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

Zixuan Meng¹, Linai Kuang¹*, Zhiping Chen², Zhen Zhang², Yihong Tan², Xueyong Li² and Lei Wang¹,²*

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

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. Zhang 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 COMPARTMENTS 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 \( \text{OppiN} = \{N_p, E_p\} \) represent a newly constructed original PPI network, where \( N_p = \{p_1, p_2, ..., p_{N_p}\} \) is the set of protein nodes in \( \text{OppiN} \) and \( E_p \) is the set of edges between protein nodes in \( \text{OppiN} \). Here, for two given proteins \( p_i \) and \( p_j \) in \( N_p \), there is an edge \( e_d(p_i, p_j) \) between them in \( E_p \), if and only if there is a known interaction between these two proteins. Based on the original PPI network \( \text{OppiN} \), we can further obtain an \( O \times O \) dimensional adjacency matrix \( \text{OppiM} \) as follows: For any two protein \( p_i \) and \( p_j \) in \( N_p \), there is \( \text{OppiM}(i,j) = 1 \), only if there is a known interaction between \( p_i \) and \( p_j \), otherwise there is \( \text{OppiM}(i,j) = 0 \).

Next, based on \( \text{OppiN} \), for any given protein \( p \) with a known gene expression sequence in \( N_p \), let \( \text{Gep}(p) = <\text{Gep}(p,1), \text{Gep}(p,2),..., \text{Gep}(p,M)> \) represent the gene expression sequence of \( p \), where \( \text{Gep}(p,t) \) is the degree of gene expression at \( t^{th} \) time. As \( \text{Gep}(p) \) 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 \( \text{Gep}(p) \) to a \( 1 \times N \) dimensional spectrum vector \( \text{DF}(p) = <|\text{DGep}(0)|, |\text{DGep}(1)|, ..., |\text{DGep}(N-1)|> \) as follows:

\[
\text{DGep}(t) = \sum_{y=1}^{M} \text{Gep}(p,y) \times W_M^y, \text{where } t = 0, 1, ..., N-1
\]

\[
W_M = e^{-i \frac{2\pi t}{M}}
\]

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:

\[
\text{GK}(p_i, p_j) = \exp\left(-\alpha_p \frac{||\text{DF}(p_i) - \text{DF}(p_j)||^2}{\text{P}_\text{Gep}}\right)
\]

Here, \( \alpha_p \) is the adjustment coefficient for the kernel bandwidth, which is defined as follows:

\[
\alpha_p = \frac{1}{|\text{P}_\text{Gep}|} \sum_{k=1}^{|\text{P}_\text{Gep}|} \frac{\alpha'}{||\text{DF}(p_k)||^2}
\]

In the above formula (4), \( \text{P}_\text{Gep} \) is the set of proteins with gene expression sequences in \( \text{OppiN} \).
Additionally, for any two given proteins $p_i$ and $p_j$ without gene expression sequences in $\text{OppiN}$, we adopt the topological features extracted from the original PPI network $\text{OppiN}$ to calculate the possibility of an association between them. Thus, the weight between $p_i$ and $p_j$ can be calculated as follows:

$$
\text{TFP} \left( p_i, p_j \right) = \frac{|\text{Com} \left( p_i, p_j \right)| + 1}{\left( |N_p(p_i)| + 1 \right) \ast \left( |N_p(p_j)| + 1 \right)}
$$

(5)

Here, $N_p(p_i)$ and $N_p(p_j)$ denote the set of neighboring protein nodes of $p_i$ and $p_j$ in $\text{OppiN}$ separately, $\text{Com}(p_i, p_j)$ represents the set of common neighbors between $p_i$ and $p_j$ in $\text{OppiN}$, 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 $\text{OppiN}$, we can calculate the possibility of an association between them, $b$, as follows:

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

(6)

