Identification of Essential Proteins based on Non-negative Matrix Factorization

Zhihong Zhang
Changsha University
Sai Hu
Changsha University
Wei Yan
Changsha University
Bihai Zhao (Bihaizhao@163.com)
Changsha University
Lei Wang
Changsha University

Research Article

Keywords: Matrix Factorization, Protein-protein interaction, Essential protein

Posted Date: January 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1237007/v1

Identification of Essential Proteins based on Non-negative Matrix Factorization

Zhihong Zhang[1,2], Sai Hu[1], Wei Yan[1], Bihai Zhao[1,2*], Lei Wang[1,2]

[1]College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, Hunan 410022, China

[2]Hunan Provincial Key Laboratory of Nutrition and Quality Control of AquaticAnimals,Department of Biological and Environmental Engineering, Changsha University, Changsha, Hunan 410022, China

[*]Corresponding author: Bihai Zhao (Bihaizhao@163.com)

Email addresses:
ZHZ: zhzhang@ccsu.edu.cn
SH:husaiccsu@163.com
WY: 1150273378@qq.com
BHZ: bihaizhao@163.com
LW: wanglei@xtu.edu.cn

Abstract

Background

Identification of essential proteins will provide valuable information for medicine and other related disciplines, especially for the diagnosis and treatment of diseases and drug design. Various different computational methods have been proposed to identify essential proteins based on protein-protein interaction (PPI) networks. However, there has been reliable evidence that a huge amount of false negatives and false positives exist in PPI data. Therefore, it is necessary to reduce the influence of false data on accuracy of essential proteins prediction by integrating multi-source biological information with PPI networks.

Results

In this paper, we proposed a non-negative matrix factorization and multiple biological information based model (NDM) for identifying essential proteins. The first stage in this progress was to construct a weighted PPI network by combing the information of protein domain, protein complex and the topology characteristic of the original PPI network. Then, the non-negative matrix factorization technique was used to reconstruct an optimized PPI network with whole enough weight of edges. In the final stage, the ranking score of each protein was computed by the PageRank algorithm in which the initial scores were calculated with homologous and subcellular localization information. In order to verify the effectiveness of the NDM method, we compared the NDM with other state-of-the-art essential protein prediction methods. The comparison of the results obtained from different methods indicated that our NDM model has better

performance in predicting essential proteins.

Conclusion

Employing the non-negative matrix factorization method and integrating multi-source biological data can effectively improve quality of the PPI network, which resulted in the led to optimization of the performance essential protein identification. This will also provide a new perspective for other prediction based on protein-protein interaction networks.

Keywords: Matrix Factorization, Protein-protein interaction, Essential protein

Background

Essential genes and their products (essential proteins) are necessary for the survival of the organism, whose functions are thought to be the basis of life. The identification of essential proteins can help us understand the basic requirements for sustaining life forms. Besides that, it plays an important role in the emerging field of synthetic biology, which aims to create a cell with the smallest genome [1]. It can also provide important reference information for biology, medicine and other disciplines [2]. At present, a variety of methods have been proposed to identify essential proteins through biological experiments, such as single-knock-out [3] and RNA interference (RNAi) [4]. Nonetheless, there are some limitations with these experimental methods, for instance high cost and time consuming. Although large-scale experimental techniques for identifying essential proteins have been greatly improved, there is still a large gap between computational methods for detecting essential proteins and genome sequences.

In recent years, many computational methods have been proposed to identify essential proteins. Based on topological features of PPI network, many centrality models have been proposed to predict essential proteins including Degree Centrality (DC) [5], Information Centrality (IC) [6], Closeness Centrality (CC) [7], Betweenness Centrality (BC) [8], Subgraph Centrality (SC) [9] and sum of Edge Clustering Coefficient Centrality (NC) [10]. Singh et al. [11] proposed the Graph Fourier Transform Centrality (GFT-C) to quantify the importance of nodes in complex networks. Wang et al. [12] designed a new efficiency centrality (EffC) sorting algorithm, which identified

influential nodes by considering the change degree of the whole network efficiency after deleting each node respectively. Li et al. [13] found that the frequency of essential proteins appearing in triangular structures was significantly higher than that of non-essential proteins and proposed a new pure centricity measure named Neighborhood close centricity (NCC). However, most of these methods make use of the topological characteristics of essential proteins, whose performance is largely dependent on the reliability of the PPI network.

In order to make up for the limitations of incompleteness of PPI networks, many research groups have combined PPI networks with other biological information to improve the accuracy of essential protein identification. Tew et al. [14] integrated the functional information of proteins and network topology attributes when designing essential protein identification methods. Zhang et al. [15] combined network topology characteristics, gene expression information and GO annotation information and proposed an essential protein prediction method called TEO. Lei et al. [16] put forward an essential protein identification method called RWEP by integrating gene expression profile, functional annotation information with PPI networks. Ren et al. [17] proposed a new centrality measure to identify essential proteins by combining PPI network topology and protein complex information, due to the stronger clustering between protein complexes and essential proteins. By integrating topological features of PPI networks and gene expressions information, Zhang et al. and Li et al. proposed two different models to predict essential proteins, named CoEWC [18] and PeC [19] respectively. Considering the modularity of essential proteins, Zhao et al. proposed an

essential protein identification method called POEM [20] in 2014. The ION method [21] combined homogeneous information of proteins and PPI networks for essential proteins discovery, which is predicated on the fact that essential proteins are more conserved than nonessential proteins. Peng et al. designed an essential protein detection method called UDoNC[22] after analyzing the relationships between essential proteins and domains in 13 species. According to the similarities of active PPI networks of each time, Zhang et al. [23] established a novel PPI network. Based on the constructed network and orthologous information of proteins, they developed a dynamic protein-protein interactive networks-based model called FDP to discover essential proteins. In spite of significant advances in network-based essential protein prediction methods, it still remains a challenge to effectively improve the PPI network quality with multi-omics data integration and enhance the performance of essential protein identification.

