IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 4, 2020, accepted December 17, 2020, date of publication December 21, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046254

An Iterative Method for Identifying Essential Proteins Based on Non-Negative Matrix Factorization

JIN LIU, XIANGYI WANG, ZHIPING CHEN, YIHONG TAN, XUEYONG LI, ZHEN ZHANG, AND LEI WANG

College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China

Corresponding author: Lei Wang (wanglei@xtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873221, in part by the Research Foundation of Education Bureau of Hunan Province under Grant 20B080, and in part by the Natural Science Foundation of Hunan Province under Grant 2018JJ4058 and Grant 2019JJ70010.

ABSTRACT In recent years, with the development of high-throughput technologies, lots of computational methods for predicting essential proteins based on protein-protein interaction (PPI) networks and biological information of proteins have been proposed successively. However, due to the incompleteness of PPI networks, the prediction accuracy achieved by these methods is still unsatisfactory, and it remains to be a challenging work to design effective computational models to identify essential proteins. In this manuscript, a novel Prediction Model based on the Non-negative Matrix Factorization (PMNMF for abbreviation) is proposed. In PMNMF, an original PPI network will be constructed first based on PPIs downloaded from any given benchmark database. And then, based on topological features of protein nodes, the original PPI network will be further converted to a weighted PPI network. Moreover, in order to overcome the incompleteness of PPI networks, the NMF (Non-negative Matrix Factorization) method will be implemented on the weighted PPI network to obtain a transition probability matrix. And then, by integrating biological information including the gene expression information, homologous information and subcellular localization information of proteins, a unique initial score will be calculated and assigned to each protein node in the weighed PPI network, based on which, an improved Page-Rank algorithm will be designed to infer potential essential proteins. Finally, in order to evaluate the performance of PMNMF, it will be compared with 14 state-of-the-art prediction models, and experimental results show that PMNMF can achieve the best identification accuracy.

INDEX TERMS Essential protein prediction, iteration method, non-negative matrix factorization.

I. INTRODUCTION

Essential proteins are found in large numbers in protein complexes, and their absence will lead to the loss of functions of related protein complexes, and make it impossible for organisms to survive or develop. Identifying essential proteins is important for the understanding of the process of cell growth and regulation, and can provide valuable information to the researches of disease analysis and drug design etc. In recent years, with the rapid development of high-throughput techniques, more and more protein-protein interactions (PPIs) have been detected successively, based on which, PPI networks are established and applied widely in designing computational models for inferring essential proteins. For instance, based on the topological characteristic of centrality [1], [2] of PPI networks, a series of calculation models including CC(Closeness Centrality) [3], DC(Degree Centrality) [4], BC(Betweenness Centrality) [5], SC[Subgraph Centrality] [6], NC(Neighbor Centrality) [7] have been proposed to discover basic proteins. Besides, Li M *et al* [8] designed an identification model named LAC to identify key proteins based on the Local Average Connectivity of protein nodes in PPI networks [9]. Qi Yi *et al* [10] designed a prediction model to infer basic proteins based on the Local Interaction Density (LID) of protein nodes in PPI networks. Chen B *et al* [11] proposed an essential protein identification method based on multiple topological structures of PPI networks. In all these methods mentioned

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

above, it is only considerate the topological properties of PPI networks, thus, due to the incompleteness of current PPI networks, the prediction accuracy of these methods is still not satisfactory. Then in order to improve the prediction accuracy of computational models, some new identification models have been proposed for the past few years by combining the topological characteristics of PPI networks and the biological information of proteins. For example, through integrating PPI networks with the gene expression data of proteins, M Li *et al* [12] and Xiwei Tang *et al* [13] proposed two prediction models called Pec and WDC respectively. W Peng *et al* designed one prediction model based on the orthologous information of proteins and PPI networks [14], and another prediction model based on the domain information of proteins and PPI networks [15] to infer essential proteins respectively. X Zhang *et al* [16] introduced an identification method called CoEWC by combining topological features of PPI networks with the co-expression properties of proteins. BH Zhao *et al* [17] designed a prediction model called POEM by integrating gene expression data of proteins with topological features of PPI networks. J Luo *et al* [18] put forward a computational method for essential protein prediction based on the local interaction density of PPI networks and biological features of protein complexes. Seketoulie Keretsu *et al* [19] presented an identification model of protein complexes based on weighted edge by clustering and the gene expression profiles of proteins. M Li *et al* [20], [21] proposed two necessary protein identification methods by integrating PPI networks with subcellular localization information and complex centrality of proteins separately. J Luo *et al* [22] introduced a method to detect essential proteins based on protein complex co-expression data and ECC (edge clustering coefficient) of PPI networks. Bihai Zhao *et al* proposed a model based on Multiplex Biological Networks [23] and a model based on Diffusion Distance Networks [24] to predict essential proteins separately. S. Li *et al* [25] proposed one iteration method called CVIM to predict essential protein, based on topological and functional features. Lei X *et al* presented one essential protein prediction methods called AFSOEP [26] to infer protein complexes by AFSO (Artificial Fish Swarm Optimization). Bihai Zhao et al [27] designed an iterative method for identifying potential key proteins from heterogeneous PPI networks. Dai W *et al* [28] proposed a method to discover essential genes based on protein-protein interaction network embedding. Fengyu Zhang *et al* [29] introduced a model called FDP to predict essential Genes by fusing dynamic PPI networks. Chen Z *et al* [30] proposed a prediction model called NPRI based on one heterogeneous network, the heterogeneous Protein-Domain network are established in accordance with initial PPI network, Protein-Domain network and gene expression data. All these above mentioned methods have demonstrated that it can improve the prediction accuracy of calculative models by combining the biological information of proteins with the topological features of PPI networks.

In general, these existing essential protein prediction methods are mainly designed by combining the topological characteristics of PPI networks with biological features of proteins. However, due to the incompleteness of PPI networks, the prediction accuracy of these methods is still not very satisfactory. Hence, inspired by the ideas of existing state-of-the-art models, in this paper, a novel prediction model called PMNMF is designed to infer essential proteins. In PMNMF, an original PPI network will be constructed first based on known PPIs downloaded from benchmark databases. And then, based on topological features of protein nodes, the original PPI network will be transformed to a weighted PPI network. Next, through adopting the Non-negative Matrix factorization (NMF) method, a transition probability matrix will be obtained. Finally, by combining the gene expression information, orthologous information and subcellular localization information of proteins, an iterative algorithm will be designed and implemented on the weighted PPI network to detect potential essential proteins. Moreover, in order to evaluate the performance of PMNMF, it will be compared with some competitive methods. Experimental results show that PMNMF can achieve reliable identification accuracies of 98.04%, 85.10%, 69.74%, 60.10%, 55.05% and 51.22% in top 1%, top 5%, top10%, top15%, top20% and top 25% of predicted potential essential proteins respectively, which predictive performance is better than all these state-of-the-art competing models.

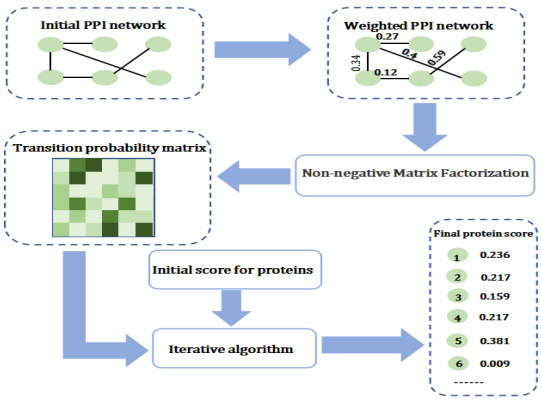


FIGURE 1. Flowchart of PMNMF.

