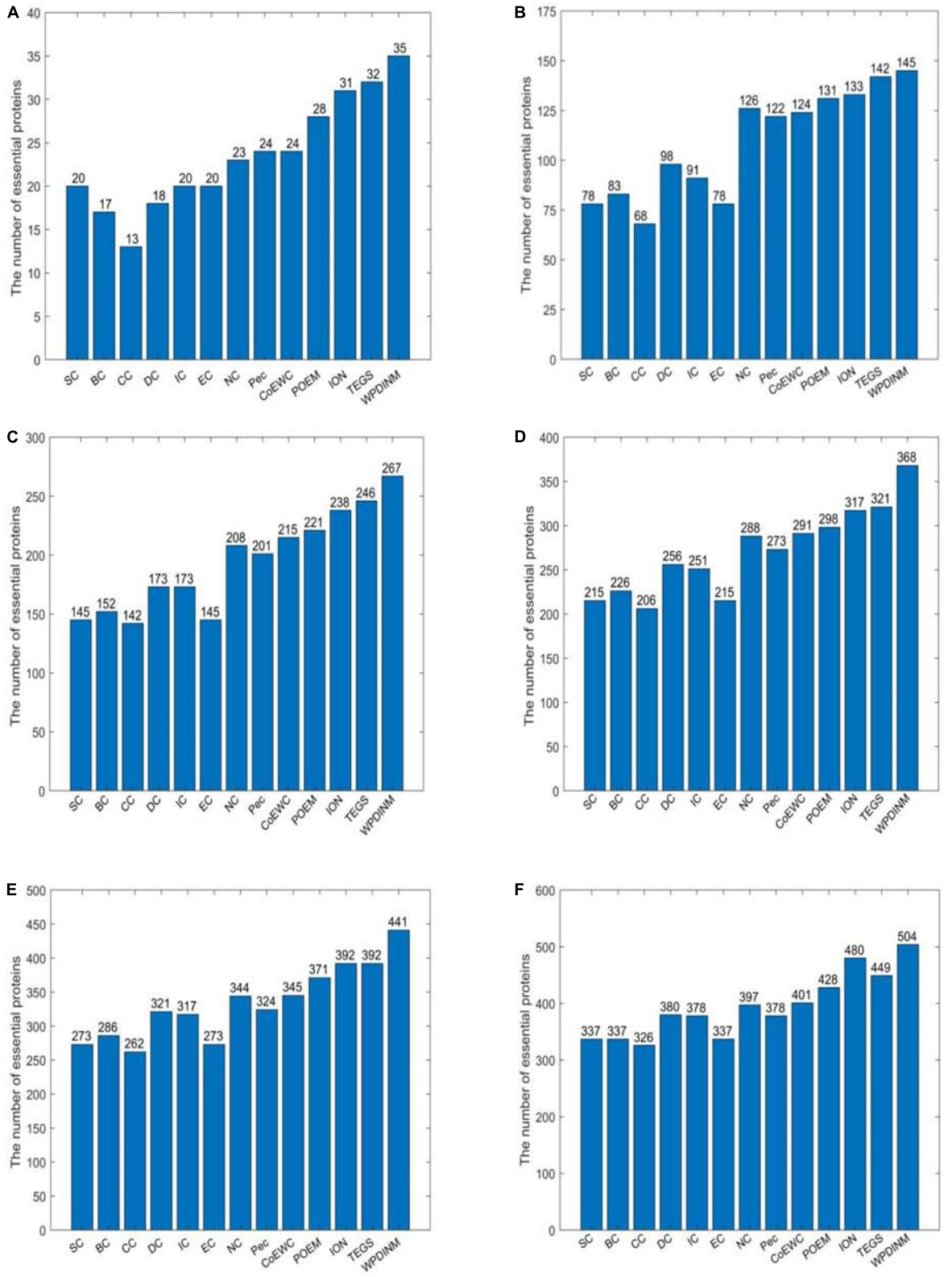

FIGURE 3 | (A) Top 1% ranked proteins, **(B)** Top 5% ranked proteins, **(C)** Top 10% ranked proteins, **(D)** Top 15% ranked proteins, **(E)** Top 20% ranked proteins, **(F)** Top 25% ranked proteins. This bar chart shows the comparison of the number of essential proteins predicted by WPDINM and other models, such as SC, BC, CC, DC, IC, EC, NC, Pec, CoEWC, POEM, ION, TEGS based on the Krogan database.