Here, we developed a novel prediction model called NDM for the task of essential proteins discovery. NDM takes a full account of the protein-protein interactions in PPI network and other multi-source biological information, such as protein complex, protein domain, homologous and subcellular localization information. Specially, the matrix factorization technique is employed in the NDM to construct a more tolerant of false negative protein interaction network than that of other methods. On that basis, we adopt the PageRank algorithm to score and rank all proteins. The comparative performance experiments were conducted for NDM and other state-of-the-art methods by using yeast data sets. The experimental results indicated that NDM obtains better

performance compared with other methods and can be effectively applied in the discovery of essential proteins.

Methods

The NDM method consists of three steps. (1) Constructing multiple networks to represent the complex relationships among proteins from topological features of PPI network, protein complex and domain information, and integrating into a reliable weighted network. (2) Reconstructing a comprehensive protein interactome network by using the non-negative matrix factorization technique to discover potential protein interactions from the weighted network. (3) Scoring and ranking proteins through random walk on the above comprehensive network.

Construction of a reliable weighted network based on multi-source biological data

To reduce the negative impact of false positives on predict performance, we transform the original PPI network to a reliable weighted PPI network (*rPPI*) by combining the topological features of original PPI networks and multi-source biological data, such as protein complexes and protein domains information. Here, three types of associations between nodes have been built and denoted as *Neighbor_PPI*, *Domain_PPI* and *Complex_PPI* respectively. The formation of these will be described in detail below.

Due to the limitation of high throughput technology, a significant proportion of experimental PPI data contains errors. Many prediction algorithms [24] explore common neighbors between pairs of proteins in PPI networks to measure the reliability of the interactions between proteins. Obviously, the more common neighbors between

two proteins, the more likely they will interact with each other. In this paper, the proteins p_i and p_j would be considered to be interconnected in the *Neighbor_PPI* network, if they have at least one common neighbor. This kind of connections between proteins is the first type of relationships in the NDM, whose reliability can be calculated as follows:

$$Neighbor\_PPI(p_i, p_j) = \begin{cases} \dfrac{\left|NS_i \cap NS_j\right|^2}{(\left|NS_i\right|-1)*(\left|NS_j\right|-1)} & if\ \left|NS_i\right|>1\ and\ \left|NS_j\right|>1 \\ 0 & otherwise \end{cases} \quad (1)$$

Where NS_i and NS_j denotes the neighborhood sets of p_i and p_j respectively.

Proteins are usually composed of one or more domains that have independent functions. Researchers [15] studied the relationship between essential proteins and their domain composition and found that the number of protein domain types is closely related to their importance. Especially, they also pointed out that more types of domains would appear in essential proteins, while non-essential proteins contained fewer types of domains. From this, we can think that the proteins p_i and p_j would be considered to be interconnected in the *Domain_PPI* network, if they have the same type of domain. The weight of this kind of connection between proteins can be calculated as follows:

$$Domain\_PPI(p_i, p_j) = PD(p_i) \times PD(p_j) \quad (2)$$

Here, $PD(p_i)$ is the domain score of the protein p_i and independent of each other, which can be calculated as follows:

$$PD(p_i) = \frac{\sum_{j=1}^{|D|} \frac{1}{NP_j} \times t_{ij} - \min_{1\le k\le |V|} (\sum_{j=1}^{|D|} \frac{1}{NP_j} \times t_{kj})}{\max_{1\le k\le |V|} (\sum_{j=1}^{|D|} \frac{1}{NP_j} \times t_{kj}) - \min_{1\le j\le |V|} (\sum_{j=1}^{|D|} \frac{1}{NP_j} \times t_{kj})} \quad (3)$$

Where V is the set of proteins, $D=\{d_1, d_2, \ldots, d_m\}$ is the set of different domains in all proteins and NP_j is the number of proteins with domain d_j. If the protein contains the domain d_j, the value of $t_{ij}$is set to 1 or else to 0.

Protein complex is composed of a variety of proteins that interact with each other. It is the basic unit of macromolecular organization and plays a crucial role in the execution of cellular functions. Many proteins perform life functions only after they are assembled in protein complexes and interact with other proteins. Studies [21] have demonstrated that if two proteins are presented in the same protein complex, they tend to have the same or similar biological functions. Therefore, the weight of interactions in the *Complex_PPI* network is measured by the number of shared protein complexes between pairs of proteins:

$$Complex\_PPI(p_i, p_j) = \frac{\left|CS_{p_i} \cap CS_{p_j}\right|}{\left|CS_{p_i}\right| \times \left|CS_{p_j}\right|} \tag{4}$$

In this formula, CS_{pi} and CS_{pj} is the set of protein complexes which contains p_i and p_j respectively. The numerator represents the common set of protein complexes which p_i and p_j belong to.

After constructing the above three kinds of protein interaction networks, we integrate these three networks and get a weighted PPI network (*rPPI*) by equation 5.

$$\begin{aligned} rPPI(p_i, p_j) &= Neighbor\_PPI(p_i, p_j) \\ &\quad + Domain\_PPI(p_i, p_j) + Complex\_PPI(p_i, p_j) \end{aligned} \tag{5}$$

The Figure 1, shown below, summarizes the overall transform process.