II. MATHOD

As shown in Fig.1, the process of PMNMF consists of the following 3 main steps:

Step 1: First, based on the dataset of known PPIs downloaded from any given benchmark database, an original PPI network will be constructed. And then, based on the topological features of protein nodes, the original PPI network will be further converted to a weighted PPI network.

Step 2: Next, based on the gene expression formation, homologous information and subcellular location

information of proteins, a unique initial score will be calculated and assigned to each protein node in the weighted PPI network.

Step 3: Finally, based on the initial scores of proteins and the transition probability matrix obtained by adopting the non-negative matrix factorization method, an improved page-rank algorithm will be designed to calculate a final score for each protein, which can be utilized to evaluate the essentiality of the protein effectively.

A. CONSTRUCTION OF THE WEIGHTED PPI NETWORK

Let Ψ denote the dataset of known PPIs downloaded from any given benchmark database, $N_P = \{p_1, p_2 \ldots .p_O\}$ be the set of all these different proteins in Ψ. For any two given proteins p_i and p_j in N_P, we define that there is an edge $e(p_i, p_j)$ between them, if and only if there is a known interaction between p_i and p_j in Ψ. And for convenience, let E_P represent the set consisting of all these edges between proteins in Ψ. Then, it is apparent that we can obtain an original PPI network $OppiN = \{N_P, E_P\}$, and based on which, we can further obtain an O×O dimensional adjacency matrix $OppiM$ as follows: for any two given proteins p_i and p_j in N_P, there is $OppiM\ (i, j) = 1$, if and only if there is a known interaction between them in Ψ, otherwise there is $OppiM\ (i, j) = 0$. In addition, let $NB(p_i)$ represent the set of nodes neighboring to p_i in $OppiN$, i.e., there are edges between these nodes and p_i in $OppiN$. Let $|NB\ (p_i)\ |$ denote the number of different nodes in $NB(p_i)$, $NB(p_i) \cap NB(p_j)$ be the set of nodes neighboring to both p_i and p_j in $OppiN$, and $|NB\ (p_i) \cap NB\ (p_j)\ |$ represent the number of different nodes in $NB(p_i) \cap NB(p_j)$, then based on the assumption that for any two given proteins, if they interact with one or more other common proteins at the same time, the interaction between these two proteins will be more reliable [31], we can define the Edge Aggregation Coefficient between p_i and p_j as follows:

$$EAC\ (p_i, p_j) = \begin{cases} \dfrac{\left|NB\ (p_i) \cap NB\ (p_j) + 1\right|}{\min\left(\left|NB\ (p_i)\right|, \left|NB\ (p_j)\right|\right)} : & if\ OppiM\ (p_i, p_j) = 1 \\ 0 : & else \end{cases} \tag{1}$$

From observing above formula (1), it is easy to see that, for any two given proteins p_i and p_j in N_P, the more common neighboring nodes between them, the bigger the value of $EAC\ (p_i, p_j)$ will be. Hence, to some degree, the Edge Aggregation Coefficient between p_i and p_j can reflect the degree of interaction between them effectively.

Moreover, it is reasonable to assume that if a protein node has known interactions with more proteins, then it will be more reliable. Hence, for any given protein p_i in N_P, let $ENB\ (p_i)$ denote the number of known interactions between it and all the other proteins in N_P, then we can define the Point Aggregation Coefficient of p_i as follows:

$$PAC\ (p_i) = \frac{ENB\ (p_i)}{\frac{|NB(p_i)|*(|NB(p_i)|-1)}{2}} \tag{2}$$

Based on above two formulas, for any two given proteins p_i and p_j in N_P, it is reasonable to assume that the potential interaction between them varies directly with both the value of the Edge Aggregation Coefficient between them and the values of their Point Aggregation Coefficients. Hence, we can define the Degree of Potential Interaction between p_i and p_j as follows:

$$DPI\ (p_i, p_j) = EAC\ (p_i, p_j) * \left(PAC\ (p_i) + PAC\ (p_j)\right) \tag{3}$$

Obviously, based on above formula (3), an $O \times O$ dimensional interaction matrix DPI can be obtained. However, through considering the limited number of known interactions between proteins and the definition of the Edge Aggregation Coefficient between proteins illustrated in above formula (1), it is easy to know that DPI will be a sparse matrix. Hence, we can adopt the Non-negative Matrix Factorization (NMF) method [32-34] to predict unknown weights, convert it to the product of two non-negative matrixes $W \in R^{O\times k}$ and $H \in R^{k\times O} (k \ll O)$ as follows:

$$DPI^* = WH^T \tag{4}$$

Here, the matrixes W and H satisfy the following target function:

$$TF = \min_{W,H} \left|\left|DPI - WH^T\right|\right|_E^2 \quad s.t.\ W \geq 0\ and\ H \geq 0 \tag{5}$$

Here, $||.||_E$ represents the Euclid paradigms.

From observing above two formulas, it is easy to see that NMF aims to find two non-negative matrixes W and H, whose product WH can provide the optimal approximation to the original matrix DPI. As for above target function illustrated in formula (5), by adopting the iterative update algorithm proposed by Lee *et al* [35], the matrixes W and H can be iteratively obtained according to the following formulas:

$$W_{ik} \leftarrow W_{ik} \times \frac{\left(DPI * H^T\right)_{ik}}{\left(WHH^T\right)_{ik}} \tag{6}$$

$$H_{ki} \leftarrow H_{ki} \times \frac{\left(W^T * DPI\right)_{ki}}{\left(W^T WH\right)_{ki}} \tag{7}$$

B. CALCULATION OF INITIAL SCORES FOR PROTEINs

In this section, we will combine the gene expression information, subcellular localization information and homologous information of proteins to calculate a unique initial score for each protein in the weighted PPI network as follows:

Firstly, for any given protein p_i, let $GE(p_i) = \{GE(p_i, 1), GE(p_i, 2), \ldots, GE(p_i, n)\}$ denote the gene expression data of p_i at n different time points, where $GE(p_i, t)$ represent the level of gene expression of p_i at the time point t. Then, based on the method of PCC (Pearson Correlation Coefficient) [36], we can calculate a PCC-based initial score for p_i as follows:

$$GScore\ (p_i) = \frac{gscore\ (p_i) - \min_{1\leq j\leq n}\{gscore(p_j)\}}{\max_{1\leq j\leq n}\{gscore(p_j)\} - \min_{1\leq j\leq n}\{gscore(p_j)\}} \tag{8}$$

Here,

$$gscore\,(p_i) = \sum_{p_l \in NB(p_i)} PCC\,(p_i, p_l) \tag{9}$$

$$PCC\,(p_i, p_j) = \frac{1}{n-1} \sum_{t=1}^{n} \left(\frac{GE\,(p_i, t) - \overline{GE\,(p_i)}}{\sigma\,(p_i)} \right) \times \left(\frac{GE\,(p_j, t) - \overline{GE\,(p_j)}}{\sigma\,(p_j)} \right) \tag{10}$$

Here, $\overline{GE(p_i)}$ denotes the average expression level of p_i at all these n time points, $\sigma\,(p_i)$ is the standard variance of gene expression levels of p_i at all these n time points, and $PCC\,(p_i, p_j)$ represents the Pearson Correlation Coefficient between p_i and p_j.

