American Journal of Biochemistry and Biotechnology

Original Research Paper

# Essential Proteins Discovery Methods based on the Protein-Protein Interaction Networks

[1]LingZhi Zhu, [2]Junling Zhang, [1]Lingya He, [1]Jun Wang, [1]Zhenwu Peng and [1]Zixin Jian

[1]*Department of Computer and Information Science, Hunan Institute of Technology, Hengyang 421008, Hunan, China*
[2]*Offices of Transformation and Development, Hengyang Normal University, Hengyang 421008, Hunan, China*

*Article history*
Received: 06-11-2017
Revised: 11-12-2017
Accepted: 27-12-2017

Corresponding Author:
Lingya He
Department of Computer and Information Science, Hunan Institute of Technology, Hengyang 421008, Hunan, China
Email: 40215417@qq.com

**Abstract:** Essential proteins are closely related to biological survival or reproduction and have important application value in respect of the location of disease genes, disease diagnosis and treatment, drug design. In order to discovery essential proteins, the researchers have proposed experimental approaches, but these methods require laborious and time consuming. With the development of high-throughput sequencing techniques, plenty of computational methods based on the Protein-Protein Interaction Networks have been put forward to identify essential proteins. In this study, we firstly have introduced the basic characteristics and data set of essential proteins and essential proteins discovery method in the Protein-Protein Interaction Networks. Following that, we have analyzed and compared the difference of various existing strategies and then have pointed out the merits and demerits of them in detailed. At last, we give several important problems and development trend about essential proteins discovery methods, which provide a strong foundation for the further research.

**Keywords:** Protein-Protein Interaction Networks, Essential Proteins, Network Centrality, Network Topology, Subcellular Localization Information, Protein Complexes

## Introduction

Protein is an essential component of all cellular and organizational structures (Glass *et al*., 2009). However, different proteins have different importance to life activities. Usually that those essential proteins are removed causes loss of protein complex function and leads to the death and development of proteins (Furney *et al*., 2006). Therefore, Biologists are mainly devoting all of their time to the identification of essential proteins from two purposes. From theoretical perspective, identifying essential proteins helps to understand the minimum requirements for cell survival and development. The recognition of essential proteins has played a vital role in the recent emergence of synthetic biology. Because the purpose of synthetic biology is to create a cell with the smallest genome. From a practical point of view, essential proteins are essential for some bacteria to survive, because they are also targets for new antibiotics. In addition, the results show that essential proteins are closely related to the human gene and are also known to help identify the disease-causing genes. Therefore, there is important application prospect of identifying essential proteins to provide valuable information not only for biology, but for medical and other related disciplines (Jeong *et al*., 2001; Karen and Zelen, 1989), especially for disease diagnosis and drug design.

In biology, essential proteins and disease-causing genes are identified primarily by the means of biomedical experiments (Clatworthy *et al*., 2007). Mutation localization is an classic method to identify essential proteins and disease genes (Wuchty and Stadler, 2003) by detecting pathogenic mutations. With the development of large-scale genome sequencing project, RNA interference is a new identification and analysis technology of essential proteins and disease gene by making the degradation of specific genes to target genes (Cullen and Arndt, 2005). In addition, it is also a widely used technique to identify essential proteins by targeted gene knockout to produce function deletion (Roemer *et al*., 2003). These biological

Science Publications

experiments are clear and effective to identify essential proteins, but they are quite high costs, very low efficiency and very few species. Therefore, the high reliability computational methods are needed to identify essential proteins urgently.

In recent years, with the rapid development of bioinformatics, the essential proteins discovery method based on computer science and mathematics theory has become a new development field. Especially a large number of protein interaction data is provided by the development of high throughput proteomic technology: The yeast two hybrid, tandem affinity purification, mass spectrometry, protein chip and phage display for example.

Many studies have shown that life activities are generally the macroscopic manifestations of complex interacting networks of proteins (genes) formed at the micro level. The interaction of all proteins in a living organism constitutes the protein-protein interaction networks (PPIs).

In PPIs, nodes represent protein molecules and edges represent the interactions between protein and protein.

And then those networks are studied to identify the essential proteins in the way of graph theory. This approach is not only convenient for biologists to observe the data in a straightforward way, but also as a kind of mathematical model of the standardized method is applied to the bioinformatics analysis and research of the relationship of interaction. The essential proteins are closely related to their topological position in PPIs.

Figure 1 shows the interaction relationship between different important nodes in the Yeast of PPIs and then the red dot, blue dots and yellow dots separately represent essential, non-essential and essential unknown proteins.