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

$$
\text{WppiM} \left( p_i, p_j \right) = 
\begin{cases} 
\beta \ast \text{PA} \left( p_i, p_j \right) + (1-\beta) \ast \text{OppiM} \left( p_i, p_j \right) & \text{if } \text{OppiM} \left( p_i, p_j \right) = 1 \\
\text{TFP} \left( p_i, p_j \right) \ast \text{GK} \left( p_i, p_j \right) & \text{otherwise} 
\end{cases}
$$

(7)

In the above formula (7), $\beta$ 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 $\text{OpdiN} = \{N_{PD}, E_{PD}\}$, where $N_{PD} = N_P \cup N_D$, $N_D = \{d_1, d_2, \ldots, d_l\}$ is the set of domain nodes in $\text{OpdiN}$, 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 node $p_i$ 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 $\text{OpdiN}$, we can further obtain an $O \times Q$ dimensional adjacency matrix $\text{OpdiM}$ as follows: for a given protein node $p_i$ and a given domain node $d_j$ in $N_{PD}$, there is $\text{OpdiM}(i, j) = 1$, if and only if there is $p_i$ belonging to $d_j$, otherwise, there is $\text{OpdiM}(i, j) = 0$.

**Construction of the Weighted DDI Network**

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

$$
\text{WddiM} \left( d_i, d_j \right) = \exp \left( -\delta_d \ast \| IP_d (d_i) - IP_d (d_j) \|^2 \right)
$$

(8)

Here, $IP_d(d_i)$ denotes the vector at the $i$th column of the matrix $\text{OpdiM}$, and $\delta_d$ is an adjustment coefficient for the kernel.
Construction of the Weighted PDI Network

In this section, through combining the weighted PPI network WpdiN and original PDI network OpdiN with the weighted DDI network WddiN, we can calculate two O × Q dimensional matrices WpdiM and WdpM as follows:

\[
WpdiM\left( t_i, t_j \right) = \left\{ \begin{array}{ll}
WppiM\left( t_i, t_j \right) & \text{if } t_i \in N_p \text{ and } t_j \in N_p \\
WddiM\left( t_i, t_j \right) & \text{else if } t_i \in N_D \text{ and } t_j \in N_D \\
\sum_{k=1}^{Q} O_{pdi}(t_i, t_j) WddiM(t_i, t_k) & \text{else if } t_i \in N_p \text{ and } t_j \in N_D \\
\sum_{k=1}^{Q} WddiM(t_i, t_k) & \text{else if } t_i \in N_D \text{ and } t_j \in N_p \\
\sum_{k=1}^{Q} WddiM(t_i, t_k) WddiM(t_k, t_j) & \text{else if } t_i \in N_D \text{ and } t_j \in N_D
\end{array} \right.
\]

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 NpD, we can obtain a Q × O dimensional allocation probability matrix APM as follows:

\[
APM\left( d_j, p_i \right) = \frac{WpdiM\left( d_j, p_i \right)}{\sum_{p_k \in d_j} WpdiM\left( d_j, p_k \right)}
\]

Next, for simplicity, let the initial score vector for all domains in WpdiN be \( S_d = \begin{bmatrix} 1, 1, 1 \end{bmatrix}^T \). 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:

\[
PSD = APM^T * S_d
\]

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:

\[
S_{SL}(j) = \frac{N_{SL}(j)}{S_{avg}}
\]

Where:

\[
S_{avg} = \sum_{j=1}^{N_{SL}} \frac{N_{SL}(j)}{N_{SL}}
\]

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

\[
FS_{SL}(p_i) = \sum_{j \in SL(p_i)} S_{SL}(j)
\]

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 OpdiN 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:

\[
Np(p_i) = \{ q \mid ed(p_i, q) \in E_d \}
\]

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

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

\[
TRI(p_i) = \sum_{q \in Np(p_i)} TRI(p_i, q)
\]

\[
Avg_{TRI}(p_i) = \frac{TRI(p_i)}{|Np(p_i)|}
\]

Where the \( TRI(p_i) \) is the set of triangles related to the protein \( p_i \) and \( |Np(p_i)| \) 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 \( p_i \):

\[
FS_{TRI}(p_i) = \frac{Avg_{TRI}(p_i)}{\max_{1 \leq j \leq O} Avg_{TRI}(p_j)}
\]