Next, based on the homologous information of proteins, for any given protein p_i, let $O\,(p_i)$ denote the homologous information of p_i, then we can obtain another homologous information based initial score for p_i as follows:

$$OScore\,(p_i) = \frac{O\,(p_i) - \underset{1 \le j \le n}{MIN}\{O(p_j)\}}{\underset{1 \le j \le n}{MAX}\{O(p_j)) - \underset{1 \le j \le n}{MIN}(O(p_j)\}} \tag{11}$$

Moreover, based on the subcellular location information of proteins, we can calculate the third subcellular location based initial score for p_i as follows:

$$SScore\,(p_i) = \underset{s_i \in Pro\_s(p_i)}{MAX}\{Subcell\,(s_i)\} \tag{12}$$

$$Subcell\,(s_i) = \frac{Sub\_p(s_i) - \underset{1 \le j \le m}{MIN}\{Sub\_p(s_j)\}}{\underset{1 \le j \le m}{MAX}\{Sub\_p(s_j)\} - \underset{1 \le j \le m}{MIN}\{Sub\_p(s_j)\}} \tag{13}$$

Here, $Pro\_s(p_i)$ denotes the set of all subcellular locations, in which the proteinp_i is located, Sub_p (s_i) is the number of proteins in the i-th subcellular localization, and m is the total number of all subcellular localizations.

Finally, based on above formulas, for any given protein p_i, we can define a unique initial score for it as follows:

$$PScore0\,(p_i) = \beta * GScore\,(p_i) + \gamma * OScore\,(p_i) + \delta * SScore\,(p_i) \tag{14}$$

Here, $\beta \in [0, 1]$, $\gamma \in [0, 1]$ and $\delta \in [0, 1]$ are the weights of the GScore(p_i), OScore(p_i) and SScore(p_i) separately, and in addition, there is $\beta + \gamma + \delta = 1$. During simulation, in order to obtain the appropriate combination of these parameters, all possible values of these three parameters will be tried to obtain different initial scores for proteins, among which, the combination corresponding to the highest prediction accuracy of essential proteins will be selected as the final values of these three parameters.Here, $\beta = 0.55$, $\gamma = 0.25$, $\delta = 0.2$.

C. CONSTRUCTION OF THE PREDICTION MODEL PMNMF

First, based on the following formula (15), we will transform the matrix DPI^* to a symmetrical transition probability matrix NTP as follow:

$$NTP\,(p_i, p_j) = \frac{TP(p_i, p_j)}{\sum_{k=1}^{n} TP\,(p_i, p_k)} \tag{15}$$

Here,

$$TP(p_i, p_j) = \begin{cases} \max\left(DPI^*\,(p_i, p_j)\,, DPI^*\,(p_j, p_i)\right) : & if\ i \neq j \\ DPI^*\,(p_i, p_j) : & else \end{cases} \tag{16}$$

Next, based on the transition probability matrix NTP, let PScore(0)=PScore0, then we can iteratively obtain the final scores for all proteins in the weighted PPI network as follows:

$$PScore\,(t+1) = \alpha * NTP * PScore\,(t) + (1-\alpha) * PScore\,(0) \tag{17}$$

Here, the parameter α is used to adjust the ratio of the initial score to the score of the latest iteration, and $PScore(t)$ is the scores of all proteins in the t-th round of iteration.

Finally, according to above descriptions, as shown in algorithm 1, the novel prediction method PMNMF can be presented as follows:

Algorithm 1 PMNMF

Input: Downloaded dataset of known PPIs, downloaded dataset of orthologous information, gene expression information, and subcellular location information of proteins, the iteration condition parameter ε, the dimensionality parameter K, the max iteration times T, and the proportional adjustment parameters α, β, γ and δ.

Output: Top K percent of proteins sorted by values in PScore in descending order

Step1: Generating the original and weighted PPI networks according to formulas (1)-(3);

Step2: Obtaining the non-negative matrices W and H according to formulas (4)-(7),Repeating (6)-(7) until the iteration times exceeds T;

Step3: Calculating the initial scores for proteins according to formulas (8)-(14);

Step4: Obtaining the transition probability matrix according to formulas (15)-(16);

Step5: Let t = t + 1, calculating PScore(t + 1) according to the formula (17) iteratively;

Step6: Repeating Step 5 until ||PScore(t+1) – PScore(t) || < ε;

Step7: Sorting proteins by values in PScore in the descending order;

Step8: Outputing top K percent of sorted proteins.

III. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA

In order to evaluate the predictive performance of PMNMF, in this section, we will compare it with 14 representative basic protein prediction methods including IC [1], CC [3], DC [4], BC [5], SC [6], NC [7], PeC [12], ION [14], CoEWC [16]

TABLE 1. Effects of the parameter α on predication performance of PMNMF based on the DIP database.

Rank α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Top1%	49	49	**50**	**50**	49	48	47	47	46
Top5%	207	211	213	**217**	216	216	213	210	209
Top10%	347	349	353	355	358	359	365	**369**	365
Top15%	449	450	459	466	474	474	**484**	481	479
Top20%	545	550	557	561	563	561	566	566	**568**
Top25%	639	646	649	652	644	645	650	**654**	648

TABLE 2. Effects of the parameter α on predication performance of PMNMF based on the Gavin database.

α / Ran	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Top1%	**18**	**18**	**18**	**18**	**18**	**18**	**18**	**18**	**18**
Top5%	87	**88**	**88**	**88**	**88**	**88**	**88**	**88**	87
Top10%	163	166	166	166	165	165	166	169	**170**
Top15%	221	221	221	222	225	226	**227**	226	**227**
Top20%	281	281	281	281	282	282	282	282	**284**
Top25%	319	319	319	321	321	322	327	330	**332**

TABLE 3. Effects of the parameter α on predication performance of PMNMF based on the Krogan database.

α / Ran	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Top1%	**37**	**37**	**37**	35	35	35	35	35	35
Top5%	145	145	149	153	153	**154**	151	147	141
Top10%	267	270	273	**275**	**275**	272	268	265	259
Top15%	356	360	363	366	366	367	367	**368**	365
Top20%	428	431	433	438	438	444	**445**	441	443
Top25%	499	499	504	503	504	504	501	**505**	**505**

and POEM [17], CVIM[25], NPRI[30], TEGS[20] and RWHN[27] simultaneously. During experiments, we will first download datasets of known PPIs from different benchmark databases including DIP [37], Gavin [38] and Krogan [39] respectively. After pre-processing, we obtain a dataset consisting of 24743 interactions between 5093 proteins from the DIP database, a dataset consisting of 7,669 interactions between 1,855 proteins from the Gavin database, and a dataset consisting of 14317 interactions between 3672 proteins from the Krogan database finally. In addition, according to the databases such as MIPS [40], SGD [41], DEG [42] and SGDP [43] etc., a dataset consisting of 1285 essential proteins can be further obtained, and based on which, 1,167, 714 and 929 essential proteins have been picked out from the databases of DIP, Gavin and Krogan separately. Moreover, based on the dataset provided by Tu BP *et al* [44], we obtain a dataset consisting of gene expression data of 6,776 proteins, which represent the gene expression levels of proteins over consecutive metabolic cycles. Additionally, the orthologous information of proteins will be downloaded from the Inparanoid database (Version7) that includes a collection of pair wise comparisons between 100 whole genomes [45]. After that, the number of times that proteins have orthologous information in reference organisms will be calculated to quantify the homologous information of proteins. Finally, based on the dataset downloaded from the COMPART-MENTS database [46] (downloaded at April 20, 2014), we can obtain a dataset consisting of the subcellular location information of proteins, in which, we will only keep 11 categories of subcellular localization data closely related to essential proteins such as the Endoplasmic, Cytoskeleton, Golgi, Cytosol, Vacuole, Mitochondrion, Endosome, Plasma, Nucleus, Peroxisome and Extracellular etc.