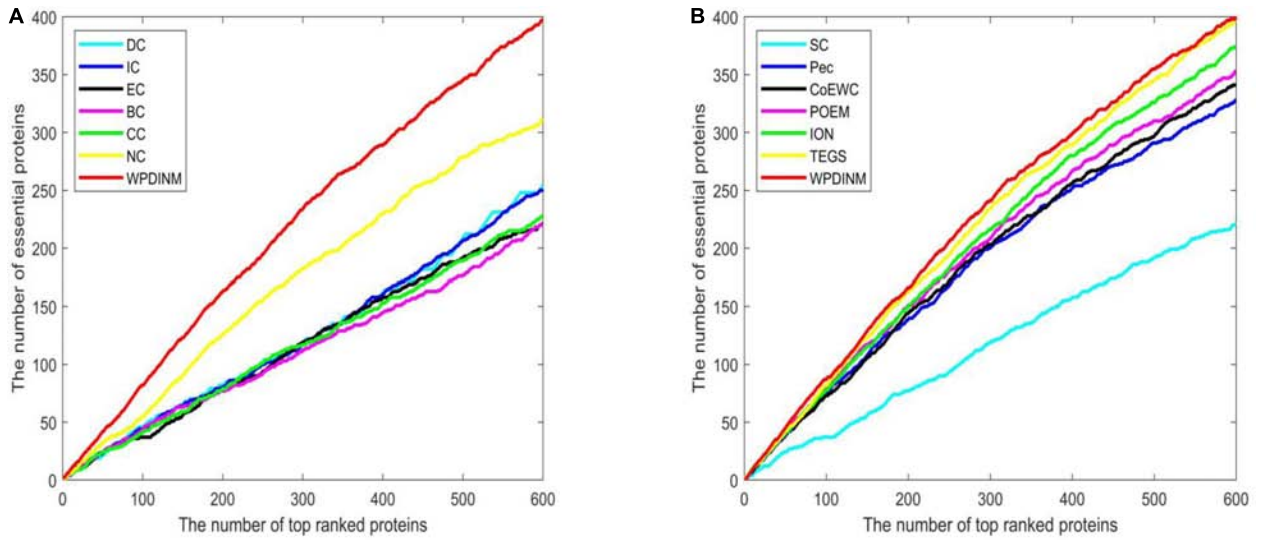

FIGURE 4 | Comparison of Jackknife curves of the WPDINM and 12 kinds of methods based on the DIP dataset. **(A)** Comparison of Jackknife curves of WPDINM and 6 other methods including DC, IC, EC, BC, CC, NC. **(B)** Comparison of Jackknife curves of WPDINM and 6 other measures such as SC, Pec, CoEWC, POEM, ION, TEGS.