Reconstruction of a comprehensive protein interactome network based on NMF

The second stage of our NDM method is to exploit potential associations between proteins from the above reliable weighted network using the Non-negative matrix factorization (NMF) technology. As an effective data representation technique, the NMF has been widely used in lncRNA-disease associations prediction [25], conserved functional modules detection [26], etc. For our purpose, we represent the reliable weighted network constructed in the first stage as an adjacency matrix $rPPI \in R^{N\times N}$. In this work, we wish to establish a new matrix $Y \in R^{N\times N}$, in which enough elements of *Y* are filled with computed values, like:

$$rPPI \approx Y = WH \tag{6}$$

where $W \in R^{N\times C}$ represents a low rank matrix, *C* denotes the number of selected features and its value far less than *N*. $H \in R^{C\times N}$is the coefficient matrix. For a given non-negative data matrix *rPPI*, the issue can be solved as the following optimization problem:

$$\min J\left(W,H\right)=\frac{1}{2}\left\|rPPI-WH\right\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. Since the objective function in Equation (7) is a joint non-convex problem, we employ the rule of multiplicative iteration to solve the objective function on the basis of using auxiliary functions. Squared Frobenius norm can be written as $\|A\|^2 = Tr(A^TA)$, therefore equation(7) equals to:

$$F=\min J\left(W,H\right)=\frac{1}{2}Tr\left(rPPI^T * rPPI-2rPPI^TWH+H^TW^TWH\right) \tag{8}$$

Its partial derivatives equations for factor *W*, *H* respectively as follow:

$$\frac{\partial F}{\partial W} = -rPPI * H^T + WHH^T$$
$$\frac{\partial F}{\partial H} = -rPPI * W^T + W^T WH \tag{9}$$

The static point can be found by the Karush-Kuhn Tucker (KKT) complementarity conditions. The KKT condition for factor W is as follows:

$$\frac{\partial F}{\partial W_{iu}} W_{iu} = 0 \tag{10}$$

In this respect, the conditions is assumed to be at work if the derivative, $\frac{\partial F}{\partial W}$, is zero:

$$\left(-rPPI * H^T + WHH^T\right)_{iu} W_{iu} = 0$$
$$W_{iu} = W_{iu} \frac{\left(rPPI * H^T\right)_{iu}}{\left(WHH^T\right)_{iu}} \tag{11}$$

Similarly, the second update rule for V can be derived as follow:

$$\frac{\partial F}{\partial H_{uj}} H_{uj} = 0$$
$$\left(-rPPI * W^T + W^T WH\right)_{uj} H_{uj} = 0$$
$$H_{uj} = H_{uj} \frac{\left(W^T * rPPI\right)_{uj}}{\left(W^T WH\right)_{uj}} \tag{12}$$

As described above, the multiplication iteration rules are as follows:

$$W_{iu} \leftarrow W_{iu} \frac{\left(rPPI * H^T\right)_{iu}}{\left(WHH^T\right)_{iu}}$$
$$H_{uj} \leftarrow H_{uj} \frac{\left(W^T * rPPI\right)_{uj}}{\left(W^T WH\right)_{uj}} \tag{13}$$

From the above equation (6-13), we can obtain an optimal matrix Y that is closest to *rPPI*. To restore the symmetry of the protein-protein interactions, the matrix Y finally need to be transformed to a symmetrical transition probability matrix *cPPI* as follow:

$$cPPI_{ij} = \begin{cases} \dfrac{\max\left(Y_{ij}, Y_{ji}\right)}{\sum_{k=0}^{N} Y_{ik}}, & \sum_{k=0}^{N} Y_{ik} \neq 0 \\ 0, & \text{else} \end{cases} \tag{14}$$

Obviously, after going through the above process, a new network *cPPI* can be set up accordingly.

Random walk on the comprehensive protein interactome network

The last stage of our NDM method is scoring proteins and generating candidate essential proteins. In this work, we run a random walk with restart (RWR) algorithm on the comprehensive protein interactome network to rank all proteins. As a part of the RWR algorithm, we first define the initial vector *IS* according to the conservative and functional features of essential proteins, which derived from homologous proteins and subcellular localization information respectively. Tang et al.[27] analysis whether all proteins in the *Saccharomyces cerevisiae* PPI network have direct homologous proteins in 99 reference species.They manifest that the more homologous proteins a protein has in reference species, the more likely it is to be the essential protein. For a given protein p_i in the comprehensive network *cPPI*, $HS(p_i)$ indicates its conservative score and can be calculated as follow [28]:

$$HS(p_i) = \frac{H(p_i)}{\max_{1 \le j \le |V|}(H(p_j))} \tag{15}$$

where $H(p_i)$ represents the number of times that protein p_i has orthologous proteins in reference organisms.

Proteins are localized at their appropriate subcellular locational compartments to

perform their biological functions. Researchers [29-30] study the relationship subcellular locational properties of proteins and find that essential proteins appear more often in specific subcellular locations. For this reason, we use the subcellular localization information to calculate functional scores of proteins (*FS*). If the protein p_i exists in the final subcellular localization dataset *R*, the occurrence frequency of each subcellular locations *r* can be calculated by the following equation:

$$OF_{p_i}(r) = \begin{cases} \dfrac{|S(r)|}{\max\limits_{1 \le k \le n}(|S(k)|)} & , \quad if \quad p_i \cap R \\ 0 & , \quad otherwise \end{cases} \tag{16}$$

where $S(r)$ represents the set of proteins corresponding to subcellular locations *r* and *n* denotes the number of whole subcellular locations related to all proteins.

For a given protein p_i, its functional score $FS(p_i)$ is defined as the max scores of all the subcellular locations in which it appears.

$$FS(p_i) = \max_{r \in C(p_i)}(OF_{p_i}(r)) \tag{17}$$

where $C(p_i)$ is the subcellular locations set corresponding to the protein p_i.