B. EFFECTS OF PARAMETER α ON PERFORMANCE OF PMNMF

In PMNMF, we set a user-defined parameter α with value between 0 and 1 to adjust the ratio of the initial protein fraction to the latest score during iterations. By setting different values to α, we can obtain different prediction accuracies of PMNMF. During simulation, we will choose the number of true essential proteins identified by PMNMF in top 1%, top 5%, top 10%, top 15%, top 20% and top 25% of predicted potential essential proteins when α is set to 0.1, 0.2, 0,3, …and 0.9 as the final results. And in detail, Table 1, Table 2 and Table 3 illustrate these results based on the databases of DIP, Gavin and Krogan respectively. From observing Table 1, it is easy to see that with the increasing

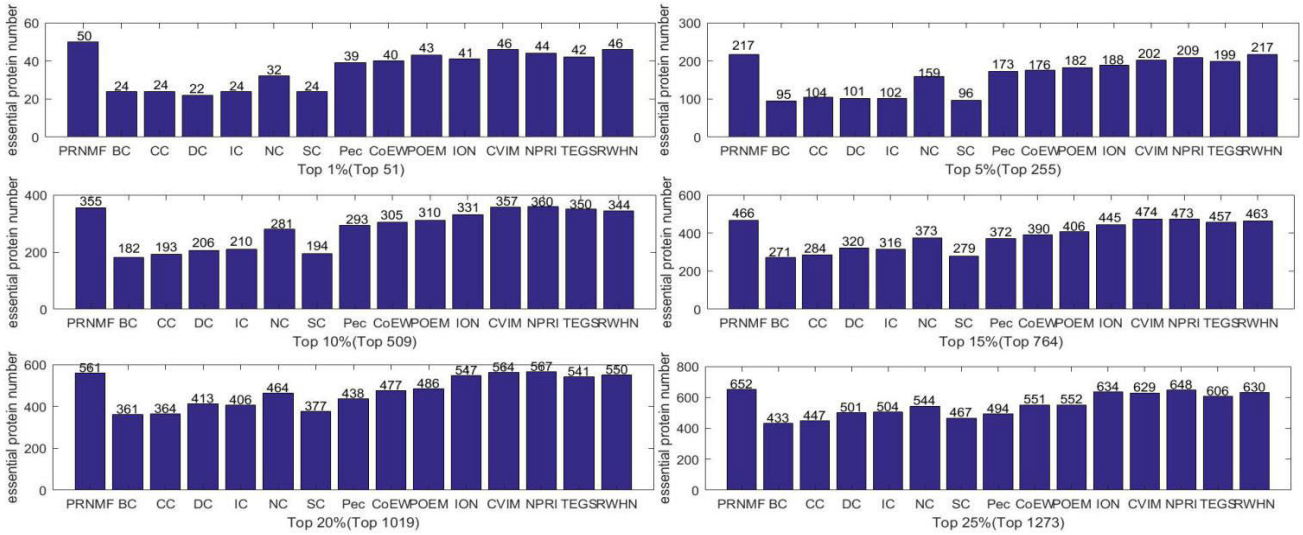


FIGURE 2. The figure shows the comparison results of the number of true essential proteins inferred by PMNMF and 14 competing identification models based on the DIP database. During experiment, proteins will be first sorted in descending order based on their scores calculated by predictive methods such as PMNMF, BC, CC, DC, IC, NC, SC, Pec, POEM, CoEWC, ION, CVIM, NPRI, TEGS and RWHN separately. And then, the top 1%, 5%, 10%, 15%, 20%, and 25% of ranked proteins will be selected as candidate essential proteins. Finally, through comparing with the downloaded dataset of known essential proteins, the number of true essential proteins identified by each method will be calculated and shown in the table, which will be adopted to evaluate the predictive ability of each method. The numbers in parentheses indicate the number of proteins ranked in each interval.

of the value of α from 0.1 to 0.9, the prediction accuracy of PMNMF will increase as well, however, when α exceeds 0.4, the prediction accuracy of PMNMF in the top 1% and 5% of predicted potential essential proteins will decrease gradually. Therefore, based on the DIP database, it will be appropriate to set α to 0.4. From observing Table.2, it is easy to see that PMNMF can achieve the best prediction results while α is set to 0.9 based on the Gavin database. From observing Table.3, it is obvious that 0.6 is a turning point. Therefore, based on the Gavin database, we consider that it will be appropriate to set α to 0.6. According to above analysis, it is easy to known that the value of the parameter α will have obvious effect on the prediction performance of PMNMF. Based on the overall performance on the three datasets, we set α to 0.4.

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

In this section, we will compare PMNMF (while $\alpha = 0.4$) with 14 state-of-the-art competing methods to evaluate its prediction performance based on the DIP database. And as shown in Fig.2, we can see that the prediction performance of PMNMF is better than that of all these 14 competitive methods. Especially, as for the top 1%, top 5%, 10% and top 15% of predicted candidate proteins, PMNMF can achieve reliable predictive accuracies of 98%, 85%, 70% and 60% separately, which are 50%, 24%, 22%, 24%, 32%, 24%, 22%, 20%, 14%, 18%, 8%, 12%, 16% and 8% higher than the predictive accuracy achieved by BC, CC, DC, IC, NC, SC, Pec, POEM, CoEWC ION,CVIM,NPRI,TEGS and RWHN respectively. Next, we further adopt the ROC (Receiver Operating Characteristic) curve and the AUCs (the area under the ROC curve) to compare the prediction performance of PMNMF with these 10 competing methods. The comparison results between PMNMF and BC, CC, DC, IC, NC and SC are illustrated in Fig.3(a), and the comparison results between PMNMF and CoEWC, PeC, POEM and ION are shown in Fig.3(b) respectively. From observing these two figures, it is easy to see that the prediction performance of PMNMF is higher than that of all these 10 competitive methods. And in addition, as shown in Table 4, from observing the AUCs achieved by PMNMF and 10 competing methods based on the DIP database, it is obvious that PMNMF can achieve the highest AUC value of 0.77, which is better than that achieved by all these 10 competitive methods as well.

