Mathematical Biosciences and Engineering

MBE, 19(6): 6331–6343.
DOI: 10.3934/mbe.2022296
Received: 15 February 2022
Revised: 26 March 2022
Accepted: 08 April 2022
Published: 20 April 2022

http://www.aimspress.com/journal/MBE

Research article

An efficient strategy for identifying essential proteins based on homology, subcellular location and protein-protein interaction information

Zhihong Zhang[1], Yingchun Luo[2,*], Meiping Jiang[2], Dongjie Wu[3], Wang Zhang[4], Wei Yan[1] and Bihai Zhao[1]

[1] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, Hunan 410022, China
[2] Department of Ultrasound, Hunan Provincial Maternal and Child Health Care Hospital, Changsha, Hunan 410008, China
[3] Department of Banking and Finance, Monash University, Clayton, Victoria 3168, Australia
[4] Department of Optoelectronic Engineering, Jinan University, Guangzhou, Guangdong 510632, China

*** Correspondence:** Email: yingchunluo123@163.com; Tel: +8673184261431; Fax: +8673184261431.

Abstract: High throughput biological experiments are expensive and time consuming. For the past few years, many computational methods based on biological information have been proposed and widely used to understand the biological background. However, the processing of biological information data inevitably produces false positive and false negative data, such as the noise in the Protein-Protein Interaction (PPI) networks and the noise generated by the integration of a variety of biological information. How to solve these noise problems is the key role in essential protein predictions. An Identifying Essential Proteins model based on non-negative Matrix Symmetric tri-Factorization and multiple biological information (IEPMSF) is proposed in this paper, which utilizes only the PPI network proteins common neighbor characters to develop a weighted network, and uses the non-negative matrix symmetric tri-factorization method to find more potential interactions between proteins in the network so as to optimize the weighted network. Then, using the subcellular location and lineal homology information, the starting score of proteins is determined, and the random walk algorithm with restart mode is applied to the optimized network to mark and rank each protein. We tested the suggested forecasting model against current representative approaches using a public database. Experiment shows high efficiency of new method in essential proteins identification. The effectiveness of this method shows that it can dramatically solve the noise problems that existing in the multi-source biological information itself and cased by integrating them.

Keywords: essential protein; protein-protein interaction; non-negative matrix symmetric tri-factorization; multiple biological information; subcellular location information; homology information

1. Introduction

Essential proteins are required for organism life, and their absence results in the loss of functional modules of protein complexes, as well as the death of the organism [1]. Essential proteins identification aids in the understanding of cell growth control mechanisms, the discovery of disease-causing genes and possible therapeutic targets, and has crucial theoretical and practical implications for drug development and disease therapy. In biological experiments, essential proteins are mainly identified by gene culling, gene suppression, transposon mutation and other methods, which cost lot of time and difficult unfortunately. Essential protein identification using computational approaches becomes achievable as high-throughput data accumulates. This identification method means utilizing the available data to find the key features that affect the importance of proteins and to determine if it is important of biological functions based on these features. The most common measuring technique is based on the topological properties of the PPI network to obtain network topology features, like Degree Centrality (DC) [2], Information Centrality (IC) [3], Closeness Centrality (CC) [4] and Subgraph Centrality (SC) [5], Betweenness Centrality (BC) [6], sum of Edge Clustering Coefficient Centrality (NC) [7]. These methods are sensitive to network structure, so false positive noise and data missing will reduce the performance of prediction easily.

In addition to characteristics of network topological, the biological characteristics involved in essential proteins identification mainly include sequence features and functional features. Zhang and Li et al. combined features of profiles of gene expression with topological features of PPI network, and proposed CoEWC [8] and PeC [9] methods respectively. Zhao et al. [10] put forward an essential protein detection model named POEM which utilize the module features of essential proteins. A weighted network with high confidence is built based on the topological structure and intrinsic characters of network and information about expression of genes, and overlap of functional modules, that coupling nature is weak and cohesive nature is strong, are discovered. In the end, the weighted density of the module to which the protein belongs was used to determine scores. Zhang et al. [11] got a new model named FDP to employs the global and local topological properties of network and protein homology information, to combine the dynamic PPI network at different times. In 2021, Zhong et al. [12] introduced a novel measuring approach named JDC that binary gene expression data with a dynamic threshold and combines the Jaccard index of similarity and degree centrality.