Finally, base on the equation (15-17), the unique initial score of p_i, $IS(p_i)$, can be defined as follow:

$$IS(p_i) = HS(p_i) \times FS(p_i) \tag{18}$$

To obtain stable and global scores for proteins, we run the random walk algorithm with restart model on the reconstructed comprehensive protein interactome network *cPPI*. After normalized, the *cPPI* network can be represented as an $n \times n$ adjacency matrix *MC*. Then, for a protein p_i in the network, its ranking score $RS(p_i)$ can be calculated by equation (19) as follow:

$$RS(p_i) = \lambda \sum_{p_j \in NS_i} MC(i,j) RS(p_j) + (1-\lambda)\ IS(p_i) \quad (19)$$

where NS_i is the set of neighbors of protein p_i and the parameter λ ($0 \leq \lambda < 1$) is used to balance the iteration information and initial scores. From the above definition, we can see that the ranking score of a protein can be regarded as a linear combination of its initial scores and neighbor correlation scores of edges in *cPPI* network. Equation (19) can be rewrite in the matrix-vector format as follows:

$$RS = \lambda * MC * RS + (1-\lambda) * IS \quad (20)$$

In order to solve equation (19), the Jacobi iterative procedure is used in our this work as follows:

$$RS^t = \lambda * MC * RS^{t-1} + (1-\lambda) * IS \quad (21)$$

where RS^t is the ranking score vector obtained in t-th iteration.

Based on above description, the overall framework of the NMF-based model for identifying essential proteins (NDM) can be descripted as the following Algorithm 1.

Algorithm 1: NDM

Input: A PPI network G, protein domains, protein complexes, subcellular localizations, homologous proteins, stopping error ε, parameters α, β and λ

Output: Top K proteins sorted by RS indescending order

Step 1: Construct the *Neighbor_PPI*, *Domain_PPI* and *Complex_PPI* according to Equation (1)-(4);

Step 2: Construct a reliable weighted network *rPPI* according to Equation(5);

Step 3: Reconstruct the comprehensive network *cPPI* by Equation (6)-(13);

Step 4:Initialize initial vector IS with $RS^0=IS$ and t=0;

Step 5: Compute RS^t according equation (21) and set t =t+1;

Step 6. If $\|RS^t-RS^{t-1}\|<\varepsilon$, then $RS=RS^t$ and terminate the algorithm. Otherwise, let t=t+1 and repeat Step 5;

Step 7. Sort proteins by the value of RS in the descending order

Step 8. Output top K of sorted proteins

Results and Discussion

The study uses qualitative analysis in order to assess the performance of NDM in predicting essential proteins. Here, other eleven state-and-art methods are used for comparison, such as DC [5], IC [6], CC [7], BC [8], SC [9], NC [10], PeC [19], CoEWC [18], POEM [20], ION [21] and the newly proposed method NTMEP [31]. Based on the authoritative list of essential proteins, we mainly analyze the accuracy of predicting essential proteins in top 100, 200, 300, 400, 500 and 600 of the ranked proteins obtained from all methods respectively. Turning now to the experimental evidence on the effectiveness of our NDM method.

3.1 Experimental data

The protein-protein interaction data are mainly concentrated in yeast, because this species has good knockout experimental characteristics, the data are the most complete and convincing. In our experiment, data related to essential protein identification mainly include benchmark essential protein dataset, PPI data and multi-source biological data. The benchmark essential proteins collected by experimental methods are mainly derived from four databases: MIPS [32], SGD [33], DEG [34], and SGDP [35], including 1,285 essential proteins. In this work, DIP data [36] is used to assess the effectiveness of our proposed method, which contains 5,093 proteins interacting with 24,743 interacted groups and 1,167 essential proteins. The commonly used data of protein complexes are CYC2008[37] and MIPS[32], including 408 and 428 complexes detected by biological experiments, respectively. The protein domains information is derived from PFAM 25.0 database [38] and contains 2,671 different types of domains.

Subcellular location data is obtained from COMPARTMENTS [39] database. To avoid specificity of data, 11 categories of sub-cellular localization are reserved, including Endoplasmic, Nucleus, Cytoskeleton, Golgi, Cytosol, Vacuole, Plasma, Mitochondrion, Endosome, Peroxisome and Extracellular [40]. For homologous protein information, we get it from the 7th edition of InParanoid database [41] which contains paired comparisons of 100 whole genomes (99 eukaryotes and 1 prokaryote).

3.2 Parameter sensitivity analysis

In terms of the computational algorithm for predicting essential proteins, the optimal parameters selection may differ from one experiment scenario to another. In this section, we mainly focus on the parameter λ for NDM. It is set to 0, 0.1, 0.2, ...,0.9 and 1, and is used to balance the iteration information and initial scores in the ultimate ranking score as described in equation (19-21). Table 1 illustrates the impact of the parameter λ on the performance of NDM. As for λ, we vary its value from 0 to 1 and the number of ranked proteins is top 100, 200, 300, 400, 500 and 600 respectively. The accuracy of the prediction is measured based on the percentage of true essential proteins in the candidate proteins. As can be seen from the table, the best prediction results are obtained when the value of λ is set to 0.3 and 0.4. Especially, for the top 100 and top 200, the best predicted accuracy (93% and 88% respectively) is achieved when the value of λ is assigned as 0.3. Therefore, the optimum λ value in this work is 0.3.

3.3 Comprehensive comparison with other methods

In this section, the comprehensive comparisons of NDM and other methods are carried out to demonstrate the effectiveness of our proposed prediction method. There are various top numbers of ranked proteins selected as candidate essential proteins after all protein scores are calculated based on each method. As can be seen from the Figure 2, the top100, 200, 300, 400, 500 and 600 of proteins are taken from the ranked results of those twelve methods as candidates, in which each protein is distinguished as essential proteins or not.