D. VALIDATION BY JACKKNIFE METHOLOGY

In this section, the jackknife methodology [47] will be implemented on top 1000 candidate essential proteins predicted by PMNMF and 10 competitive models to compare the performances between them based on the DIP database. The comparison results are illustrated in the following Fig.4, in which, the X-axis shows the number of predicted potential essential proteins in descending order according to the predicted scores of proteins, while the Y-axis denotes the cumulative count of the truly proven essential proteins. Especially, Fig.4(a) shows the comparison results among PMNMF, BC, CC, DCIC, NC and SC, from which, it can be seen that the predictive performance of PMNMF is much higher than that of all these 6 competing methods. Moreover, from observing Fig.4(a), it is easy to see that with the increasing of the number of ranked proteins, the performance gap between PMNMF

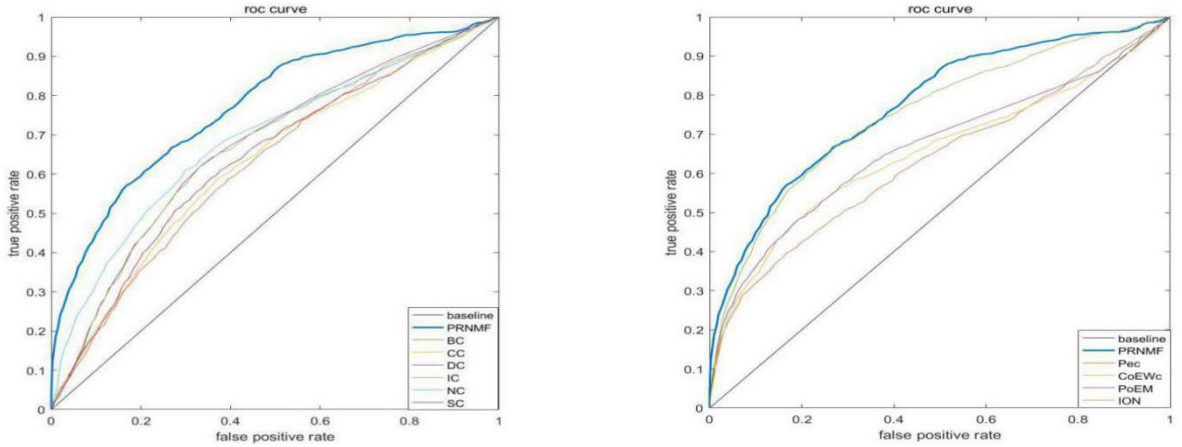


FIGURE 3. The ROC curves achieved by different prediction models based on the DIP database. (a) Comparison results among PMNMF, BC, CC, DC, IC, NC and SC. (b) Comparison results among PMNMF, CoEWC, PeC, POEM and ION.

TABLE 4. AUCs achieved by PMNMF and 10 competitive methods based on the DIP database.

method	PMNMF	BC	CC	DC	IC
AUC	0.77	0.62	0.63	0.67	0.67
NC	SC	CoEWC	PeC	POEM	ION
0.69	0.64	0.63	0.65	0.67	0.75

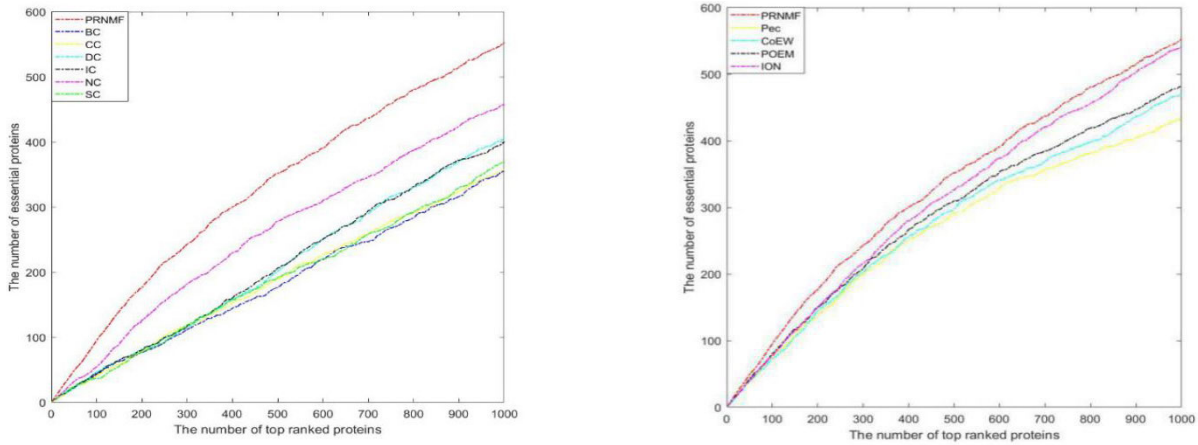
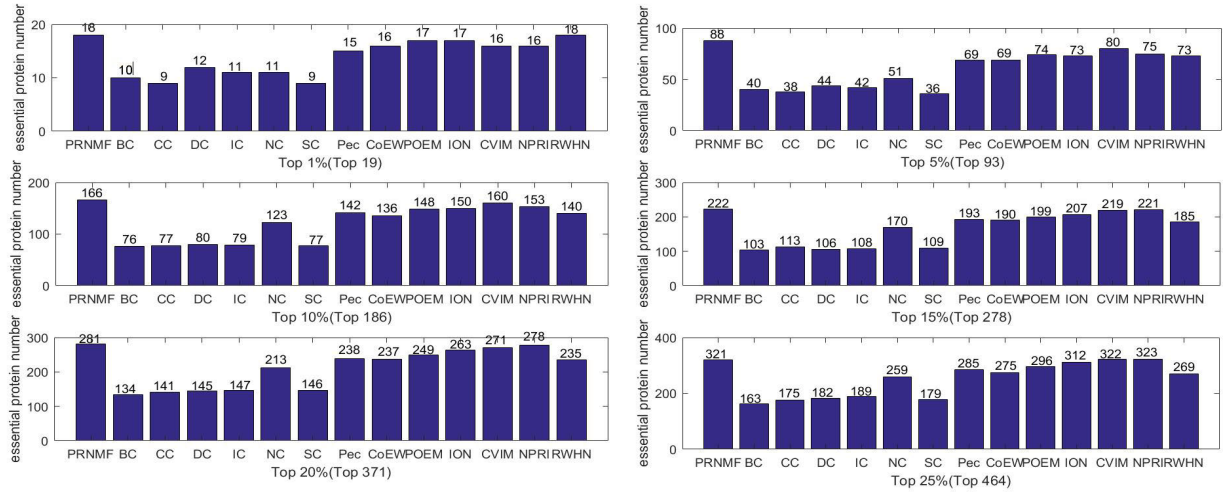


FIGURE 4. The Jackknife curves of PMNMF and 10 competing methods based on the DIP database are shown in this figure, where the X-axis represents the number of ranked potential key proteins from top 100 to top 1000, and the Y-axis is the cumulative count of the true necessary proteins identified by these models. (a) Comparison results of PMNMF, BC, CC, DC, IC, NC and SC. (b) Comparison results among PMNMF, PEC, CoEWC, POEM and ION.