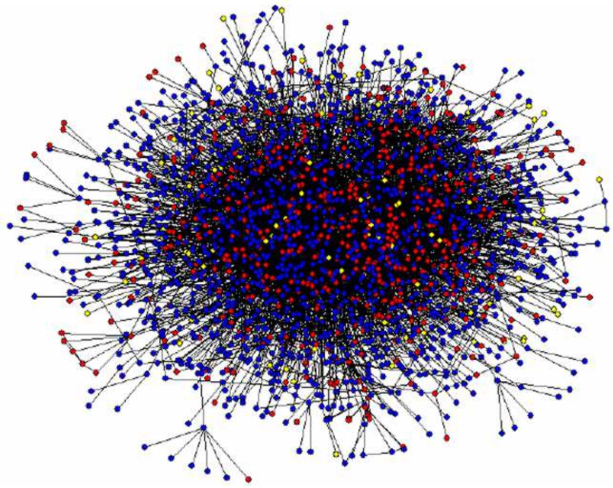

**Fig. 1:** The yeast of PPIs

## Essential Proteins Discovery

Existing studies have shown that there are thousands of errors in cells, but few organisms have serious consequences. What is the source of such resilience. For a random network, if most of the nodes are paralysed, it will inevitably lead to the fragmentation of the network, as the network inevitably breaks up into small islands that are unable to communicate with each other. Due to the topological properties of scale-free networks, protein networks will show the different situation, even in random manufacturing high proportion of mutation in cells, the change of protein will not continue our cooperation. In general, the protein network is remarkably resilient to mutations, which is essentially the non-homogeneous topological structure of the non-scale network. The methods of random removal destroy the non-essential nodes, because they are much larger than the critical nodes. Compared to the critical nodes that connect almost all nodes, non-essential nodes have only a small number of connections and removing them will not have a significant impact on the topology of the network. But the reliance on essential nodes also poses a serious problem: The protein network can be vulnerable to deliberate attacks. Through a series of simulations, the network can be broken into isolated subnet by removing the few key nodes of the scale-free network. Biological experiments on yeast protein networks also show that removing high-connectivity proteins can lead to yeast deaths more than other nodes. These critical nodes are deterministic and once they occur, mutations that make them unworkable are most likely to lead to the death of the whole cell. To avoid mass destruction, it is best way to identify key nodes (proteins) in the network and protect them. So which nodes (proteins) are essential. Because different properties describe proteins in different ways, it is a problem for choosing the right features to identify the essential proteins. In general, there are three criteria to follow:

- The data of this feature should be easy to obtain and effective standard boots biology
- The feature should have high predictive protein capability
- The characteristics have a minimal biological significance to each other

## Data Sets of Essential Proteins Discovery

It is well known that the proteins are very indispensable to all life activities and then different proteins have different importance to life activities. Usually those are removed cause loss of protein complex function and the proteins that cause organisms to be unable to survive or develop are called essential proteins. The recognition of essential proteins can provide valuable information for biology, medicine and so on at the system level. Existing researchers have proposed that protein key and evolutionary rate, expression level, subcellular localization, protein interaction network topology characteristics and so on. At present, on the data of a large number of protein interaction and orthologous protein subcellular localization data is posted on the Internet, academic institutions in published papers at the same time as the experimental data can be downloaded freely available to be used by researchers around the world. Essential proteins discovery research mainly USES data:

- Main source database of Protein interaction
    - DIP (http://dip.doe-mbi.ucla.edu)
    - MIPS (http://mips.gsf.de)
    - BIND (http://www.bind.ca)
    - GRID (http://wiki.dbgrid.org)
    - STRING (http://string.embl.de)
    - HPID (http://www.hpid.org/)
- Main source database of subcellular location
    - eSLDB(http://gpcr.biocomp.unibo.it/esldb/)
    - LOCATE (http://locate.imb.uq.edu.au/)
    - LocDB ( http://www.rostlab.org/services/locDB)
    - SUBA3(http://suba.plantenergy.uwa.edu.au/)
- Main source database of direct homologous protein
    - InParanoid (http://InParanoid.sbc.su.se)

## Essential Proteins Discovery Method

### *Essential Proteins Discovery Methods Based on Network Topology*

Jeong *et al.* (2001) proposed the centrality-lethality rule and pointed out that the essential proteins is closely related to the network topology in PPIs. Specifically, the proteins that they have more neighbors are easier to affect the entire network topology structure, which have fatal effect on the body. That is to say, the higher degree proteins have in PPIs, the more tend to be the essential proteins. The theory also becomes the basis of essential proteins discovery based on network topology (Hahn and Kern, 2005; Estrada and Rodriguez-Velazquez, 2005; Bonacich, 1987; Wang *et al.*, 2012; Joy *et al.*, 2005; Wuchty and Stadler, 2003; Karen and Zelen, 1989). At present, this approachs are based on the correlation between criticality and topological features of protein-protein interaction networks.

### *Based on Local Connectivity*

Degree Centrality (DC) (Hahn and Kern, 2005) was proposed by Hahn MW and Kern AD after studying the protein interaction between yeast, worm and fruit flies. They studied the three species of PPIs, found that after each number of protein interaction and in the network centricity related, more central place in the network of proteins evolve more slowly and more critical in terms of survival. They suggest that the key to protein in the protein interaction network is related to its interaction in the network, which is related to the degree of protein. We assume that the elements of the corresponding positions in the network adjacency matrix are specific and the degree of the node is calculated as follows:

$$DC(u_i)=\sum_{u_j} h_{u_iu_j} \quad (1)$$

Subgraph Centrality (*SC*) (Estrada and Rodriguez-Velazquez, 2005) was Estrada E and rodriguez-velazquez JA. The subgraph centrality method gives the kid a larger weight than the larger one. In comparison with the centrality, the centrality of the subgraph can be better than the nodes in the network. By studying the protein interaction network of saccharomyces cerevisiae, it is found that the centrality of subgraph is more related to the death of single protein from the protein group. We set the number of loops $\mu_l(u_i)$ that start and end with the node $u_i$ and the length l. $\left(\alpha_1,\alpha_2,...,\alpha_{u_j}\right)$ is the *N* linearly independent eigenvectors of the network adjacency matrix *A*, respectively corresponding to the eigenvalue $\lambda_1,\lambda_2,...,\lambda_{u_j}$, $\alpha_{u_j}(u_i)$ is $\alpha_{u_j}$ the first $u_i$ component of the subgraph, which is defined as follows:

$$SC(u_i)=\sum_{l=0}^{\infty}\frac{\mu_l(u_i)}{l!}=\sum_{u_j=1}^{N}\left[\alpha_{u_j}(u_i)\right]^2 e^{\lambda_{u_j}} \quad (2)$$