From Figure 2 we can see that there is a significant improved accuracy of identifying essential proteins by our proposed method, while contrasted with other eleven methods. When the top 100 to top 600 proteins selected as candidates, we can see that the NDM resulted in the higher values of percentage improvement respectively to 69.1%, 39.7%, 33.5%, 30.9%,26.5%, 28.4% than the NC which obtains the best results from the classical networks topology-based centrality methods such as DC, IC, BC, CC, SC and NC. Compared with other multi-source based predict methods (CoEWC, PeC, POEM, ION and NTMEP), NDM still achieves a full-scale improvement. One of the most significant improvement in results is that, when essential candidates are selected with top100 to top 600 respectively, the accuracies of our method improved by 1.09%, 2.92%, 2.53%, 2.03%, 1.73% and 1.79% than NTMEP which has the best performance among the other multi-source based predict methods.

3.4 Validated by precision-recall curves

Furthermore, we plot the precision-recall (PR) curve at different cutoffs to assess the performance of each method. After scores of all proteins are calculated with each method, we select top *k* proteins as the essential candidates (positive set), and others as the non-essential candidates (negative set) in descending order. Here, the value of *k* is set from 1 to 5093 (the total number of proteins). The values of recall and precision can be computed by using each method and are reported at different cutoffs respectively. The PR curves of NDM and other eleven methods are shown in Figure 3. Figure 3(a) presents the comparison of NDM with six centrality methods only based on the topology feature of PPI network, including DC, IC, BC, CC, SC and NC. Figure 3(b) shows the results obtained from the preliminary analysis of NDM and five multi-source based methods, including PeC, CoEWC, POEM, ION and NTMEP. As can be seen from the figure, the NDM group reports significantly better performance than the othereleven groups. Especially, the identification rate of NDM is 100% for the first 39 essential candidates. This is a remarkable result that cannot be accomplished by any other competitive methods.

3.5 Validated by jackknife methodology

In this subsection, the simulation is conducted to utilize the jackknife methodology to compare our NDM method with other state-of-the-art methods (DC, BC, CC, SC, IC, NC, PeC, CoEWC, POEM, ION and NTMEP). The value of area under the jackknife curve can be computed to quantify the overall performance. Figure 4 plots the

jackknife curves of the competitive methods, in which the horizontal axis denotes the top number of essential candidates ranked in descending order with each method, and the vertical axis represents the number of essential proteins identified. In order to make the result clear, we separate the comparing result to three subgraph and select ten random assortments to compare. Figure 4(a) illustrates the jackknife curves of NDM, DC, IC, SC and ten random assortments. It is evident that NDM almost always has the highest value under the same value of essential candidates. The curves of NDM, BC, CC, NC and ten random assortments are shown in Figure 4(b) which presents that NDM achieves a higher precision with respect to all other competitive methods for any given number of essential candidates. Figure 4(c) shows the comparison result of NDM and other multi-source based methods (PeC, CoEWC, POEM, ION and NTMEP). As can be seen from the Figure 4(c), NDM has lower performance than ION for the ranked 900 and more essential candidates of predicted list, but still achieves the highest value than other methods. However, a higher precision obtained from NDM comparing to other multi-source based methods for the top portion of the predicted list. This portion of predicted results is important as it identifies the potential essential proteins predicted with high confidence.

Conclusions

As mentioned in the literature review, many computational methods have been developed to predict essential proteins based on PPI network or multi-source biological data and achieve good performance. However, these methods do not take full advantage of the relationship between multiple sources of data. This work set out to develop a

model for better performance, named NDM, which integrate PPI network, protein complexes, protein domains, subcellular localizations and homologous proteins information. To get the utmost out of multi-source data, non-negative matrix factorization is introduced into our proposed method. Also, a comprehensive experiment is carried out andthe results of this experiment show that our new method has the better performance than six topology-based centrality methods (DC, BC, CC, SC, IC and NC) and five multi-source based methods (PeC, CoEWC, POEM, ION and NTMEP). A possible explanation for these results might be that there are deep relationships between multiple sources of data. These results add to the rapidly expanding field of computational methods based on multi-source biological information. It is unfortunate that the study did not take other biological data into account. This is an important issue for future research.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

Publicly available datasets were analyzed in this study. This data and the NDM program can be found here: https://github.com/husaiccsu/NDM.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

ZHZ, BHZ obtained the protein-protein interaction data, benchmark essential protein dataset, domain data, the protein complexes information, subcellular location data and homologous protein information. ZHZ, SH and BHZ designed the new method, NDM, and analysed the results. ZHZ and BHZ drafted the manuscript together. WY and LW participated in revising the draft. All authors have read and approved the manuscript.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant No. 61772089, 61873221, 61672447, Natural Science Foundation of Hunan Province (2020JJ4648), Research Foundation of Education Bureau of Hunan Province (19A048, 18A441, 18C0958, 21A0548), Educational planning key project of Hunan Province(XJK18DJA1) and Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals(2018TP1027).

References

[1] Glass JI, Hutchison CA 3rd, Smith HO, Venter JC. **A systems biology tour de force for a near-minimal bacterium**. *Mol Syst Biol*. 2009, 5:330.

[2] Parrish JR, Yu J, Liu G, Hines JA, Chan JE, Mangiola BA, Zhang H, Pacifico S, Fotouhi F, DiRita VJ, Ideker T, Andrews P, Finley RL Jr. **A proteome-wide protein interaction map for Campylobacter jejuni**. *Genome Biol*. 2007, **8**(7):R130.

[3] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, *et al*,. **Essential Bacillus subtilis genes**. *Proc Natl Acad Sci U S A*. 2003,**100**(8):4678-83

[4] Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J. **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi**. *Nature*. 2003, **421**(6920):231-7.

[5] Hahn MW, Kern AD. **Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks**. *Mol Biol Evol*. 2005, **22**(4):803-6.

[6] Björnsdottir K. **Language, research and nursing practice**. *J Adv Nurs*. 2001, **33(**2):159-66.