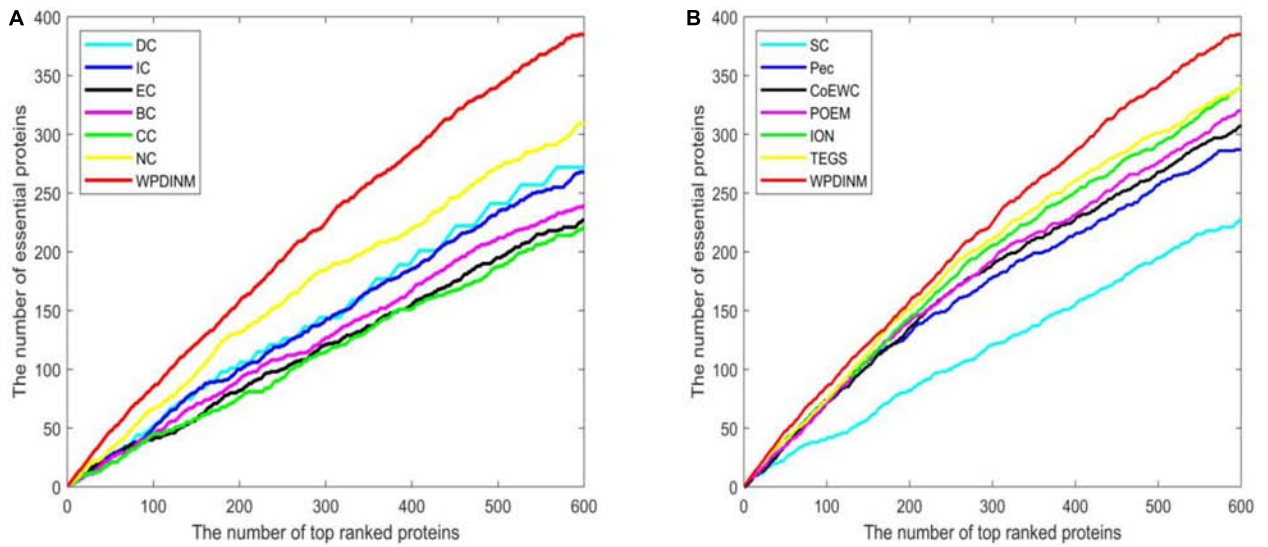

FIGURE 5 | Comparison of Jackknife curves of the WPDINM and 12 kinds of methods based on the Krogan database. **(A)** Comparison of Jackknife curves of WPDINM and 6 other methods including DC, IC, EC, BC, CC, and NC. **(B)** Comparison of Jackknife curves of WPDINM and 6 other measures such as SC, Pec, CoEWC, POEM, ION, TEGS.

Where φ,θ and τ are proportion parameters, which are used to adjust the ratio of feature score for proteins and satisfy $\varphi + \theta + \tau = 1$.

Finally, according to the above formula (13) and formula (19), for any given protein p_i, we can obtain its initial score as follows:

$$S_0\left(p_i\right) = \omega * PSD\left(p_i\right) + (1-\omega) * FS\left(p_i\right) \quad (25)$$

Here, ω is a proportion parameter.

Construction of the Prediction Model WPDINM

According to the weighted PPI network *WppiN*, let N_{p_i} and N_{p_j} be the sets of neighboring nodes of p_i and p_j, respectively, then $N_{p_i} \cap N_{p_j} = \{p_1, p_2, ..., p_T\}$ is the set of common neighbors of both p_i and p_j. Supposing that there is $WppiM\left(p_1, p_j\right) \leq WppiM\left(p_2, p_j\right) \leq ... \leq WppiM\left(p_T, p_j\right)$, then we define the allocation possibility of weight from p_i to p_j as follows:

$$WAP\left(p_i, p_j\right) = WppiM\left(p_i, p_T\right) * WppiM\left(p_T, p_j\right) \quad (26)$$

TABLE 1 | The common and differences between the WPDINM and the top 500 proteins detected by other methods (DIP).

Methods (Me)	|WPDINM ∩ Me|	|Me-WPDINM|	Percentage of essential proteins in {Me-WPDINM} (%)	|WPDINM-Me|	Percentage of essential proteins in{WPDINM-Me} (%)
DC	168	332	22.89	332	68.98
IC	170	330	23.94	330	68.79
EC	139	361	24.65	361	69.81
SC	139	361	24.65	361	69.81
BC	134	366	21.31	366	69.95
CC	133	367	24.25	367	69.48
NC	225	275	36.36	275	64
Pec	213	287	38.68	287	60.98
CoEWC	226	274	39.05	274	60.22
POEM	243	257	40.86	257	58.37
ION	307	193	41.97	193	56.99
TEGS	265	235	52.76	235	57.02

TABLE 2 | The common and differences between the WPDINM and the top 500 proteins detected by other methods (Krogan).

Methods (Me)	|WPDINM ∩ Me|	|Me-WPDINM|	Percentage of essential proteins in{Me-WPDINM} (%)	|WPDINM-Me|	Percentage of essential proteins in{WPDINM-Me} (%)
DC	195	305	32.13	305	64.92
IC	196	304	30.59	304	65.79
EC	166	334	25.75	334	69.46
SC	166	334	25.75	334	69.46
BC	169	331	31.12	331	70.09
CC	158	342	23.98	342	69.01
NC	217	283	38.52	283	62.89
Pec	193	307	34.85	307	62.21
CoEWC	199	301	37.21	301	61.46
POEM	211	289	39.10	289	61.59
ION	305	195	36.41	195	61.54
TEGS	226	274	44.52	274	59.12

Similarly, supposing that there is $WppiM\left(p_1, p_i\right) \leq WppiM\left(p_2, p_i\right) \leq \ldots \leq WppiM\left(p_T, p_i\right)$, then we define the allocation possibility of weight from p_j to $p_i$as follows:

$$WAP\left(p_j, p_i\right) = WppiM\left(p_j, p_T\right) * WppiM\left(p_T, p_i\right) \quad (27)$$

Hence, based on the above formulas, for any two given protein nodes p_i and p_j in $WppiN$, we can obtain an allocation possibility matrix of weights between them as follows:

$$WAPM\left(p_i, p_j\right) =$$

$$\begin{cases} \rho * WAP\left(p_i, p_j\right) / \sum_k WAP\left(p_i, p_k\right) : if \sum_k WAP\left(p_i, p_k\right) \neq 0 \\ 0 : \qquad \text{Otherwise} \end{cases} \quad (28)$$

Where ρ is the adjustment parameter with a value between 0 and 1.