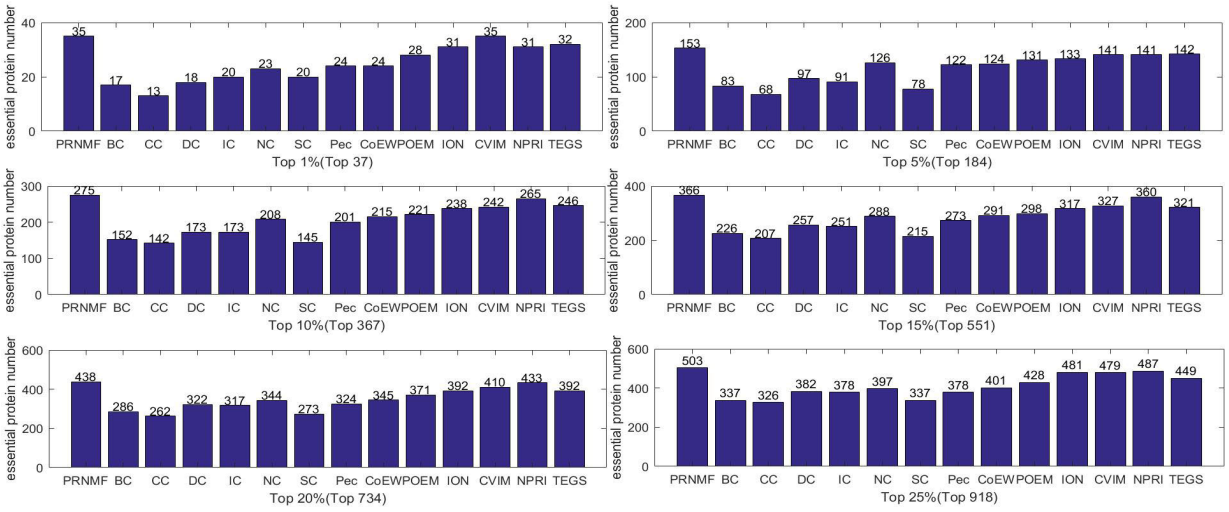
and these competitive methods will increase significantly. Fig.4(b) illustrates the comparison results among PMNMF, Pec, CoEWC, POEM and ION, from which, it can be seen that the prediction performance of PMNMF is to some degree higher than that of all these 4 competing methods as well. However, from observing Fig.4(b), it can be seen that with the increasing of the number of ranked proteins, the performance gap between PMNMF and these competitive methods will increase gradually.

E. DIFFERENCE BETWEEN PMNMF AND 10 COMPETITIVE PREDICTION METHODS

In this section, we will select top 200 proteins predicted by PMNMF and 10 competitive methods based on the DIP database to analyze the difference and commonality between them. Comparison results between PMNMF and 10 competitive methods are illustrated in the following Table.5, in which, Mi indicates one of these 10 methods. |PMNMF ∩Mi| denotes the number of common key proteins



(a)



(b)

FIGURE 5. The figure shows the comparison results of the number of true essential proteins inferred by PMNMF and 13 competing identification models based on the Gavin database and the Krogan database respectively. During experiment, proteins will be first sorted in descending order based on their scores calculated by predictive methods such as PMNMF, BC, CC, DC, IC, NC, SC, PeC, CoEWC, POEM, ION, CVIM, NPRI, TEGS or RWHN separately. And then, the top 1%, 5%, 10%, 15%, 20%, and 25% of ranked proteins will be selected as candidate essential proteins. Finally, through comparing with the downloaded dataset of known essential proteins, the number of true essential proteins identified by each method will be calculated and shown in the table, which will be adopted to evaluate the predictive ability of each method. The numbers in parentheses indicate the number of proteins ranked in each interval. (a) Comparison results based on the Gavin database. (b) Comparison results based on the Krogan database.

identified by both PMNMF and Mi. |PMNMF-Mi| means the number of key proteins detected by PMNMF but not by Mi.

|Mi- PMNMF| represents the number of proteins inferred by Mi but not by PMNMF. {PMNMF-Mi} denotes the set of true key proteins identified by PMNMF but not by Mi, and {Mi- PMNMF} denotes the set of true basic proteins inferred by Mi but not by PMNMF. From observing Table.5, it is easy to know that among these top 200 proteins, the proportions of true essential proteins predicted by PMNMF but not by competing methods are more than 80%, which indicate that PMNMF can achieve much higher identification accuracy and better prediction performance than all these 10 competitive methods

F. RECOGNITION PERFORMANCE OF PMNMF BASED ON THE GAVIN DATABASE AND KROGAN DATABASE

In order to demonstrate the universal applicability of PMNMF method, in this section, we further adopt the Gavin and Krogan databases to compare the prediction performance between PMNMF and some competitive prediction methods. The comparison results are shown in Fig.5 and Fig.6. From observing Fig.5(a), it is clear that based on the Gavin database, the prediction accuracies of PMNMF exceed 89% in the top 1%, top 5% and top 10% of ranked candidate essential proteins. From observing Fig.5(b), it is clear that based on the Krogan database, the prediction accuracies of PMNMF exceed 83% in the top 1% and top 5% of ranked