Eigenvector Centrality (*EC*) (Bonacich, 1987) is a measure of the importance of a node in the network. Each node in the network has a relative value, which is based on the principle that the contribution of high index node to a node is higher than the low index node. We hypothesize that $\alpha_{u_iu_j}\in G$ a quantization of the status support of nodes $u_j$ to $u_i$ in the network *G*. The centricity of the characteristic vector of node u is defined as:

$$EC_{u_i}=\alpha_{1u_i}x_1+\alpha_{2u_i}x_2+...+\alpha_{nu_i}x_n \quad (3)$$

In other words, the center degree of the node u is a function of the other node center degree associated with it and a series of equations defined by this equation can be expressed and solved by the matrix.

Edge Clustering Coefficient Centrality (*NC*) (Wang *et al.*, 2012) unlike the above, this method not only considers the centricity of the nodes, but also considers the centricity between the node and its neighbors. *NC* is applied to three different types of proteins of yeast which are better than DC, BC, CC, SC, EC and IC. Moreover, the essential proteins found by NC showed obvious cluster effect. We will protein network as an undirected graph $G = (V, E)$, suppose you have an edge $E(u_i, u_j)$, its endpoint is $u_i$ and $u_j$, considering how many other node $w$ and $u_i$, $u_j$, in the network and are adjacent, is the other two sides, $E(u_i, w)$, $E(u_j, w)$ and $E(u_i, u_j)$ form a closed triangles. Therefore, the margin aggregation coefficient $E(u_i, u_j)$ is defined as the ratio of the triangle $ECC(u_i, u_j)$ with the most likely composition of the side, namely:

$$ECC(u_i,u_j) = \frac{z_{u_i,u_j}}{\min\left(k_{u_i}-1, k_{u_j}-1\right)} \tag{4}$$

The number of triangles $z_{u_iu_j}$ that actually contain the edges in the network; $k_{u_i}, k_{u_j}$ represents the degree of node $u_i$ and node $u_j$ respectively; $\min\left(k_{u_i}-1, k_{u_j}-1\right)$ represents the number of possible triangles that contain the edge.

Because the essential proteins discovery is still the protein node, the Sum of Edge Clustering Coefficient is the Sum of Edge Clustering Coefficient. The calculation of $SoECC(u_i)$ is as follows:

$$SoECC(u_i) = \sum_{u_j \in N_{u_i}} ECC\left(u_i,u_j\right) \tag{5}$$

where, $N_u$ represents all the neighbor nodes of node $u$. By adding the edge aggregation coefficient and the factor that considers the node degree, it is obvious that the higher protein nodes will have higher SoECC scores.

### *Based on Global Connectivity*

Between's Centrality (BC) was presented by Joy *et al.* (2005) after studying the yeast's Protein Interaction (PPI) network. They found a large number of low-grade proteins in the yeast protein interaction network. This cannot be explained by the lack of scale features of the protein interaction network. There is no scale characteristic that low protein should be low number. These data indicate that there is a module structure in the network and the high number of low-degree proteins may play an important role in connecting two modules. They found that high Numbers of proteins are more important and proteins have a positive evolutionary history. We assume that the network nodes and the number of the shortest path have $f_{vw}$ between $v$, $w$ for, the shortest path $g^u_{wv}$ have in a pass, then the shortest path between, probability:

$$h^u_{wv} = g^u_{wv} / f_{wv} \tag{6}$$

the absolute number of mediation as follows:

$$BC_u = \sum_v^n \sum_w^n h^u_{wv},\ u \neq v \neq w\ v \prec w \tag{7}$$

According to Freeman's research, the node is at the center of a star topology with the maximum number of mediations:

$$BC_{max} = \left(n^2 - 3n + 2\right) / 2 \tag{8}$$

Therefore, the relative intermediary number is:

$$BC_{ru} = 2BE_u / \left(n^2 - 3n + 2\right) \tag{9}$$

Closeness Centrality (*CC*) (Wuchty and Stadler, 2003) was proposed by Wuchty S and Stadler PF in the "Centers of complex networks". *CC* considered the deviation, status and centroid values of three different geometries. They found that the biological networks often (not always) and pure local centrality, such as the degree of vertex, can be used to describe the centrality of biological networks. We assume that $h$ the distance between the nodes $u$ and $v$, the number of nodes of the star network, the proximity of the nodes $u_i$ is:

$$CC^{-1}_{u_i} = \sum_{u_j=1}^{n} \mu_{u_iu_j} \tag{10}$$

Due to the proximity of the core nodes of the star network was $1/(n-1)$, the relative proximity of the nodes $u_i$ is:

$$CC_{u_i} = CC^{-1}_{u_i}(n-1) \tag{11}$$

Information Centrality (*IC*) (Karen and Zelen, 1989) Mediations only consider the shortest paths between nodes and other paths may also have their importance in acquiring or passing information. Information centrality is information based on all possible paths of two nodes. This method does not require the enumeration path to be restricted by the shortest path. Newman thinks it's actually another approximation, essentially measuring the nodes. The harmonic mean length of the path of the endpoint, if it is connected to other nodes through many short paths, means that the average path length is small and the information degree (proximity) is large:

$$IC = \frac{E[G] - E[G']}{E[G]} \tag{12}$$

In this case, the network $G'$ (network $E$ number of edges $G$) of association nodes $u$ and edges is obtained $M$

after removing the strip $E-d(u_i)$ associated with the node and $G$ the efficiency $E[G]$ is defined as:

$$E[G]=\frac{1}{M(M-1)}\sum_{u_i,u_j\in G,u_i\neq u_j} d^{o}_{u_iu_j}/d_{u_iu_j} \tag{13}$$

In the upper formula, the length of $d_{u_iu_j}$ represents the shortest path between $u_i$ and $u_j$, $d^{o}_{u_iu_j}$ indicating that they are along the Euclidean distance of a straight line.

In general, the method based on network topology to build according to the interaction of protein interaction network $G = (V, E)$, $V$ is the protein in the network is a collection of vertices and $E$ is the interaction between the edge (directed or undirected while) collection and then I'm going to set up its adjacency matrix $A$, if $u_i$, $u_j$ is belong to $V$ and the value of $u_i$ and $u_j$ is 1, this means that $u_i$, $u_j$ has one side, 0 means $u_i$, $u_j$ has no edges. These methods first grade each protein in the protein network and then determine the essential proteins based on its score. The advantage of such a method is that it can predict essential proteins directly by using protein scores, without knowing some of the essential proteins in advance to train the classifier. There are three main drawbacks:

- Current Protein-Protein Interaction networks (PPIs) data is incomplete, including many false positives and false negatives. This affects essential protein recognition accuracy
- Poor recognition of essential protein in low connectivity
- The centrality method is proposed for the topological properties of the protein interaction network, ignoring the intrinsic biological significance of the essential proteins

### *Based on Sequence Data*

### *Based on Gene Expression Data*

The current protein interaction data also is not very perfect and inevitably contains a false positive data; it will affect essential proteins discovery method based on network topology characteristics of accuracy. High-throughput experiments provide a large number of other biological information that can be used not only reduce the impact of false positive data, but also characterize essential proteins from different angles. Therefore, many scholars on the basis of in-depth analysis of the topological characteristics of advantages and disadvantages, depicting the topological characteristics of essential proteins from different angles in other biological information fusion to the protein interaction network to improve the accuracy of essential proteins to predict. Li *et al.* (2012) proposed the PeC scheme to identify essential proteins. Let's assume that the sample number is the number of moments in the gene expression data, $E_{u_iu_j}(u_i,u_j)$ and $E_{u_iu_j}(u_i,u_j)$ protein, respectively the $u_i$ and $u_j$ values at the moment, $E_{u_iu_j}(u_i,i)$ and $E_{u_iu_j}(u_j,i)$ for proteins the $u_i$ and $u_j$ expressed in all the moments of the average value and expressed proteins $\sigma(u_i)$ and $\sigma(u_j)$ at all time expression values standard variance, the Pearson correlation coefficient of gene expression:

$$PCC(u_i,u_j) = \frac{1}{K-1}\sum_{t=1}^{k}\left(\frac{E_{u_iu_j}(u_i,i)-\overline{E_{u_iu_j}}(u_i)}{\sigma(u_i)}\right)\left(\frac{E_{u_iu_j}(u_j,i)-\overline{E_{u_iu_j}}(u_j)}{\sigma(u_j)}\right) \tag{14}$$

The range of values $PCC(u_i, u_j)$ is [minus 1, 1] and $PCC(u_i, u_j)<0$ indicates that the gene is negatively correlated with the expression and $PCC(u_i, u_j)>0$ indicates that the gene is positively correlated with the expression, $PCC(u_i, u_j) = 0$ indicates that there is no correlation between the gene and the absence of the gene. Based on the analysis of high non-essential proteins and the fact that essential protein nodes tend to cluster and tend to be expressed. We set $N(u_i)$ and $N(u_j)$ separately represent the collection of neighbor nodes of nodes $u$ and nodes $v$ and the probability of the same cluster is:

$$PCC^{c}(u_i,u_j)=PCC(u_i,u_j)\times\frac{|N(u_i)\cap N(u_j)|}{\min\{|N(u_i)|,|N(u_j)|\}} \tag{15}$$