The method based on multi-source data integration effectively improve the prediction's level of accuracy and robustness. The commonly used processing method is to build a highly reliable weighted PPI network through weighted summary and the features are different for different prediction methods. The processing method of simple superposition will obfuscate the complicated relationship that exists between the multi-source data and generate artificial noise. The parameter setting is also matter which will influence the practical application of the algorithm. Non-negative matrix tri-factorization (NMTF) [13] is mainly used to analyze data matrices with non-negative elements, disintegrate the input matrix into three non-negative factor matrices, and approximate the input matrix through low-rank non-negative representation. It has been widely used in many fields such as text mining [14], recommendation system [15,16] and biological data analysis [17,18].

In view of the advantages of NMTF in data analysis and integrate protein homology information and subcellular location information to improve the prediction performance of essential proteins, an approach of non-negative matrix symmetric tri-factorization (IEPMSF) is offered as an optimal method for solving the noise problems in identifying essential proteins. In order to avoid more noise caused by multi-source data integration, this paper only uses the topological features of the original protein interaction data to construct the protein weighted network. But this method is not optimal because of the existence of false negatives and false positives. To solve this problem, the traditional NMTF algorithm is optimized. The factorization process is regarded as the "soft clustering" process of proteins, to predict the potential protein-protein interactions by a non-negative matrix symmetric tri-factorization algorithm (NMSTF), thus forming the optimal protein weighted network. Finally, to achieve the goal of predicting essential proteins, the homology information and subcellular location information of proteins are combined to create an initial score for each protein, which is then used to score and order each protein in the optimized network using the restart random walk algorithm.

2. Materials and methods

2.1. Basic framework of the model

This paper builds an improved protein-weighted network using the protein-protein interaction network and the NMSTF algorithm to increase the accuracy of important protein identification, and integrates subcellular localization information with protein homology information to design an essential model to identify essential proteins, IEPMSF. The model consists of three modules: weighted network building module, weighted network optimization module, and proteins scoring and sorting module.

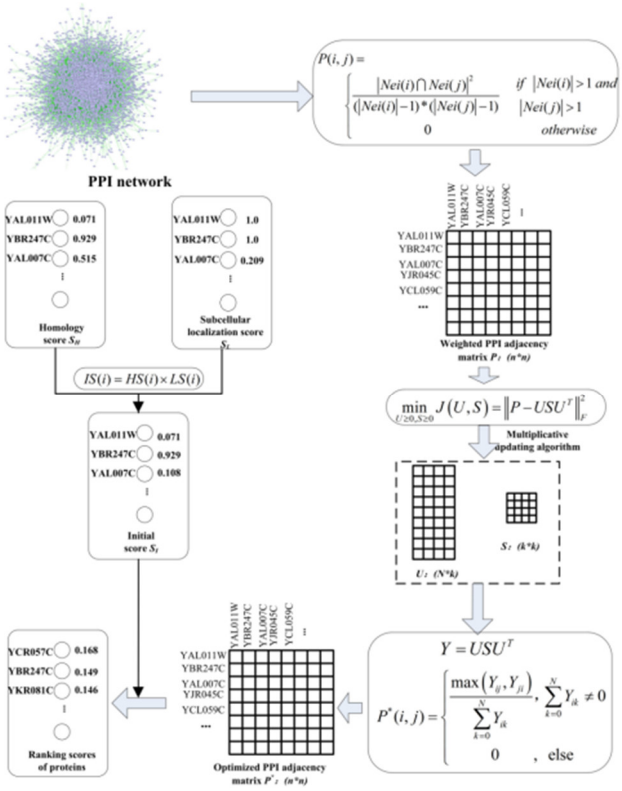


Figure 1. Overall workflow of IEPMSF for identifying essential proteins.

2.1.1. Weighted network construction module

Through topological analysis of yeast networks, the researchers found that PPI networks have small-world and non-scale characteristic [19] and that essential proteins have a strong connection with the topological properties of proteins. The co-neighbor coefficient is commonly utilized in the functional recognition [20] of proteins in PPI networks, demonstrating that the more similar neighbors two proteins in a network have, the more likely they are to interact. To measure the degree of interaction between the two proteins, we use the co-neighbor coefficient to give the edge weights of the network of protein interaction.