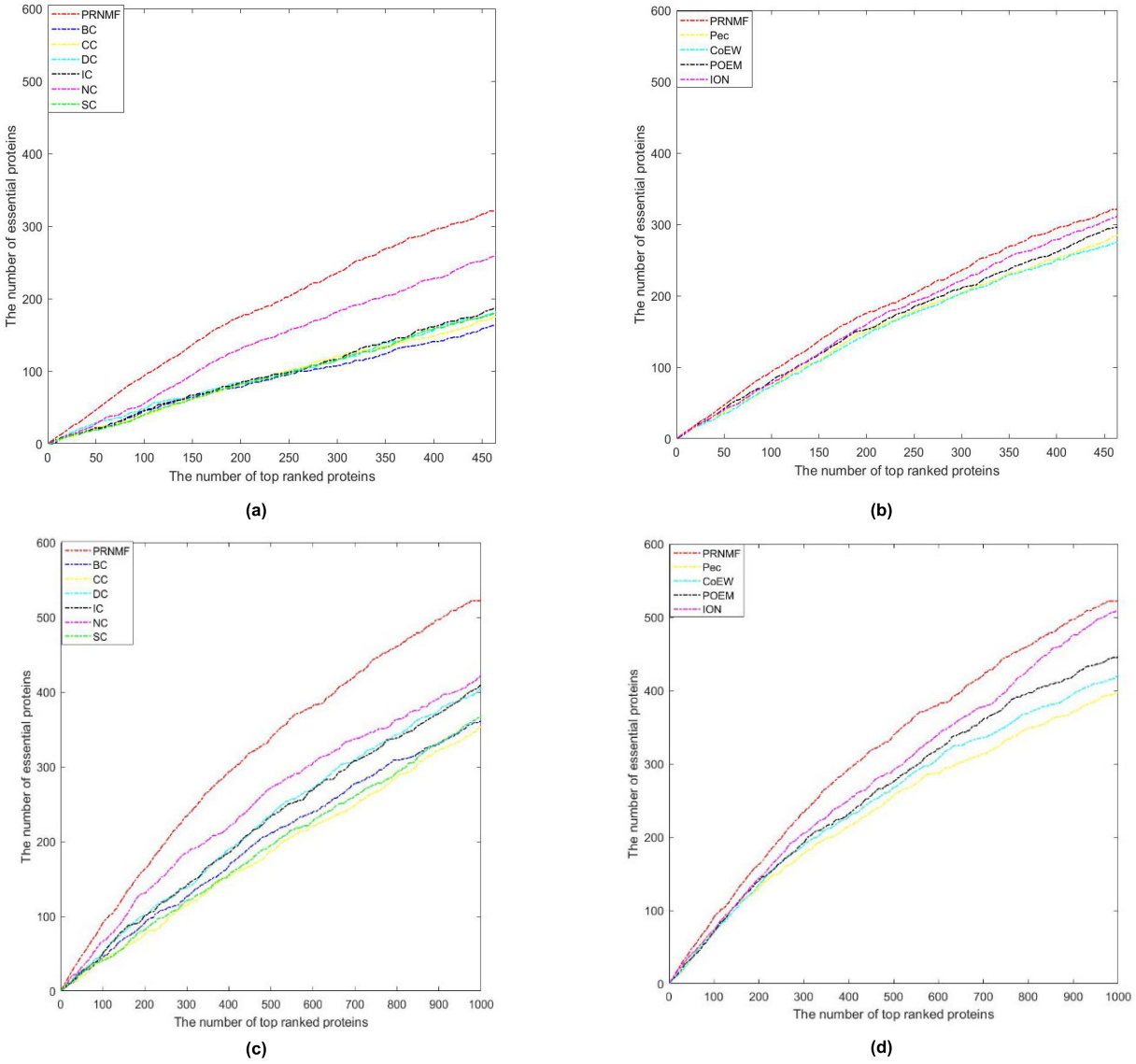


FIGURE 6. **The Jackknife curves of PMNMF and 10 competing methods based on the Gavin database and the Krogan database are shown in this figure respectively, where the X-axis represents the number of ranked potential key proteins from top 100 to top 1000, and the Y-axis is the cumulative count of the true necessary proteins identified by these models. (a) Comparison results of PMNMF, BC, CC, DC, IC, NC and SC based on the Gavin database. (b) Comparison results among PMNMF, PEC, CoEWC, POEM and ION based on the Gavin database. (c) Comparison results of PMNMF, BC, CC, DC, IC, NC and SC based on the Krogan database. (d) Comparison results among PMNMF, PEC, CoEWC, POEM and ION based on the Krogan database.**

candidate essential proteins. And based on both of these two databases, the prediction accuracies achieved by PMNMF are higher than all other competitive prediction methods. In addition, from observing Fig.6(a) and Fig.6(b), we can find that the prediction performance of PMNMF is higher than all these 10 methods based on the Gavin database, and from observing Fig.6(c) and Fig.6(d), we can find that the prediction performance of PMNMF is higher than all these 10 methods based on the Krogan database as well.

IV. DISCUSSION

Essential proteins are important for cell growth and regulation processes. In recent years, accumulating computational methods have been proposed to identify essential proteins. However, due to the effects of false positives and false negatives in original PPI data obtained by high-throughput techniques, it is still a challenging work to develop a stable and accurate essential protein prediction model. Inspired by the fact that it can improve the prediction performance of computational models by integrating PPI networks with multiple biological information of proteins, a novel prediction model called PMNMF based on the Non-negative Matrix Factorization is designed in this manuscript. In PMNMF, a weighted PPI network is first constructed by extracting the topological information of proteins from the original PPI network, and then, by applying the NMF method on the weighted PPI

TABLE 5. Differences between PMNMF and 10 competitive methods based on the top 200 proteins and the DIP database.

Centrality measures (Mi)	|PMNMF∩Mi|	| PMNMF-Mi|	Percentage of the essential proteins in {PMNMF−Mi}	Percentage of the essential proteins in {Mi−PMNMF}
BC	23	177	87.57%	31.07%
CC	22	178	87.08%	22.02%
DC	29	171	87.13%	31.58%
IC	27	173	86.71%	30.64%
NC	60	140	85.71%	49.29%
SC	26	174	86.78%	29.31%
Pec	93	107	85.05%	49.53%
CoEWC	97	103	85.44%	53.40%
POEM	97	103	81.55%	56.31%
ION	113	87	83.91%	52.87%

network and combining with biological information of proteins, an improved Page-Rank algorithm is introduced to calculate the importance scores for proteins. Experimental results show that PMNMF can achieve superior prediction results than state-of-the-art prediction models, which demonstrates that PMNMF is an effective prediction method for key protein prediction. Of course, there are still some shortcomings in current version of PMNMF. For example, more biological information of proteins being considered, the prediction performance of PMNMF may become better.

V. CONCLUSION

The main contributions of this manuscript can be summarized as follows: (1) A weighted PPI network is established based on the topological information of proteins in the original PPI network. (2) The Non-negative Matrix Factorization is introduced to obtain the transition probability matrix. (3) An improved Page-Rank algorithm is designed to estimate the critical scores of proteins. However, there are still some limitations in current version of PMNMF. For example, since a random algorithm is adopted to initialize these two non-negative factorization matrices, the result obtained by the NMF algorithm has a certain degree of randomness as well, we improve the stability of results through multiple iterations, more effective methods is worthy to be explored in the future researches.

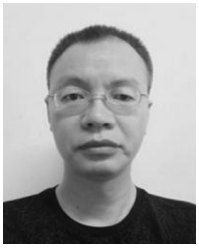
REFERENCES

[1] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social Netw.*, vol. 11, no. 1, pp. 1–37, Mar. 1989.
[2] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
[3] S. Wuchty and P. F. Stadler, "Centers of complex networks," *J. Theor. Biol.*, vol. 223, no. 1, pp. 45–53, Jul. 2003.
[4] M. W. Hahn and A. D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," *Mol. Biol. Evol.*, vol. 22, no. 4, pp. 803–806, 2004.
[5] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *J. Biomed. Biotechnol.*, vol. 2005, no. 2, pp. 96–103, 2005.
[6] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 5, May 2005, Art. no. 056103.
[7] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1070–1080, Jul./Aug. 2012.
[8] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, "A local average connectivity-based method for identifying essential proteins from the network level," *Comput. Biol. Chem.*, vol. 35, no. 3, pp. 143–150, Jun. 2011.
[9] M. Li, Y. Lu, J. Wang, F. X. Wu, and Y. Pan, "A topology potential-based method for identifying essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 372–383, Mar./Apr. 2015.
[10] Y. Qi and J. Luo, "Prediction of essential proteins based on local interaction density," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 6, pp. 1170–1182, Nov. 2016.
[11] B. Chen and F.-X. Wu, "Identifying protein complexes based on multiple topological structures in PPI networks," *IEEE Trans. Nanobiosci.*, vol. 12, no. 3, pp. 165–172, Sep. 2013.
[12] M. Li, H. Zhang, J.-X. Wang, and Y. Pan, "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data," *BMC Syst. Biol.*, vol. 6, no. 1, p. 15, 2012.
[13] X. Tang, J. Wang, J. Zhong, and Y. Pan, "Predicting essential proteins based on weighted degree centrality," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 2, pp. 407–418, Mar. 2014.
[14] W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan, "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks," *BMC Syst. Biol.*, vol. 6, no. 1, p. 87, 2012.
[15] W. Peng, J. Wang, Y. Cheng, Y. Lu, F. Wu, and Y. Pan, "UDoNC: An algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 276–288, Mar. 2015.
[16] X. Zhang, J. Xu, and W.-X. Xiao, "A new method for the discovery of essential proteins," *PLoS ONE*, vol. 8, no. 3, Mar. 2013, Art. no. e58763.
[17] B. Zhao, J. Wang, M. Li, F. Wu, and Y. Pan, "Prediction of essential proteins based on overlapping essential modules," *IEEE Trans. NanoBiosci.*, vol. 13, no. 4, pp. 415–424, Dec. 2014.
[18] J. Luo and Y. Qi, "Identification of essential proteins based on a new combination of local interaction density and protein complexes," *PLoS ONE*, vol. 10, no. 6, Jun. 2015, Art. no. e0131418.
[19] S. Keretsu and R. Sarmah, "Weighted edge based clustering to identify protein complexes in protein–protein interaction networks incorporating gene expression profile," *Comput. Biol. Chem.*, vol. 65, pp. 69–79, Dec. 2016.
[20] M. Li, W. Li, F.-X. Wu, Y. Pan, and J. Wang, "Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information," *J. Theor. Biol.*, vol. 447, pp. 65–73, Jun. 2018.
[21] M. Li, Y. Lu, Z. Niu, and F.-X. Wu, "United complex centrality for identification of essential proteins from PPI networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 370–380, Mar. 2017.
[22] J. Luo and J. Wu, "A new algorithm for essential proteins identification based on the integration of protein complex co-expression information and edge clustering coefficient," *Int. J. Data Mining Bioinf.*, vol. 12, no. 3, p. 257, 2015.
[23] B. Zhao, S. Hu, X. Liu, H. Xiong, X. Han, Z. Zhang, X. Li, and L. Wang, "A novel computational approach for identifying essential proteins from multiplex biological networks," *Frontiers Genet.*, vol. 11, Apr. 2020.