Considering that the height of the node tends to be the essential proteins, it will $PCC^{c}(u_i, u_j)$ be considered as the weight of the edge and the weight of the connection edge of the node is the central measure of the node:

$$PeC(u_i)=\sum_{u_j\in N(u_i)} PCC^{c}(u_i,u_j) \tag{16}$$

### *Based on Subcellular Localization Information*

At present, many non-supervision methods based on network level are proposed to identify essential proteins. These methods can predict essential proteins directly and do not need to know some of the known essential proteins. However, because the existing methods ignore the spatiotemporal characteristics of protein interactions, the calculated central score cannot measure the key of protein effectively. In addition, many machine learning methods are designed to be too much of a essential proteins in a species and may have poor predictive performance in other species. Peng *et al.* (2015) using subcellular localization information, build the cell interval (PSLIN) protein interaction network, this paper proposes a combining a centricity method for arbitrary in the cell interval to recalculate the protein interaction network centricity score to identify the essential proteins (LSED). We assume that Smax represents the largest PSLIN, $S_i$ represents the scale of PSLIN $S_i$ and the reliability of PSLINSi is the scale of its scale and $S_{\max}$ scale as follows:

$$C(S_i) = \frac{|S_i|}{|S_{maxi}|} \tag{17}$$

For each protein, its *LC* is calculated based on its central score in the sorted PSLIN and the reliability of these PSLIN, which is calculated as follows:

$$LC(\rho) = \begin{cases} LC(\rho), & LC(\rho) \geq Ess(S_i,\rho) \\ LC(\rho) + \left(Ess(S_i,\rho) - LC(\rho) \times C(S_i)\right), & LC(\rho) \prec Ess(S_i,\rho) \end{cases} \tag{18}$$

On the basis of this method, Peng *et al.* (2015) have different importance on the interaction of proteins based on different subcellular intervals and construct a weighted protein interaction network. Then, a Centrality method based on the Importance of subcellular interval is Compartment Importance Centrality (CIC) to detect the essential proteins. We assume that $C_{\max}$ represents the largest range and |I| represents the scale of the interval I. The importance of interval *I*(*i*) is defined as the ratio of the scale of this interval to the largest interval scale. The importance of subcellular interval is as follows:

$$I(i) = \frac{|i|}{|C_{\max}|} \tag{19}$$

For an interaction of two proteins ($u_i$, $u_j$), the interaction of two proteins ($u_i$, $u_j$) subcellular interval information can be defined:

$$SL(u_i,u_j) = Loc(u_j) \cap Loc(u_j) \tag{20}$$

That is the subcellular *u* and *v* interval for proteins and sharing. *SL*(*u*, *v*) may contain zero, one or more subcellular intervals.

If one or more subcellular intervals may be included, *SL*($u_i$, $u_j$) then the importance of the maximum interval ($u_i$, $u_j$) importance as interaction is calculated as follows:

$$w(u_i,u_j) = \begin{cases} \max(I(i)), i \in SL(u_i,u_j) \\ I(C_{\min}), \text{otherwise} \end{cases} \tag{21}$$

The CIC fraction of a protein depends on its importance in the interaction of different sub cells, as shown in formula (18). Among them, the protein of *N*($u_j$) represents the interaction with protein $u_j$, *w*($u_i$, $u_j$) represents the importance of interaction ($u_i$, $u_j$):

$$CIC(u_j) = \sum_{u_i \in N(u_j)} w(u_i,u_j) \tag{22}$$

Compared to CIC method and other methods of centricity in yeast, human, mice and fruit flies on protein interaction network prediction performance, results show that the CIC method on four species better predict the performance of essential proteins. In particular, unlike other ways of over fitting a particular species, the CIC method can be used to predict the essential proteins of different species.

### *Based on Protein Complex Mining*

Numerous studies have shown a close link between protein complexes and essential proteins. Hart *et al.* (2007) through research is pointed out that the key of protein is an attribute of a protein complex, often a lot of focus on some essential proteins function module and in some function modules only exist in very small amounts of essential proteins and on a high-credible yeast protein network, a protein complex was excavated using MCL algorithm. In the Fig. 2, yellow nodes and red nodes represent essential proteins and non-essential proteins. Consequently it is obvious from Fig. 2 that the essential proteins is concentrated in some complex.

On this basis, the Zotenko *et al.* (2008) proposed Essential Complex in Modules, the concept of a group with the same or similar Biological functions of highly connected network of protein function module, thought tendency and the key hub node of reason is because they are a large number of key compounds exists in the module, to participate in more Biological functions. The experimental results also confirm that the essential proteins is abundant in this key complex module. Wang *et al.* (2011) used the theory of Radicchi to measure the importance of a protein in the complex inside the complex.

The protein node *u* in the protein complex *C* is the number of edges connected to the internal node of the complex *C*, which is the inner degree:

$$k_{in}(u_i,C_{u_i}) = \left|E\ (u_i,u_j)\right| u_i,u_j \in V(C) \tag{23}$$

In this case, the *E*($u_i$, $u_j$) of node $u_i$ and node $u_j$ is *V*(*C*) the set of all the protein nodes in the complex *C*. Set $C_{u_i}$ all node $u_i$ compounds, compounds containing protein, protein node *u* in the participation of all the protein complexes In degrees within the Sum of the In-Degree:

$$SoID(u_i) = \sum_{u_i \in C_{u_i}} uk_{in}(u_i,C_{u_i}) \tag{24}$$

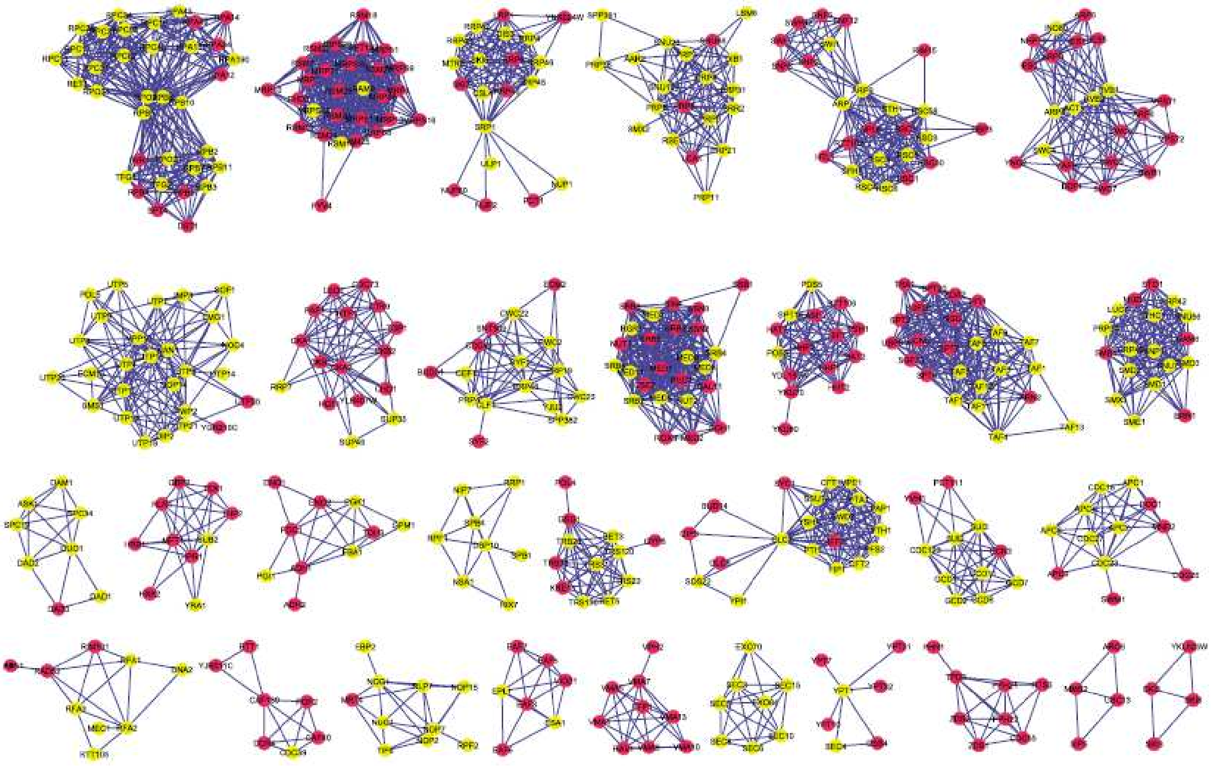

**Fig. 2:** Enrichment of essential proteins in complex (Hart *et al*., 2007)

It is obvious that the protein nodes involved in multiple complexes have relatively higher SoID scores. Therefore, it can be clearly seen from the definition that the index of SoID comprehensively considers the topological properties of proteins within the complex and in the overall situation.

### *Based on Multiple Data Resource Fusion*

### *Content and Application Scope*

Except the network topology feature in PPIs. Essential proteins, essential genes and the following characteristics was used to combine to multiple data fusion: Such as fluctuation in mRNA expression, GC content, Condon adaptation index, predicted sub-cellular localization, Condon usages, biased distribution of essential genes in leading and lagging strands, homologous search, evolutionary rate, phylogenetic conservation.

For each feature, there is a corresponding application scope and biological significance. Such as homologous search, evolutionary rate, phylogenetic conservation means the conservative protein can be essential proteins. Flux balance analysis is widely used to evaluate the critical of genes. But the method requires a clear definition of nutrient availability and biomass yield in a given environment. The concept of important functions of load points and choke points is used to estimate whether an enzyme is critical. The evolutionary rate and topology of proteins are analyzed in combination, such as the protein in the center of the protein evolve slowly and thus it more likely to be the essential protein. Gene sequence characteristics, such as codon usage, GC content and localization signals are successfully used to deduce essential genes of s.m. ikatae from S.c erevisiae. Phylogenetic conservation is used as a essential predictive feature of genes in and such as genes in e. coli and B.s ubtilis, the essential genes of leading and lagging strands have been widely known to be distributed.

### *Methods based on Machine Learning*

The method of computing multivariate information in identifying essential proteins is generally combined with the essential protein characteristics discovered by machine learning methods. The individual essential protein characteristic is not obviously, in the adjustment of various information, it can dig out the essential protein in a better way. The machine learning methods can carry out the identification or classification of essential genes effectively. There are many specific machine learning methods that can be used to solve this problem, such as SVM, deep learning and decision etc. Most studies when modeling the identification of proteins, it authenticate problems as a supervised secondary classification problem. The classification problem can be defined as follows: Input vector $\{x_n\}$, $n = 1...,N$, identifying classification label $\{y_i\}$, as general labeling value is {essential, non-essential}. The data read by the secondary classifier is generally a matrix of n rows (each protein) m column (each eigenvalue of the protein). Among them, n genes constitute the sample and the m column is the characteristic of the input protein.

But the accuracy and performance of the training set classification depend on the training dataset and how the training data sets designed for classifiers represent the actual data sets. That means the degree to which a large degree of identification depends on the value of the key protein identifies the key protein.

### *Based on Multiple Data Binding Model*

From the perspective of system biology, the method of identifying essential proteins based on multiple data binding models is explored. Specifically, we unified the expression of the data and designed the scoring function of each data by the characteristics analysis of a variety of related data. According to these related data score, an effective computational model that combines a variety of related data is proposed to identify essential proteins, as shown in Fig. 3.