A simple undirected graph $G = (V, E)$ can be a model of a PPI network. Here, the nodes set $V = \{v_1, v_2, \cdots\}$ as proteins, the edges set $E = \{e_1, e_2, e_3 \cdots\}$ is a representation for the interaction of two different proteins. Defining a weighted network is $WG = (V, E, P)$, where $P(i, j)$ indicating the likelihood of the interaction of the v_i and v_j proteins, can be computed using the equation below :

$$P(i,j) = \begin{cases} \frac{|Nei(i) \cap Nei(j)|^2}{(|Nei(i)|-1)*(|Nei(j)|-1)} & if\ |Nei(i)| > 1\ and\ |Nei(j)| > 1 \\ 0 & otherwise \end{cases} \tag{1}$$

where $Nei(i)$ and $Nei(j)$ respectively represent collection of neighbor nodes of the v_i and v_j, $|Nei(i) \cap Nei(j)|$ represent the number of common neighbors. If there are not any common neighbor proteins between the v_i and v_j, then $P(i, j) = 0$. We are going to assume that the probability of the interaction, the co-neighbor coefficient between the proteins, is independent of each other, and it's going to be in the range of 0 to 1.

2.1.2. Weighted network optimization module

As previously stated, false positives and false negatives can be found in PPI networks derived from high-throughput biological research. In other words, there are still some uncertainties in the construction of weighted networks based on protein interactions. NMTF was proposed by Ding in 2006 [13], which is an effective tool applied to recommendation systems successfully. Therefore, we can exploit the potential new protein interactions based on the existing protein and protein interaction data by using NMTF technology.

The traditional NMTF is the decomposition of the correlation matrix Y^{n*n} into three low-rank sub-matrices, $F \in R^{n*k}$, $S \in R^{k*k}$ and $G \in R^{k*n}$, by which to approximate the original input matrix, as follows:

$$P \approx Y = FSG^T \tag{2}$$

Where the parameter k represents the factorization level and reflects the total number of possible vectors in the column spaces and row spaces. After being weighted to the protein interaction network with co-neighbor coefficients, the association matrix of the network can be constructed to represent the connection relationship between proteins. The elements in the correlation matrix are the co-neighbor values for each edge. Due to the singularity of nodes in the protein interaction network and the resulting correlation matrix is a symmetry matrix, the simple utilization of the conventional NMTF technology is not reasonably explanatory. Hart [21] pointed out that essential proteins often gather together, and the criticality of proteins is related to protein complexes rather than dependent on a single

protein, which indicates that essential proteins have modular properties. Specifically, given a non-negative input matrix P, factor matrix S can be seen as the cluster index [14] of the vertex. Based on this, this paper proposes an improved NMTF algorithm called a non-negative matrix symmetric three-factors decomposition to rewrite Eq (2) into the following form:

$$P \approx Y = USU^T \tag{3}$$

Among them, $U \in R^{n*k}$ can be seen as "soft" clustering labels of proteins, and $S \in R^{k*k}$ as a correlation matrix between protein modules, $S = S^T$. Then we can design the loss objective function of Eq (3) as follows:

$$D = \min_{U \geq 0, S \geq 0} J(U,S) = \|P - USU^T\|_F \tag{4}$$

Where $\|\cdot\|_F$ refers to the Frobenius specification. We use the multiplication update iteration technique to derive the objective function on the basis of employing the auxiliary function because the object function is a joint nonconvex problem. According to the rules of Squared frobenius norm we can know $||X||_2 = Tr(X^TX)$, which can solve D as follows:

$$D = Tr(P^TP - 2P^TUSU^T + US^TU^TUSU^T) \tag{5}$$

Solve partial differential equations for U and S factors in Eq (5) respectively:

$$\frac{\partial D}{\partial U} = -4PUS + 4USU^TUS$$

$$\frac{\partial D}{\partial S} = -2U^TPU + 2U^TUSU^TU \tag{6}$$

Followed as Karush-Kuhn Tucker (KKT) complementary condition, we can find a static point, the KKT condition for U and S. These rules can be written as follows:

$$\frac{\partial D}{\partial U_{ik}} U_{ik} = 0 \tag{7}$$

By Eq (7), we can get:

$$(USU^TUS - PUS)_{ik} U_{ik} = 0$$

$$U_{ik} = U_{ik} \frac{(PUS)_{ik}}{(USU^TUS)_{ik}} \tag{8}$$

Similarly, the S can be calculated using the same procedure:

$$S_{ik} = S_{ik} \frac{(U^TPU)_{ik}}{(U^TUSU^TU)_{ik}} \tag{9}$$

These rules can be expressed in a matrix form:

$$U_{ik} \leftarrow U_{ik} \frac{(PUS)_{ik}}{(USU^T US)_{ik}}$$

$$S_{ik} \leftarrow S_{ik} \frac{(U^T PU)_{ik}}{(U^T USU^T U)_{ik}} \tag{10}$$

According to the above multiplication update iteration rules, the final U and S can be calculated, so as to obtain the optimal $Y = USU^T$ approximating the original input matrix.