[24] B. Zhao, X. Han, X. Liu, Y. Luo, S. Hu, Z. Zhang, and L. Wang, "A novel method to predict essential proteins based on diffusion distance networks," *IEEE Access*, vol. 8, pp. 29385–29394, 2020.
[25] S. Li, Z. Chen, X. He, Z. Zhang, T. Pei, Y. Tan, and L. Wang, "An iteration method for identifying yeast essential proteins from weighted PPI network based on topological and functional features of proteins," *IEEE Access*, vol. 8, pp. 90792–90804, 2020, doi: 10.1109/ACCESS.2020.2993860.
[26] X. Lei, X. Yang, and F. Wu, "Artificial fish swarm optimization based method to identify essential proteins," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 495–505, Mar./Apr. 2020, doi: 10.1109/TCBB.2018.2865567.
[27] B. Zhao, Y. Zhao, X. Zhang, Z. Zhang, F. Zhang, and L. Wang, "An iteration method for identifying yeast essential proteins from heterogeneous network," *BMC Bioinf.*, vol. 20, no. 1, p. 355, Dec. 2019.
[28] W. Dai, Q. Chang, W. Peng, J. Zhong, and Y. Li, "Network embedding the protein–protein interaction network for human essential genes identification," *Genes*, vol. 11, no. 2, p. 153, 2020.
[29] F. Zhang, W. Peng, Y. Yang, W. Dai, and J. Song, "A novel method for identifying essential genes by fusing dynamic protein–protein interactive networks," *Genes*, vol. 10, no. 1, p. 31, 2019.
[30] Z. Chen, Z. Meng, C. Liu, X. Wang, L. Kuang, T. Pei, and L. Wang, "A novel model for predicting essential proteins based on heterogeneous protein-domain network," *IEEE Access*, vol. 8, pp. 8946–8958, 2020.
[31] X. Lei and X. Yang, "A new method for predicting essential proteins based on participation degree in protein complex and subgraph density," *PLoS ONE*, vol. 13, no. 6, 2018, Art. no. e0198998.
[32] K. Hosoda, M. Watanabe, H. Wersing, E. Körner, H. Tsujino, H. Tamura, and I. Fujita, "A model for learning topographically organized parts-based representations of objects in visual cortex: Topographic nonnegative matrix factorization," *Neural Comput.*, vol. 21, no. 9, pp. 2605–2633, Sep. 2009.
[33] D. S. Huang and C. H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.
[34] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 599–607, Jul. 2009.
[35] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[36] D. Horyu and T. Hayashi, "Comparison between pearson correlation coefficient and mutual information as a similarity measure of gene expression profifiles," *Jpn. J. Biometrics*, vol. 33, no. 2, pp. 125–143, 2013.
[37] I. Xenarios, "DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, Jan. 2002.
[38] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, and M. Boesche, "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, Mar. 2006.
[39] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchnko, and J. Li, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.
[40] H. W. Mewes, "MIPS: Analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D169–D172, Jan. 2006.
[41] J. Cherry, "SGD: Saccharomyces genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, Jan. 1998.
[42] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Res.*, vol. 37, pp. D455–D458, Jan. 2009.
[43] *Saccharomyces Genome Deletion Project*. Accessed: Jun. 20, 2012. [Online]. Available: http://yeastdeletion.stanford.edu/
[44] B. P. Tu, "Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, Nov. 2005.
[45] G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer, "InParanoid 7: New algorithms and tools for eukaryotic orthology analysis," *Nucleic Acids Res.*, vol. 38, pp. D196–D203, Jan. 2010.
[46] J. X. Binder, S. Pletscher-Frankild, K. Tsafou, and C. Stolte, "COMPARTMENTS: Unification and visualization of protein subcellular localization evidence," *Database*. to be published, doi: 10.1093/database/bau012.
[47] A. G. Holman, P. J. Davis, J. M. Foster, C. K. Carlow, and S. Kumar, "Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi," *BMC Microbiol.*, vol. 9, no. 1, p. 243, 2009.

JIN LIU is currently pursuing the bachelor's degree in information and computing science with Changsha University. Her current research interest is bioinformatics.

XIANGYI WANG is currently pursuing the bachelor's degree in information and computing science with Changsha University. Her current research interest is bioinformatics.



ZHIPING CHEN received the B.S. degree in computer science and technology from Xiangtan University, in 1994, and the M.S. and Ph.D. degrees in computer science and technology from Hunan University, in 1997 and 2003, respectively. From 1997 to 2009, he has taught in Hunan University. He is currently a Professor with Changsha University. His current research area includes bioinformatics mainly.

YIHONG TAN received the Ph.D. degree in computer science and technology from Hunan University, in 2012. He is currently a Full Professor with Changsha University. His current research area is mainly bioinformatics.

XUEYONG LI received the M.S. degree in computer science and technology from Hunan University, in 2003, and the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, in 2012. He is currently a Professor with Changsha University. His current research area is mainly information theory.

ZHEN ZHANG received the B.S. degree from Anhui Agricultural University, Hefei, Anhui, in 2006, the M.S. degree from Central South University, in 2009, and the Ph.D. degree from Northwestern Polytechnical University, in 2019. He is currently an Associate Professor with Changsha University, China. His research area focuses on big data mainly.

LEI WANG received the Ph.D. degree in computer science from Hunan University, China, in 2005. From 2005 to 2007, he was a Postdoctoral Fellow with Tsinghua University, China. After that, he moved to USA and Canada as a Visiting Scholar in Duke University and Lakehead University successively. From 2009 to 2011, he was an Associate Professor with Hunan University. From 2011 to 2018, he was a Full Professor with Xiangtan University. He is currently a Full Professor and an Academic Leader of computer engineering with Changsha University, China. He has published more than 100 peer-reviewed articles. His main research areas include bioinformatics and the Internet of Things.