[7] Wuchty S, Stadler PF. **Centers of complex networks**. J *Theor Biol*. 2003, **223**(1):45-53.

[8] Joy MP, Brock A, Ingber DE, Huang S. **High-betweenness proteins in the yeast protein interaction network**. *J Biomed Biotechnol*. 2005, **2005**(2):96-103.

[9] Estrada E, Rodríguez-Velázquez JA. **Subgraph centrality in complex networks**. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2005, **71**(5 Pt 2):056103.

[10] Wang J, Li M, Wang H, Pan Y. **Identification of essential proteins based on edge clustering coefficient**. *IEEE/ACM Trans Comput Biol Bioinform*. 2012, **9**(4):1070-80.

[11] Singh R, Chakraborty A, Manoj B S. **GFT centrality: A new node importance measure for complex networks**. *Physica A: Statistical Mechanics and its*

Applications. 2017, **487**(1): 185-95.

[12] Wang S, Du Y, Deng Y. **A new measure of identifying influential nodes: Efficiency centrality**. *Communications in Nonlinear Science and Numerical Simulation*. 2017, **47**: 151-63.

[13] Li G, Li M, Wang J, Li Y, Pan Y. **United Neighborhood Closeness Centrality and Orthology for Predicting Essential Proteins**. *IEEE/ACM Trans Comput Biol Bioinform*. 2020, **17**(4):1451-58.

[14] Tew KL, Li XL, Tan SH. **Functional centrality: detecting lethality of proteins in protein interaction networks**. *Genome Inform*. 2007, **19**:166-77.

[15] Zhang W, Xu J, Li Y, Zou X. **Detecting Essential Proteins Based on Network Topology, Gene Expression Data, and Gene Ontology Information**. *IEEE/ACM Trans Comput Biol Bioinform*. 2018, **15**(1):109-116.

[16] Lei X, Yang X, Fujita H. **Random walk based method to identify essential proteins by integrating network topology and biological characteristics**. *Knowledge-Based Systems*. 2019, **167**(1):53-67.

[17] Ren J, Wang JX, Li M, Wang H, et al, **Prediction of Essential Proteins by Integration of PPI Network Topology and Protein Complexes Information**. *Bioinformatics Research and Applications*. 2011, **6674**:12-24.

[18] Zhang X, Xu J, Xiao WX. **A new method for the discovery of essential proteins**. *PLoS One*. 2013, **8**(3):e58763.

[19] Li M, Zhang H, Wang JX, Pan Y. **A new essential protein discovery method based on the integration of protein-protein interaction and gene expression**

data. *BMC Syst Biol*. 2012, **6**:15.

[20] Zhao B, Wang J, Li M, Wu FX, Pan Y. **Prediction of essential proteins based on overlapping essential modules**. *IEEE Trans Nanobioscience*. 2014, **13**(4):415-24.

[21] Peng W, Wang J, Wang W, Liu Q, Wu FX, Pan Y. **Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks**. *BMC Syst Biol*. 2012, **6**:87.

[22] Peng W, Wang J, Cheng Y, Lu Y, Wu F, Pan Y. **UDoNC: An Algorithm for Identifying Essential Proteins Based on Protein Domains and Protein-Protein Interaction Network**s. *IEEE/ACM Trans Comput Biol Bioinform*. 2015, **12**(2):276-88.

[23] Zhang F, Peng W, Yang Y, Dai W, Song J. **A Novel Method for Identifying Essential Genes by Fusing Dynamic Protein**–**Protein Interactive Networks**. *Genes (Basel)*. 2019, **10**(1):31.

[24] Zhao B, Hu S, Li X, Zhang F, Tian Q, Ni W. **An efficient method for protein function annotation based on multilayer protein networks**. *Hum Genomics*. 2016, **10**(1):33.

[25] Fu G, Wang J, Domeniconi C, Yu G. **Matrix factorization-based data fusion for the prediction of lncRNA-disease associations.** *Bioinformatics*. 2018, **34**(9):1529-1537.

[26] Wang P, Gao L, Hu Y, Li F. **Feature related multi-view nonnegative matrix factorization for identifying conserved functional modules in multiple**

biological networks. *BMC Bioinformatics*. 2018, **19**(1):394.

[27] Tang X, Li X, Hu S, et al. **A framework for identifying functional modules in dynamic networks**. *International Journal of Data Mining and Bioinformatics*. 2018, **21**(1): 1-17.

[28] Zheng CH, Huang DS, Zhang L, Kong XZ. **Tumor clustering using nonnegative matrix factorization with gene selection**. *IEEE Trans Inf Technol Biomed*. 2009, **13**(4):599-607.

[29] Zhao B, Wang J, Li X, Wu FX. **Essential protein discovery based on a combination of modularity and conservatism**. *Methods*. 2016, **110**:54-63.

[30] Li G, Li M, Wang J, Wu J, Wu FX, Pan Y. **Predicting essential proteins based on subcellular localization, orthology and PPI networks**. *BMC Bioinformatics*. 2016, **17** (Suppl 8):279.

[31] Zhang Z , Jiang M, Wu D, et al. **A Novel Method for Identifying Essential Proteins Based on Non-negative Matrix Tri-Factorization**. *Frontiers in Genetics*, 2021: 1373.

[32] Mewes HW, Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stümpflen V. **MIPS: analysis and annotation of proteins from whole genomes in 2005**. *Nucleic Acids Res*. 2006, **34**(Database issue):D169-72.

[33] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. **SGD: Saccharomyces Genome Database**. *Nucleic Acids Res*. 1998, **26**(1):73-9.

[34] Zhang R, Lin Y. **DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes**. *Nucleic Acids Res*. 2009, **37**(Database issue):D455-8.

[35] Saccharomyces Genome Deletion Project, http://www-sequence.stanford.edu/group

[36] Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res*. 2002, **30**(1):303-5.