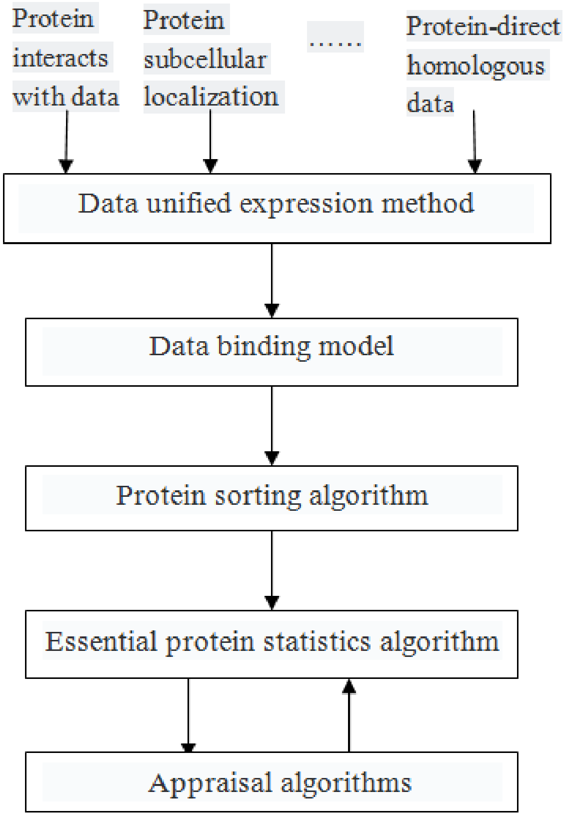
Protein interacts with data
Protein subcellular localization
……
Protein-direct homologous data
Data unified expression method
Data binding model
Protein sorting algorithm
Essential protein statistics algorithm
Appraisal algorithms

**Fig. 3:** Multiple data binding model

### *Based on Weighted Protein-Protein Interaction Networks*

However, the protein interaction data currently available contains a large number of false positives, which greatly reduce the accuracy of essential proteins discovery. Tang *et al.* (2011) proposed a method WDC for the centrality centrality. This chapter is based on this consideration, this paper puts forward a kind of combined with logistic regression model and GO the interaction of the weighted semantic similarity method, weighted protein network, to establish a high degree of confidence on the weighted network using centricity measure used to identify the essential proteins and verify the effectiveness of the proposed method. Lin (1998) defines the similarity of two GO annotations as the maximum amount of information of their common ancestor with the average amount of information of these two notes, namely:

$$sim(c_1,c_2)=\frac{2\max_{c\in C_T(c_1,c_2)}(\log\rho(c))}{\log\rho(c_1)+\log\rho(c_2)} \qquad (25)$$

Among them, $\rho(c)$ represents the probability of the occurrence of the comment phrase $c$ in the whole GO data set, $c_T(c_1,c_2)$ representing the collection of all common ancestors.

On this basis, Li *et al.* (2010) defined the reliability of interaction $E(u_i, u_j)$ as the functional similarity (functional similarity, FS) of protein $u_i$ and $u_j$, namely:

$$FSim(u_i,u_j)=\max_{c_1\in F_{u_i},c_2\in F_{u_j}}\left(sim(c_1,c_2)\right) \qquad (26)$$

Among them, $F_{u_i}$, $F_{u_j}$ the function annotation set representing protein $u$ and protein $v$ respectively. Since the reliability of each side represents the probability of interaction, between 0 and 1, the normalization of $FS(u_i, u_j)$ should be made:

$$FSim^*(u_i,u_j)=\frac{FSim^*(u_i,u_j)-\min(FSim)}{\max(FSim)-\min(FSim)} \qquad (27)$$

Among them, max($FSim$) and man($FSim$) respectively represent the maximum and minimum value of the functional similarity of all proteins interacting with each other. Taking into account the above factors, Li *et al.* (2010) constructed the weighted protein network based on the logistic regression model and the credibility of GO semantic similarity.

The weight of the interaction between protein $u_i$ and protein $u_j$ is defined $W(u_i, u_j)$ as follows:

$$W\left(u_i,u_j\right)=\frac{L\left(u_i,u_j\right)+FSim^*\left(u_i,u_j\right)}{2} \tag{28}$$

## Summary

The identification of essential proteins is significant and can promote the application of synthetic biology effectively, seeking candidate drug targets and searching for disease-causing genes. The calculation method of identifying essential proteins makes up for the long time, low efficiency and high cost of the biological experiment method and the experimental object has the risk, which can only be implemented on a small number of species. The existing method of identifying essential proteins at present is mainly of protein interaction network topology structure, low identification accuracy and limited to specific species. Fusion of many kinds of protein data is to identify the essential proteins development trend, many methods to discovery essential proteins has obtained certain result, but still needs to improve in such aspects as identification accuracy, a series of key scientific problems and challenges, which remain to be researched specific as follows:

- The coming of the era of big data accumulated the massive amounts of different types of biological data, the current lack of effective method to integrate the data used in the study of key aspects of protein identification
- More than the existing methods to build a single protein network used to identify essential proteins and the construction of essential proteins network reliability is not high, the lack of perfect and reliable essential proteins discovery method and model
- The current researchers is to build a more simplified protein information and then identify essential proteins and then a simplified protein may ignore some key information, the lack of perfect multiple information fusion model
- The existing calculation methods are mainly based on static network, which fails to reflect the dynamic characteristics of protein and lacks the dynamic network with time/blank characteristics
- Some methods only identify essential proteins in specific species and perform well in other species

According to the current situation, it is urgent to find the calculation method of high accuracy and multiple species.