Based on the above allocation possibility matrix $WAPM$, let a possibility vector S_{t+1} denote the vector of scores of proteins at the $(t+1)^{th}$ iteration, then we can calculate the proteins ranks iteratively as follows:

$$S_{t+1} = \mu * WAPM * S_t + (1-\mu)S_0 \quad (29)$$

Where $\mu \in (0, 1)$ is a scale parameter for adjusting the proportion of the current score vector S_t and initial score vector S_0.

Based on the above descriptions, the algorithm WPDINM can be briefly described as follows.

Algorithm 1:WPDINM

Input: domain data, Original PPI network, original protein-domain network, subcellular data, orthologous data, iterative error value ε, the proportion regulation parameters β, φ, θ, τ, ω, μ

Output: proteins score vector S

Step 1: Establishing weighted PPI network based on formulas (1–7);

Step 2: Establishing weighted domain-domain network based on formulas (8, 9);

Step 3: Establishing weighted protein-domain network based on formulas (10–12);

Step 4: Initializing proteins scores based on formulas (13–25);
Step 5: Establishing allocation network based on formulas (26–28);
Step 6: Calculating the S_{t+1}based on the formula (29),let $t = t+1$;
Step 7: Repeating Step 6 until there is $||S_{t+1}-S_t||^2 <\varepsilon$;
Step 8: Sorting the proteins scores for vector S_{t+1} through descending order.

RESULTS

Comparison of Twelve Essential Proteins Prediction Measures

The data presented by the bar chart illustrates that the identification performance of WPDINM exceeds the other measures by comparing the forecast accuracy from top 1% to top 25% proteins. It's apparent from **Figure 2** that, with the comparison of prediction accuracy in the top 1% proteins, 90.19% of the true key proteins are detected by the WPDINM method. By deferring the top 5% of proteins, the identification precision of WPDINM is up to 81.96%. The prediction result from the top 10% of proteins shows that the percentage of essential proteins identified by WPDINM is 70.58%. The prediction accuracies of WPDINM are 27.4, 19.6, 15.3, 13.2, 10.3, and 8.4% higher than the NC method from the top 1% to top 25%. By comparing it with the TEGS method, the precision of WPDINM increase by 3.6% from the top 25% of proteins.

Figure 3 shows the identification accuracy in the Krogan database. By observing the top 1% of proteins, the true essential proteins predicted by WPDINM make up 95%. With the top 5% proteins, 145 essential proteins detected. For the top 10% proteins, the proportion of essential proteins detected by the WPDINM is 5.7% observably higher than TEGS. For the top 15% and top 20% of proteins, the WPDINM can acquire 66.7% and 60% of the identification accuracy. In particular, in the top 25% candidate proteins, the prediction accuracy of WPDINM is, respectively, enhanced by 5.9% by comparing with TEGS. From what has been analyzed above, we can conclude that, whether in the DIP dataset or the Krogan database, the prediction performance of WPDINM is superior to these methods.

TABLE 3 | The AUCs of WPDINM and nine different methods in the Krogan database and DIP database.

Method	AUC (DIP)	AUC (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
ION	0.7522	0.7413
TEGS	0.7386	0.7287
WPDINM	0.7714	0.778

Validated by Jackknife Methodology

To further assess the prediction effect for WPDINM, the Jackknife Methodology is adopted to compare WPDINM with other methods. **Figure 4** shows the comparison results from the top 600 ranked proteins in the DIP dataset between the WPDINM method and other methods. As is revealed by **Figure 4A**, we can see that the WPDINM has more advantages than six prediction methods including IC, DC, CC, NC, BC, and EC. **Figure 4B** indicates that the performance of WPDINM exceeds the six methods: SC, Pec, CoEWC, POEM, ION, TEGS, respectively.

Figure 5 indicates the comparison result from the top 600 ranked proteins between the WPDINM and other measures in the Krogan dataset. From **Figure 5A**, it can be seen that the curve of WPDINM is above the curves of other competitive methods, containing DC, IC, EC, BC, CC, and NC. From **Figure 5B**, we can observe that the WPDINM is superior to the six methods including SC, Pec, CoEWC, POEM, ION, and TEGS.