After the above data processing, we construct an optimized network association matrix, and conduct the corresponding standardization processing as follows:

$$P^*(i,j) = \begin{cases} \frac{\max(Y_{ij}, Y_{ji})}{\sum_{k=0}^{N} Y_{ik}}, & \sum_{k=0}^{N} Y_{ik} \neq 0 \\ 0, & else \end{cases} \tag{11}$$

The cumulative sum of each row of i in the matrix P^* is 0 or 1.

2.1.3. Proteins scoring and sorting module

We give an initial score to every protein from protein interaction network given by direct homology information and sub-cell localization information to improve the accuracy of essential protein prediction.

Studies have shown that when a protein has more homologous proteins in a reference species, it is highly likely to be an essential protein. The direct homology score of protein node v_i is calculated by the equation below:

$$HS(i) = \frac{HP(i)}{\max\limits_{1 \le j \le |V|} HP(j)} \tag{12}$$

where $HP(i)$ represent how many direct homologous proteins in the reference species collection SC node v_i has, as follows:

$$HP(i) = \sum_{m \in SC} TN_i \quad \text{where } TN_i = \begin{cases} 1 & if\ v_i \in XS_m \\ 0 & otherwise \end{cases} \tag{13}$$

where the XS_m represents a collection of proteins with direct homologous proteins and is a subset of V. For those proteins that possess homologous proteins in all reference species, their direct homology score of 1 is given. Instead, if a protein does not have a direct homologous protein in all reference species, it has a score of 0.

Previous research has revealed that the essential state of proteins is not simply linked to the biological properties of PPI networks, but also to their location in space. Therefore, making full use of subcellular localization information is important for essential proteins prediction. Studies have shown that essential proteins are found in higher concentrations in certain subcellular locations than non-essential, and evolve more conserved [22]. Let $L(R)$ be the protein set appearing at subcellular location r, and the frequency of protein appearing at it is possible to calculate each subcellular location r, as shown below:

$$OF(r) = \frac{|L(r)|}{\max\limits_{k \in R} |L(k)|} \tag{14}$$

Where $|L(r)|$ represents the number of proteins present at subcellular location r, and R represents the set containing each subcellular location. For a protein v_i, let $C(i)$ be the set of subcellular sites in which it occurs, and the definition of subcellular localization score *LS(i)* is the score of the maximum frequency of its occurrence at all subcellular locations by using the following equation:

$$\mathrm{LS(i)} = \max_{r \in C(i)} OF(r) \tag{15}$$

Combined with the direct homology score and subcellular localization score obtained by Eqs (12) and (15), the initial value score, $IS(i)$, which is possible to compute the v_i of each protein in the protein interaction network, with following equation:

$$HS(i) = HS(i) \times \mathrm{LS}(i) \tag{16}$$

Based on the weighted network constructed previously and the initial score based on the multi-source biological information, the final score, $FS(i)$, of a protein v_i from network can be calculated as bellow:

$$FS(i) = \alpha \sum_{j \in Nei(i)} P^*(i,j) FS(j) + (1-\alpha) IS(i) \tag{17}$$

where, $Nei(i)$ shows the set of neighbor nodes of v_i.

As can be seen from Eq (17), a protein's final score may be thought of as a linear combination of its multi-source bioinformatics mark and its neighboring correlation mark. Among them, the percentage of these two scores are adjusted using parameter a. When a is equal to 0, the final protein score is only related to the multi-source biological information score, and when the value of a is 1, the score is only related to the common neighbor properties of a protein. However, the amount of protein in the network is numerous and they have great computational complexity. Therefore, we can rewrite Eq (17) into the form of a matrix vector:

$$FS(i) = \alpha * P^* * FS + (1-\alpha) * IS \tag{18}$$

Finally, the Jacobi iterative method can be used to quantitatively solve the Eq (18):

$$FS^t = \alpha * P^* * FS^{t-1} + (1-\alpha) * IS \tag{19}$$

The number of iterations is represented by t = (0,1,2, …).

3. Expeiments and results

3.1. Experimental data source

The validity of the IEPMSF model was evaluated by using the basic data of essential protein. The dataset incorporates essential protein dataset, PPI network dataset, protein homology information dataset, and subcellular location dataset. The benchmark essential proteins involved in the datasets are 1199 essential proteins, mainly from databases of MIPS [23], SGD [24], DEG [25], and SGDP [26]. The DIP [27] database is used to get the PPI network data. Excluding repeated protein interactions and the protein itself interactions, there are 5093 proteins and 24,743 interactions in the collection. The

subcellular location data was downloaded from the COMPARTMENTS [28] database, which integrates MGD [29], SGD [24], UniProtKB [30], WormBase [31] and FlyBase [32] databases and eventually obtains 3923 proteins with subcellular location information. The homologous protein data is gathered from the InParanoid database's 7th edition [33], which included pair-wise comparisons of entire genomes of 99 eukaryotes and 1 prokaryote.