Based on the orthologous information obtained from the InParanoid 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:

\[
FS_{ORT}(p_i) = \frac{f_{oth}(p_i)}{\max_{1 \leq j \leq O} f_{oth}(p_j)}
\]

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

\[
FS(p_i) = \varphi * FS_{SL}(p_i) + \theta * FS_{TRI}(p_i) + \tau * FS_{ORT}(p_i)
\]
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.
Where $\varphi, \theta$ and $\tau$ 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:

$$S_0(p_i) = \omega * PSD(p_i) + (1-\omega) * FS(p_i)$$  \hspace{1cm} (25)$$

Here, $\omega$ 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 $WppiM(p_i, p_j) \leq WppiM(p_2, p_j) \leq ... \leq WppiM(p_T, p_j)$, then we define the allocation possibility of weight from $p_i$ to $p_j$ as follows:

$$WAP(p_i, p_j) = WppiM(p_i, p_T) * WppiM(p_T, p_j)$$  \hspace{1cm} (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        | 40.66                                            | 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
\[ WppiM(p_1, p_i) \leq WppiM(p_2, p_i) \leq ... \leq WppiM(p_T, p_i) \]
then we define the allocation possibility of weight from \( p_j \) to \( p_i \) as follows:

\[
WAP(p_j, p_i) = WppiM(p_j, p_T) \times WppiM(p_T, p_i) \tag{27}
\]

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:

\[
WAPM(p_i, p_j) = \begin{cases} 
\rho \times WAP(p_i, p_j) / \sum_k WAP(p_i, p_k) & : \text{if } \sum_k WAP(p_i, p_k) \neq 0 \\
0 & : \text{Otherwise} 
\end{cases} \tag{28}
\]

Where \( \rho \) 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:

\[
S_{t+1} = \mu \times WAPM \times S_t + (1-\mu)S_0 \tag{29}
\]

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 \( \varepsilon \), the proportion regulation parameters \( \beta, \psi, \theta, \tau, \omega, \mu \).

**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 TEGS. For the top 5% proteins, 145 essential proteins detected. For the top 10% proteins predicted by WPDINM make up 95%. With 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.

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, I, 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

| 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 | 0.6683 |
| IC | 0.6657 | 0.6673 |
| EC | 0.6384 | 0.6167 |
| BC | 0.625 | 0.6248 |
| SC | 0.6384 | 0.6167 |
| CC | 0.6291 | 0.6114 |
| NC | 0.6879 | 0.6684 |
| 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 |
FIGURE 6 | Comparison of PR curves and ROC curves between WP DINM 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 CoEWIC, POEM, ION, TEGS. (F) The ROC curves of CoEWIC, 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           |

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 $\mu$ 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 $\mu$ is set to a different value. It can be seen that the forecast accuracy slightly fluctuates, with the value of $\mu$ 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 $\mu$ 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 $\beta, \gamma$ to the computing formula (7). $\beta$ and $\gamma$ are adopted to regulate the ratio of two kinds of similarity between proteins. When the values of $\beta$ and $\gamma$ are set to 0.5, the WPDINM method obtains the best prediction effect. In formula (19), the parameters $\varphi, \theta$ and $\tau$ 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 $\varphi, \theta$ and $\tau$ to 0.25, 0.35, and 0.45, respectively. Moreover, in formula (25), when the value of $\omega$ is set to 0.7, the WPDINM obtains the best performance.

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.

REFERENCES

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32:D214–D220.

Binder, J. X., Sune, P. F., Kalliopi, T., Christian, S., O’Donoghue Seán, J., Reinhard, S., et al. (2014). Departments: 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., Chernvitz, 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., Sanjitt, 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 biological networks. *Nature* 406, 43–47. doi: 10.1038/35075138

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

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

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

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 ppi networks. *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: hubs 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). iDriver: 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

Miews, 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/gkl418

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).
Meng et al. (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/gkn588

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.

Copyright © 2021 Meng, Kuang, Chen, Zhang, Tan, Li 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.