Differences Between WPDINM and Other Methods

To compare the differences between WPDINM and other methods, we select the top 500 ranked proteins to compare the WPDINM with the 11 methods. The results of the comparison are shown in **Tables 1**, **2**. The Methods (Me) is a series of measures compared with the WPDINM methods. WPDINMMe is a set of proteins identified by the WPDINM and one of the Me. |WPDINMMe|denotes elements numbers of the WPDINMMe. WPDINM−Me represents the proteins detected by WPDINM except the proteins detected by both WPDINM and one of Me. Similarity, Me−WPDINM denotes the the proteins detected by one of Me except the proteins detected by both WPDINM and one of Me. |WPDINM−Me| and |Me−WPDINM| are the numbers of proteins in WPDINM−Me, Me−WPDINM, respectively. The data provided in **Table 1** shows the distinction between the WPDINM method and the eleven kinds of methods in the DIP dataset. It can be found from the second column of the table that the numbers of proteins identified by WPDINM and DC,IC, CC,BC,IC,EC are fewer than 200 proteins. In terms of the data for NC, the numbers of common proteins detected by both WPDINM and NC is just less than half. The proportions of overlapping proteins predicted by WPDINM and Pec, CoEWC, POEM are not more than half. **Table 2** reflects the differences of between WPDINM methods and other methods in the Krogan database. From **Table 2** we can see that the proportion of key proteins in {WPDINM-Me} is higher than one of the methods.

We further employ the receiver operating characteristic curve (ROC) and Precision-recall curve (PR) to test the prediction ability of the WPDINM model. The larger the area under the ROC curve (AUC), the better the prediction effect of the