To determine the significance of proteins in the protein interaction network, proteins are compared with results derived by the algorithm IEPMSF or other existing ways, DC [2], IC [3], CC [4], BC [5], SC [6], NC [7], PeC [9], CoEWC [8], POEM [10], FDP [11] and JDC [12] for example.

3.1.1. Influence of the parameter *a* on the capability of the IEPMSF method

In IEPMSF, the ordering score of the proteins are different depending on the a. To study the impact of parameter *a* on the capability of IEPMSF method, we experimented with several values ranging from 0 to 1 to see how they affected the accuracy of essential proteins prediction of IEPMSF. Table 1 contains detailed experimental data. The range of essential candidates selected is from the top 100 to the top 600. The ratio of actually essential proteins predicted determines predictive accuracy.

As shown in Table 1, it is shown that when $a = 0$, the predicted essential protein only considers the direct homology of the protein, while when $a = 1$, the predicted essential protein only considers the co-neighbor information. When $a = 0$ or $a = 1$, the IEPMSF performs worse than the values of 0 to 1. This means that combination of the direct homologues of proteins and their neighbours can predict the required proteins more accurately than if only one of these properties is considered. To compare with other algorithms, as $a = 0.1$, when the top 100 ranking proteins are chosen as essential protein candidates, the accuracy can reach 0.97, as shown in the experimental findings in Figure 2.

Table 1. Influence of the parameter *a* on the accuracy of IEPMSF prediction.

a	Top 100	Top 200	Top 300	Top 400	Top 500	Top 600
0	78.00%	77.00%	73.70%	72.30%	67.00%	63.00%
0.1	97.00%	84.50%	78.00%	74.00%	68.40%	64.50%
0.2	92.00%	85.50%	79.70%	74.50%	69.80%	65.30%
0.3	89.00%	86.00%	78.30%	72.80%	69.20%	64.80%
0.4	87.00%	83.00%	76.00%	71.80%	68.60%	65.00%
0.5	87.00%	78.00%	74.00%	70.00%	67.20%	64.30%
0.6	86.00%	77.00%	71.30%	69.00%	64.80%	63.00%
0.7	85.00%	75.00%	69.00%	66.80%	63.80%	60.00%
0.8	82.00%	74.00%	67.30%	64.50%	62.00%	59.20%
0.9	83.00%	75.00%	65.30%	62.80%	59.60%	57.30%
1	81.00%	71.00%	64.70%	59.80%	55.80%	53.20%

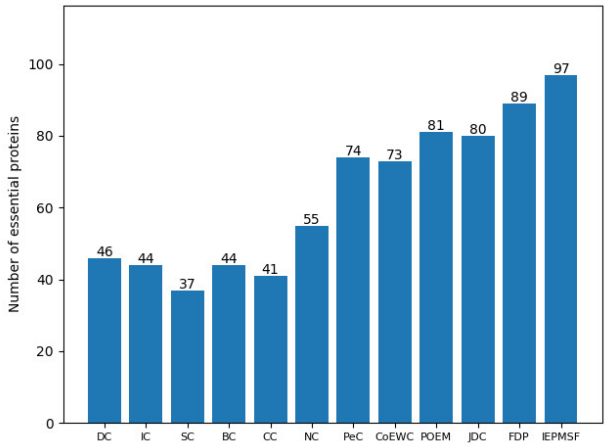
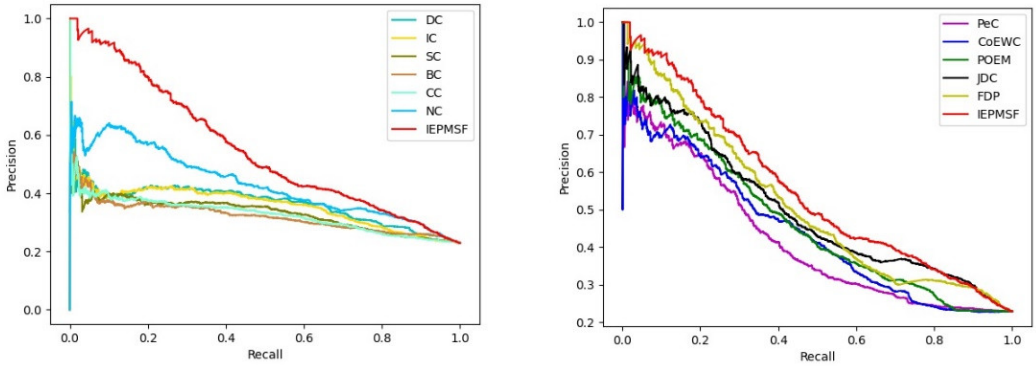