[37] Pu S, Wong J, Turner B, Cho E, Wodak SJ. **Up-to-date catalogues of yeast protein complexes**. *Nucleic Acids Res*. 2009, **37**(3):825-31.

[38] Bateman A, Coin L, Durbin R, et al. **The Pfam protein families database**. *Nucleic Acids Res*. 2004, **32**(suppl 1):D138–41

[39] Binder JX, Pletscher-Frankild S, Tsafou K, et al. **COMPARTMENTS: unification and visualization of protein subcellular localization evidence**. Database. 2014, **2014**:bau012.

[40] Peng,X, Wang J, Zhong,J, Luo J, and Pan Y, **An efficient method to identify essential proteins for different species by integrating protein subcellular localization information**. *In Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Washington, DC, USA, Nov. 2015: 277–280

[41] Gabriel O, Thomas S, Kristoffer F, et al. **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis**. *Nucleic Acids Res*. 2010, **38**(Database issue):D196–203.

Figure Legends

Figure 1- Illustration of the reliable weighted PPI network construction

This figure illustrates construction of the reliable weighted PPI network *r*PPI based on multiple networks, including the *Neighbor_PPI* network, *Domain_PPI* network and *Complex_PPI* network.

Figure 2- Number of correctly essential proteins identified by NDM and other eleven previously competitive methods.

Proteins are ranked in the descending order based on their ranking scores computed by NDM, DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, ION and NTMEP. Then, top 100, 200, 300, 400, 500 and 600 of the ranked proteins are selected as candidates for essential proteins. According to the list of benchmark essential proteins, the number of true essential proteins is used to judge the performance of each method. The figure shows the number of true essential proteins identified by each method.

Figure 3- Performance comparison between NDM and other eleven approaches in terms of PR curves.

The proteins ranked in top K (cut-off value) by each method (NDM, DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION and NTMEP) are selected as candidate essential proteins (positive data set) and the rest of proteins are regarded as candidate nonessential proteins (negative data set). With different values of K selected, the values of precision and recall are computed for each method. The values of precision and recall are plotted in PR curves with different cut-off values. (a) shows the PR curves of

NDM, DC, IC, SC, BC, CC and NC. (b) shows the PR curves of NDM and other five methods: CoEWC, PeC, POEM, ION and NTMEP.

Figure 4- Jackknife curves for NDM and other eleven approaches.

The x-axis represents proteins ranked by NDM and ten eleven methods, ranked from left to right as strongest to weakest prediction of essentiality. The Y-axis is the cumulative count of essential proteins encountered moving left to right through the ranked. The areas under the curve for NDM and the eleven other methods are used to compare their prediction performance. In addition, the 10 random assortments are also plotted for comparison. (a) shows the comparison results of NDM, IC, SC and DC. (b) shows the comparison results of NDM, BC, CC and NC. (c) shows the comparison results of NDM and other five methods: PeC, CoEWC, POEM, ION and NTMEP.