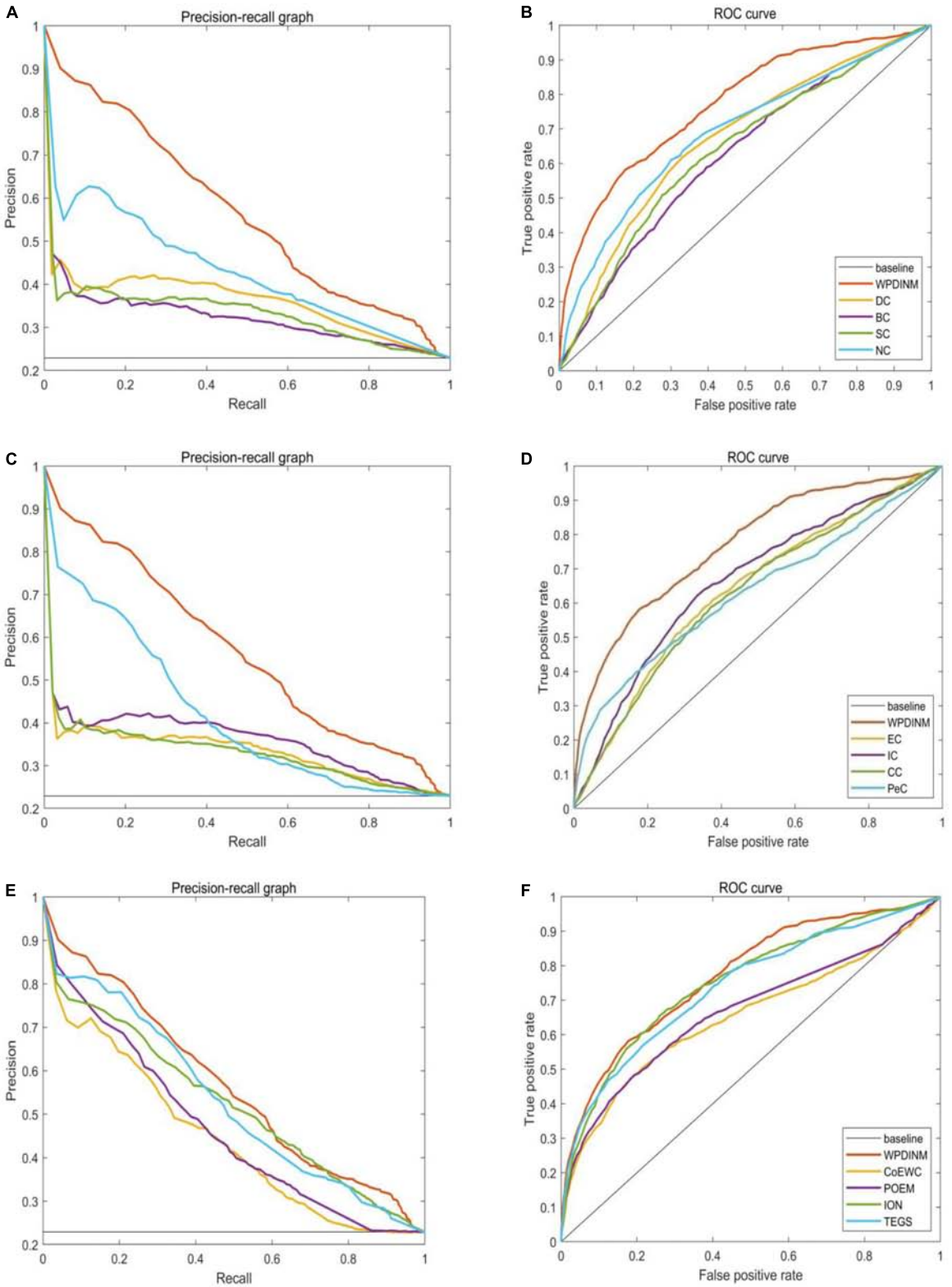

FIGURE 6 | Comparison of PR curves and ROC curves between WPDINM and the other competing methods based on the DIP database. **(A)** The PR curves of DC, BC, SC, NC. **(B)** The ROC curves of DC, BC, SC, NC. **(C)** The PR curves of EC, IC, CC, PeC. **(D)**The ROC curves of EC, IC, CC, PeC. **(E)** The PR curves of CoEWC, POEM, ION, TEGS. **(F)** The ROC curves of CoEWC, POEM, ION, TEGS.