Figure 2. Number of true essential proteins predicted by DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, JDC, FDP and IEPMSF, when top 100 ranked proteins as candidates and $a = 0.1$.

When essential protein candidates with higher scores at different ratios (top 100, 200, 300, 400, 500, and 600) are chosen, their highest values are 97% ($a = 0.1$), 86% ($a = 0.3$), 79.7% ($a = 0.2$), 74.5% ($a = 0.2$), 69.8% ($a = 0.2$) and 65.3% ($a = 0.2$) respectively. The maximum level of accuracy is centered at $a = 0.2$ as the number of candidate proteins grows. Therefore, we set a as 0.2 to carry out the following experiments.

3.1.2. The precision-recall curve (PR curve) analysis predicted by various methods

The PR curve is applied to further validate the capability of the various approaches. Firstly, according to the final scores computed by each technique, proteins in the protein interaction network are sorted in descending order. The preceding K proteins are considered essential proteins (positive dataset), whereas the remaining proteins are considered non-essential proteins (negative dataset), where the threshold K ranges from 1 to 5093. As the K values be changed, to produce the PR curve, the corresponding precision and recall values for each approach are computed, as illustrated in Figure 3. The PR curves of IEPMSF are compared with PR curves of centrality algorithms (DC, IC, CC, BC, SC, and NC) and of multi-source information fusion methods (PeC, CoEWC, POEM, JDC, and FDP) in Figure 3(a) and (b) respectively. As seen in Figure 3, the PR curve of the IEPMSF has much higher value than that of other algorithms.

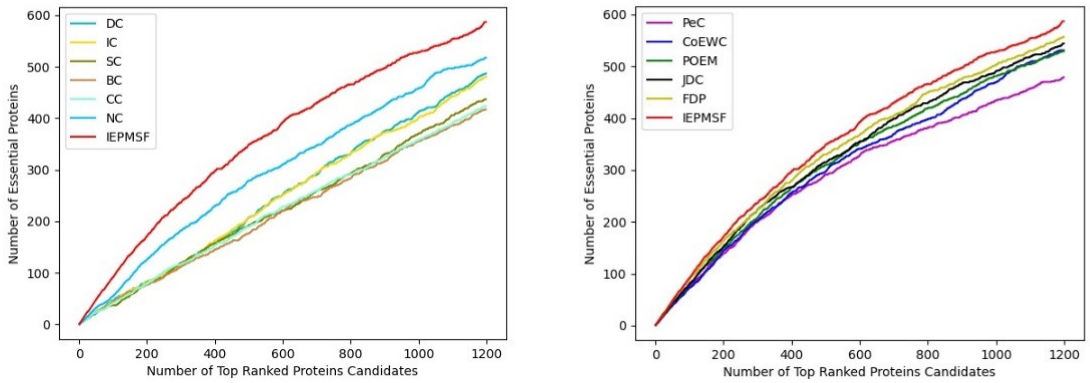


(a) (b)

Figure 3. Compared IEPMSF with other eleven approaches from the point of PR curves. (a) Curves of DC, IC, SC BC, CC, NC and IEPMSF; (b) Curves of PeC, CoEWC, POEM, JDC, FDP and IEPMSF.

3.1.3. The jackknife curve analysis predicted by various methods

To further examine the prediction performance of IEPMSF and other approaches, we apply the jackknife method. Figure 4 depicts the experimental outcomes. The number of putative essential proteins ranked first by each approach is represented on X-axis and the real number of important proteins found is represented on Y-axis. Performance of each method is compared in the area below the center line. Figure 4(a) demonstrate the outcome of a comparison between DC, IC, CC, BC, SC, NC and IEPMSF. From Figure 4(a), we see that the IEPMSF prediction of essential proteins is significantly more accurate than that of NC. Figure 4(b) shows the comparison of IEPMSF and existing methods based on multi-source information fusion (PeC, CoEWC, POEM, JDC and FDP). According to all of the experimental data. the accuracy of IEPMSF in predicting essential proteins is greater than the other 11 approaches, according to all of the experimental data.