Table Legends

Table 1 - Impact of the parameter λ on the performance of NDM

Figures

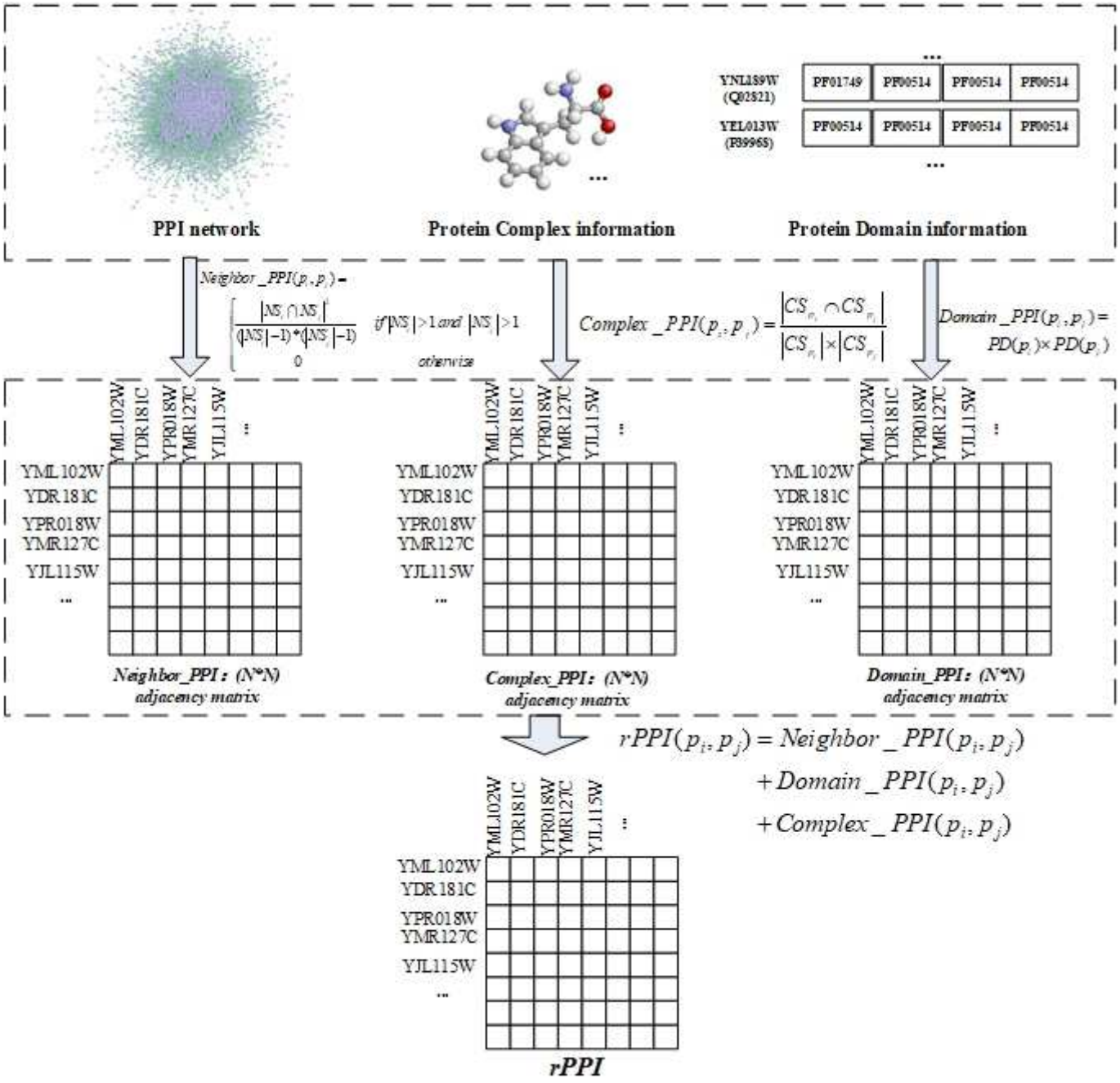

Figure 1

Illustration of the reliable weighted PPI network construction

This figure illustrates construction of the reliable weighted PPI network *r*PPI based on multiple networks, including the *Neighbor_PPI* network, *Domain_PPI* network and *Complex_PPI* network.

Figure 2

Number of correctly essential proteins identified by NDM and other eleven previously competitive methods.

Proteins are ranked in the descending order based on their ranking scores computed by NDM, DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, ION and NTMEP. Then, top 100, 200, 300, 400, 500 and 600 of the ranked proteins are selected as candidates for essential proteins. According to the list of benchmark essential proteins, the number of true essential proteins is used to judge the performance of each method. The figure shows the number of true essential proteins identified by each method.

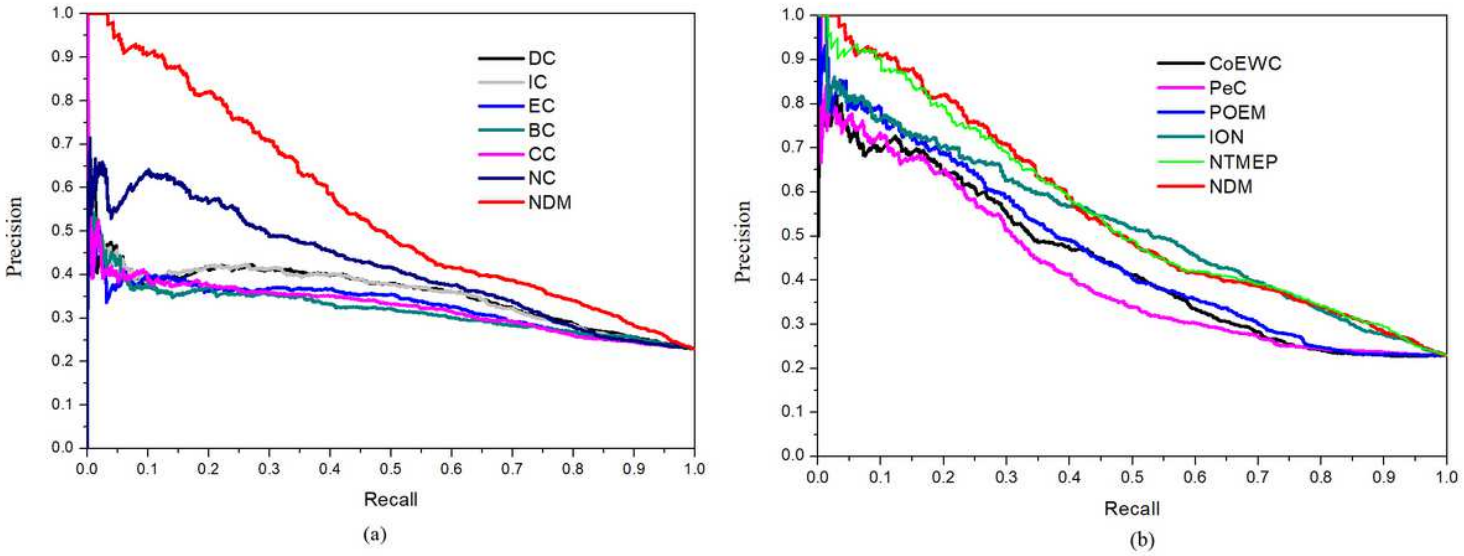

Figure 3

Performance comparison between NDM and other eleven approaches in terms of PR curves.

The proteins ranked in top K (cut-off value) by each method (NDM, DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION and NTMEP) are selected as candidate essential proteins (positive data set) and the rest of proteins are regarded as candidate nonessential proteins (negative data set). With different values of K selected, the values of precision and recall are computed for each method. The values of precision and recall are plotted in PR curves with different cut-off values. (a) shows the PR curves of NDM, DC, IC, SC, BC, CC and NC. (b) shows the PR curves of NDM and other five methods: CoEWC, PeC, POEM, ION and NTMEP.

Figure 4

Jackknife curves for NDM and other eleven approaches.

The x-axis represents proteins ranked by NDM and ten eleven methods, ranked from left to right as strongest to weakest prediction of essentiality. The Y-axis is the cumulative count of essential proteins

encountered moving left to right through the ranked. The areas under the curve for NDM and the eleven other methods are used to compare their prediction performance. In addition, the 10 random assortments are also plotted for comparison. (a) shows the comparison results of NDM, IC, SC and DC. (b) shows the comparison results of NDM, BC, CC and NC. (c) shows the comparison results of NDM and other five methods: PeC, CoEWC, POEM, ION and NTMEP.