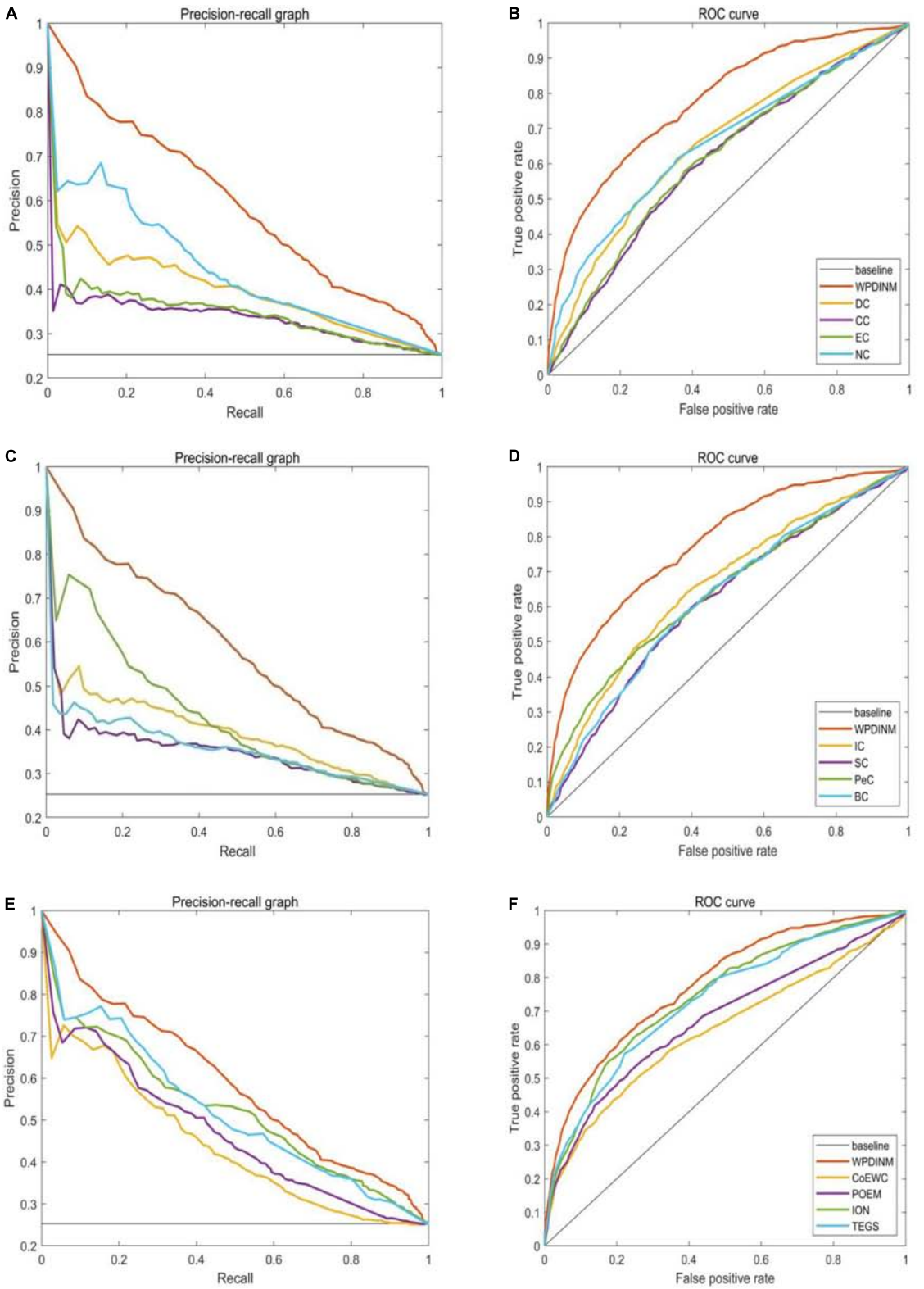

FIGURE 7 | Comparison of PR curves and ROC curves between WPDINM and the other competing methods based on the Krogan database. **(A)** The PR curves of DC, CC, EC, NC. **(B)** The ROC curves of DC, CC, EC, NC. **(C)** The PR curves of IC, SC, PeC, BC. **(D)** The ROC curves of IC, SC, PeC, BC. **(E)** The PR curves of CoEWC, POEM, ION, TEGS. **(F)** The ROC curves of CoEWC, POEM, ION, TEGS.

TABLE 4 | Number of essential proteins identified by WPDINM and 11 methods based on the GAVIN database.

Methods	Top 1% (19)	Top 5% (93)	Top 10% (196)	Top 15% (279)	Top 20% (371)	Top 25% (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
WPDINM	17	83	156	223	281	333

measure. The AUC data for all methods are collected in **Table 3**. **Figures 6**, **7** show the ROC curves and PR curves of the WPDINM method and various methods based on the DIP database and the Krogan database, respectively. As depicted in **Figure 6F**, although the ROC curves of WPDINM and ION have a little overlap, the AUC of WPDINM from **Table 3** is higher than the ION model. **Figure 7** shows that the ROC curve of WPDINM is higher than other competitive measures in the Krogan database.

As shown in **Table 4**, when comparing with the other 12 measures, the prediction accuracy of WPDINM is highest from top1% to top 25%. This reveals that the indication effect of the WPDINM model is better than 12 competing methods and that the WPDINM method has applicability to a large extent.

The Analysis of Parameters

Because the prediction precision needs to be enhanced, we set a proportions parameter $\mu \in (0, 1)$ in iterative formula (29). As is demonstrated in **Table 5**, we can see that in the DIP dataset, different values of parameter μ can have various influences on the experiment result. The statistics show the prediction accuracy in the top 1% to the top 25% proteins, when the parameter μ is set to a different value. It can be seen that the forecast accuracy slightly fluctuates, with the value of μ increasing. We repeat the same operation in the Krogan database. The data in **Table 6** presents the prediction performance from the Krogan database when parameter changing. Finally, because the prediction result is most competitive when the value of μ is 0.4, we choose to compare it with other methods.

For the sake of achieving higher prediction accuracy, we set a series of parameters. When calculating the weighted protein-protein network, we add two parameters β, γ to the computing formula (7). β and γ are adopted to regulate the ratio of two kinds of similarity between proteins. When the values of β and γ are set to 0.5, the WPDINM method obtains the best prediction effect. In formula (19), the parameters φ, θ and τ are employed to adjust the proportion of three features such as subcellular localization, orthologous information, and triangles features. The best experimental result is obtained by setting φ, θ and τ to 0.25, 0.35, and 0.45, respectively. Moreover, in formula (25), when the value of ω is set to 0.7, the WPDINM obtains the best performance.

TABLE 5 | Influence of the parameter μ on WPDINM's predication accuracy (DIP).

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rank									
Top 1%	46	46	46	46	46	46	46	46	46
Top 5%	212	211	210	209	209	209	209	210	210
Top 10%	361	360	361	360	359	359	359	359	357
Top 15%	472	471	470	474	476	476	478	478	479
Top 20%	566	568	568	569	568	569	570	570	569
Top 25%	651	651	651	651	651	652	652	652	649

TABLE 6 | Influence of the parameter μ on WPDINM's predication accuracy (Krogan).

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rank									
Top 1%	35	35	35	35	35	35	35	35	35
Top 5%	147	147	146	145	146	146	145	145	148
Top 10%	268	267	268	267	267	265	264	264	264
Top 15%	369	369	367	368	366	367	369	367	365
Top 20%	440	441	441	441	439	436	437	436	437
Top 25%	500	502	502	504	503	504	504	503	501

DISCUSSION

Essential proteins perform a crucial role in medicine and disease research, which deepen understanding of biological life processes. Accordingly, the prediction of key proteins has become a popular topic in recent years and deserves close attention. Recently, most computational models combining PPI networks and biological information are designed so that simple use of PPI data is unfavorable for achieving prediction accuracy.