(a) (b)

Figure 4. Compared IEPMSF with other eleven approaches from the point of jackknife curves. (a) Curves of DC, IC, SC BC, CC, NC and IEPMSF; (b) Curves of PeC, CoEWC, POEM, JDC, FDP and IEPMSF.

4. Conclusions and discussion

The essential proteins identifying is not only a prerequisite in comprehending organism survival, but it is also critical for the discovery of disease-causing genes and possible therapeutic targets. An essential proteins identification model IEPMSF was designed in this paper. In order to avoid more noise caused by multi-source data integration, to build the weighted network, the model only uses the common neighbor topology properties of the nodes in the network from original PPI data. Considering the issue of false positive and false negative PPI data caused by high-throughput trials, and the clustering function of NMTF, the weighted network was optimized using the non-negative matrix symmetric tri-factorization (NMSTF) technique to uncover probable protein-protein interactions. Finally, the starting score of each protein node was calculated using the subcellular location and homologous proteins information, and the restart random walk method was used to score and rank each protein in the network. Compared with the topological centrality method and the traditional multi-source information integration method, the experimental findings reveal that the suggested essential proteins prediction approach, IEPMSF, significantly improves the performance of essential proteins prediction. On the basis of the existing work, how to design a more effective method to construct a weighted network based on multi-source information integration is the future research direction of essential proteins identification. In long term, we will investigate including more biological data during the weighted network construction step, and try to apply the model to other species.

Acknowledgments

This project is partially funded by the National Natural Science Foundation of China (61772089, 62006030), Natural Science Foundation of Hunan Province (2020JJ4648), Major Scientific and Technological Projects for collaborative prevention and control of birth defects in Hunan Province (2019SK1010).

Conflict of interest

The authors declare no competing interests.

References

1. M. Li, R. Zheng, H. Zhang, J. Wang, Y. Pan, Effective identification of essential proteins based on priori knowledge, network topology and gene expressions, *Methods*, **67** (2014), 325–333. https://doi.org/10.1016/j.ymeth.2014.02.016
2. M. W. Hahn, A. D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol. Biol. Evol.*, **22** (2005), 803–806. https://doi.org/10.1093/molbev/msi072
3. K. Björnsdottir, Language, research and nursing practice, *J. Adv. Nurs.*, **33** (2001), 159–166. Available from: https://pubmed.ncbi.nlm.nih.gov/11168697/.
4. S. Wuchty, P. F. Stadler, Centers of complex networks, *J. Theor. Biol.*, **223** (2003), 45–53. https://doi.org/10.1016/S0022-5193(03)00071-7

5. E. Estrada, J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, *Phys. Rev. E.*, **71** (2005), 056103. https://doi.org/10.1103/PhysRevE.71.056103
6. M. P. Joy, A. Brock, D. E. Ingber, S. Huang, High-betweenness proteins in the yeast protein interaction network, *Biomed. Res. Int.*, **2005** (2005), 96. https://doi.org/10.1155/JBB.2005.96
7. J. Wang, M. Li, H. Wang, Y. Pan, Identification of essential proteins based on edge clustering coefficient, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **9** (2012), 1070–1080. https://doi.org/10.1109/TCBB.2011.147
8. X. Zhang, J. Xu, W. Xiao, A new method for the discovery of essential proteins, *PLoS One*, **8** (2013), e58763. https://doi.org/10.1371/journal.pone.0058763
9. M. Li, H. Zhang, J. Wang, Y. Pan, A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data, *BMC Syst. Biol.*, **6** (2012), 15. https://doi.org/10.1186/1752-0509-6-15
10. B. Zhao, J. Wang, M. Li, F. Wu, Y. Pan, Prediction of essential proteins based on overlapping essential modules, *IEEE Trans. Nanobioscience*, **13** (2014), 415–424. https://doi.org/10.1109/TNB.2014.2337912
11. F. Zhang, W. Peng, Y. Yang, W. Dai, J. Song, A novel method for identifying essential genes by fusing dynamic protein–protein interactive networks, *Genes*, **10** (2019), 31. https://doi.org/10.3390/genes10010031
12. J. Zhong, C. Tang, W. Peng, M. Xie, Y. Sun, Q. Tang, et al., A novel essential protein identification method based on PPI networks and gene expression data, *BMC Bioinf.*, **22** (2021), 248. https://doi.org/10.1186/s12859-021-04175-8
13. C. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2006), 126–135. https://doi.org/10.1145/1150402.1150420
14. A. Hassani, A. Iranmanesh, N. Mansouri, Text mining using nonnegative matrix factorization and latent semantic analysis. *Neural Comput. Appl.*, **33** (2021), 13745–13766. https://doi.org/10.1007/s00521-021-06014-6
15. Z. Khan, N. Iltaf, H. Afzal, H. Abbas, Enriching non-negative matrix factorization with contextual embeddings for recommender systems, *Neurocomputing*, **380** (2020), 246–258. https://doi.org/10.1016/j.neucom.2019.09.080
16. Y. Qing, C. Jun, N. AI-Nabhan, Data representation using robust nonnegative matrix factorization for edge computing, *Math. Biosci. Eng.*, **19** (2022), 2147–2178. https://doi.org/10.3934/mbe.2022100
17. Y. Qiu, W. Ching, Q. Zou, Matrix factorization-based data fusion for the prediction of RNA-binding proteins and alternative splicing event associations during epithelial-mesenchymal transition, *Briefings Bioinf.*, **22** (2021), bbab332. https://doi.org/10.1093/bib/bbab332
18. Y. Man, G. Liu, Y. Kuo, X. Zhou, SNFM: A semi-supervised NMF algorithm for detecting biological functional modules, *Math. Biosci. Eng.*, **16** (2019), 1933–1948. https://doi.org/10.3934/mbe.2019094
19. N. Pržulj, D. A. Wigle, I. Jurisica, Functional topology in a network of protein interactions, *Bioinformatics*, **20** (2004), 340–348. https://doi.org/10.1093/bioinformatics/btg415
20. B. Zhao, S. Hu, X. Li, F. Zhang, Q. Tian, W. Ni, An efficient method for protein function annotation based on multilayer protein networks, *Hum. Genomics*, **10** (2016), 33. https://doi.org/10.1186/s40246-016-0087-x