The present study proposes a prediction algorithm to detect key proteins by integrating a PPI network and series of protein feature data. Firstly, we construct the weighted PPI network

based on the original PPI network and gene expression data processed by DFT. Next, the weighted domain-domain network is established based on the original protein-domain network. Then, by integrating the weighted domain-domain network with the weighted PPI network, the new weighted protein-domain network is further constructed. After that, we assign the initial scores for each protein by combining the topological feature and some biological information such as orthologous information, and subcellular information. Finally, we design a novel iteration algorithm based on the PageRank algorithm to compute protein scores iteratively. As a result, to testify the performance of the WPDINM algorithm, the WPDINM method is applied to three datasets including the DIP database, the Krogan database, and the Gavin database. The experimental result shows that the WPDINM achieves better indication than competitive methods.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

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

FUNDING

This research 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.2021.645932/full#supplementary-material

REFERENCES

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffithsjones, S., et al. (2004). The pfam protein families database. *Nucleic Acids Res.* 42, D222–D230.

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. Sociol.* 92, 1170–1182. doi: 10.2307/2780000

Chen, J., and Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* 22, 2283–2290. doi: 10.1093/bioinformatics/btl370

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

Ernesto, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* 71.(5 Pt 2):056103. doi: 10.1103/PhysRevE.71.056103

Gabriel, O., Thomas, S., Kristoffer, 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

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

Jop, M. P., Brock, A., Lngber, 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

Kim, W. (2012). Prediction of essential proteins using topological properties in go-pruned PPI network based on machine learning methods. *Tsinghua Technol.* 17, 645–658.

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

Li, G., Li, M., Wang, J. X., Wu, J., Wu, F. X., and Pan, Y. (2016). Predicting essential proteins based on subcellular localization, orthology and ppinetworks. *BMC Bioinform.* 17(Suppl 8):279. doi: 10.1186/s12859-016-1115-5

Li, M., Lu, Y., Wang, J. X., Wu, F. X., and Pan, Y. (2015). A topology potential-based method for identifying essential proteins from ppi networks. *IEEE ACM Trans. Comput. Biol. Bioinform.* 12, 372–383. doi: 10.1109/TCBB.2014.2361350

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

Lin, C. Y., Chin, C. H., Wu, H. H., Chen, S. H., Ho, C. W., and Ko, M. T. (2008). Hubba: hub objects analyzer-a framework of interactome hubs identification for network biology. *Nucleic Acids Res.* 36, 438–443.

Lu, X., Wang, X., Ding, L., Li, J., Gao, Y., and He, K. (2020). frDriver: a functional region driver identification for protein sequence. *IEEE ACM Trans. Computat. Biol. Bioinform.* 1–10. doi: 10.1109/TCBB.2020.3020096

Lu, X. G., Qian, X., Ling, X., Miao, Q. M., and Peng, S. L. (2019). Dmcm: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Luo, J. W., and Kuang, L. (2014). A new method for predicting essential proteins based on dynamic network topology and complex information. *Comput. Biol. Chem.* 52, 34–42. doi: 10.1016/j.compbiolchem.2014.08.022

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

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

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

Shang, X., Wang, Y., and Chen, B. (2016). Identifying essential proteins based on dynamic protein-protein interaction networks and rna-seq datasets. *Sci China Inf. Sci.* 59:070106. doi: 10.1007/s11432-016-5583-z

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

Tang, X. W., Wang, J. X., Zhong, J. C., 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., and Pan, Y. (2011). "A new method for identifying essential proteins based on edge clustering coefficient," in *Bioinformatics Research and Applications.* ISBRA 2011. Lecture Notes in Computer Science, Vol. 6674, eds J. Chen, J. Wang, and A. Zelikovsky (Berlin: Springer), 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, Sul-Min, 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

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. doi: 10.1093/nar/gkn858

Zhang, W., Xu, J., Li, Y., and Zou, X. (2018). 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, X., Wang, X., Acencio, M. L., Lemke, L., and Wang, X. (2016). An ensemble framework for identifying essential proteins. *BMC Bioinform.* 17:322. doi: 10.1186/s12859-016-1166-7

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

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. NanoBioscience* 13, 415–424. doi: 10.1109/TNB.2014.2337912

Zhong, J., Wang, J. X., Peng, W., Zhang, Z., and Pan, Y. (2013). Prediction of essential proteins based on gene expression programming. *BMC Genomics* 14 Suppl 4(Suppl 4):S7. doi: 10.1186/1471-2164-14-S4-S7

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.

The handling editor QZ declared a past co-authorship/collaboration with one of the authors LW.