21. G. T. Hart, I. Lee, E. M. Marcotte, A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality, *BMC Bioinf.*, **8** (2007), 236. https://doi.org/10.1186/1471-2105-8-236
22. G. Li, M. Li, J. Wang, J. Wu, F. Wu, Y. Pan, Predicting essential proteins based on subcellular localization, orthology and PPI networks, *BMC Bioinf.*, **17** (2016), 279. https://doi.org/10.1186/s12859-016-1115-5
23. H. W. Mewes, D. Frishman, K. F. X. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, et al., MIPS: analysis and annotation of proteins from whole genomes in 2005, *Nucleic Acids Res.*, **34** (2006), D169-D172. https://doi.org/10.1093/nar/gkj148
24. J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, et al., SGD: Saccharomyces genome database, *Nucleic Acids Res.*, **26** (1998), 73–79. https://doi.org/10.1093/nar/26.1.73
25. R. Zhang, Y. Lin, DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes, *Nucleic Acids Res.*, **37** (2009), D455–D458. https://doi.org/10.1093/nar/gkn858
26. W. Peng, J. Wang, W. Wang, Q. Liu, F. Wu, Y. Pan, Iteration method for predicting essential proteins based on ontology and protein-protein interaction networks, *BMC Syst. Biol.*, **6** (2012), 87. https://doi.org/10.1186/1752-0509-6-87
27. I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S. Kim, D. Eisenberg, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.*, **30** (2002), 303–305. https://doi.org/10.1093/nar/30.1.303
28. J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, et al., COMPARTMENTS: unification and visualization of protein subcellular localization evidence, *Database*, **2014** (2014), bau012. https://doi.org/10.1093/database/bau012
29. J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse, *Nucleic Acids Res.*, **40** (2012), D881–D886. https://doi.org/10.1093/nar/gkr974
30. M. Magrane, UniProt Consortium, UniProt Knowledgebase: a hub of integrated protein data, *Database*, **2011** (2011), bar009. https://doi.org/10.1093/database/bar009
31. T. W. Harris, I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W. J. Chen, et al., WormBase: a comprehensive resource for nematode research, *Nucleic Acids Res.*, **38** (2010), D463–D467. https://doi.org/10.1093/nar/gkp952
32. P. McQuilton, S. E. St. Pierre, J. Thurmond, the FlyBase Consortium, FlyBase 101–the basics of navigating FlyBase, *Nucleic Acids Res.*, **40** (2012), D706–D714. https://doi.org/10.1093/nar/gkr1030
33. G. Östlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, et al., InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, *Nucleic Acids Res.*, **38** (2010), D196–D203. https://doi.org/10.1093/nar/gkp931

AIMS Press