## Funding Information

This work is supported by the National College Students' Innovation and Entrepreneurship Training Program ([2017]40-14748), The Project Development Plan of Science and Technology of Hengyang City (No. 2017KJ246) and the Hunan Province College Students' Research Learning and Innovative Experiment Plan Fund (Xiang Jiao Tong [2017]205-845).

## Author's Contributions

**Lingzhi Zhu and Junling Zhang:** Contributed to the planning and implementation of this article and wrote the article of the first three chapters as well as the third sections of the fourth chapters.

**Lingya He and Zhenwu Peng:** Wrote the article of the second sections of the fourth chapters and revising the article.

**Jun Wang and Zixin Jian:** Wrote the article of the first sections of the fourth chapters.

## Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

## References

Bonacich, P., 1987. Power and centrality: A family of measures. Am. J. Sociol., 92: 1170-1182.

Clatworthy, A.E., E. Pierson and D.T. Hung, 2007. Targeting virulence: A new paradigm for antimicrobial therapy. Nat. Chem. Biol., 3: 541-548. DOI: 10.1038/nchembio.2007.24

Cullen, L.M. and G.M. Arndt, 2005. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol. Cell Biol., 83: 217-223. DOI: 10.1111/j.1440-1711.2005.01332.x

Estrada, E. and J.A. Rodriguez-Velazquez, 2005. Subgraph centrality in complex networks. Phys. Rev. E, 71: 056103-056103. DOI: 10.1103/PhysRevE.71.056103

Furney, S.J., M.M. Alba and N. Lopez-Bigas, 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genom., 7: 165-165. DOI: 10.1186/1471-2164-7-165

Glass, J.I., C.A. Hutchison, H.O. Smith and J.C. Venter, 2009. A systems biology tour de force for a near-minimal bacterium. Mol. Syst. Biol., 5: 330-330. DOI: 10.1038/msb.2009.89

Hahn, M.W. and A.D. Kern, 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol., 22: 803-806. DOI: 10.1093/molbev/msi072

Hart, G.T., I, Lee and E.M. Marcotte, 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinform., 8: 236-236. DOI: 10.1186/1471-2105-8-236

Jeong, H., S.P. Mason, A.L. Barabási and Z.N. Oltvai, 2001. Lethality and centrality in protein networks. Nature, 411: 41-42. DOI: 10.1038/35075138

Joy, M.P., A. Brock, D.E. Ingber and S. Huang, 2005. High-betweenness proteins in the yeast protein interaction network. J. Biomed. Biotechnol., 2005: 96-103. DOI: 10.1155/JBB.2005.96

Karen, S. and M. Zelen, 1989. Rethinking centrality: Methods and examples. Soc. Netw., 11: 1-37. DOI: 10.1016/0378-8733(89)90016-6

Li, M., H. Zhang, J.X. Wang and Y. Pan, 2012. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst. Biol., 6: 15-15. DOI: 10.1186/1752-0509-6-15

Li, M., J.X. Wang, H. Wang and Y. Pan, 2010. Essential proteins Discovery from weighted protein interaction networks. Proceedings of the International Symposium on Bioinformatics Research and Applications, (BRA' 10), Springer, Berlin, pp: 89-100. DOI: 10.1007/978-3-642-13078-6_11

Lin, D., 1998. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, Jul. 24-27, Morgan Kaufmann Publishers Inc., USA, pp: 296-304.

Peng, X., J.X. Wang, Z. Jiancheng, J. Luo and Y. Pan, 2015. An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Nov. 9-12, IEEE Xplore Press, Washington, DC, USA, pp: 277-280. DOI: 10.1109/BIBM.2015.7359693

Roemer, T., B. Jiang, J. Davison, T. Ketela and K. Veillette *et al*., 2003. Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. Mol. Microbiol., 50: 167-181. DOI: 10.1046/j.1365-2958.2003.03697.x

Tang, X., J. Wang, B. Liu, M. Li and G. Chen *et al*., 2011. A comparison of the functional modules identified from time course and static PPI network data. BMC Bioinformat., 12: 339-339. DOI: 10.1186/1471-2105-12-339

Wang, H., M. Li, J.X. Wang and Y. Pan, 2011. A new method for identifying essential proteins based on edge clustering coefficient. Proceedings of the International Symposium on Bioinformatics Research and Applications, (BRA' 11), Springer, Berlin, pp: 87-98. DOI: 10.1007/978-3-642-21260-4_12

Wang, J.X., M. Li, H. Wang and Y. Pan, 2012. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans. Comput. Biol. Bioinform., 9: 1070-1080. DOI: 10.1109/TCBB.2011.147

Wuchty, S. and P.F. Stadler, 2003. Centers of complex networks. J. Theor. Biol., 223: 45-53. DOI: 10.1016/S0022-5193(03)00071-7

Zotenko, E., J. Mestre, D.P. O'Leary and T.M. Przytycka, 2008. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. PLoS Comput. Biol., 4: e1000140-e1000140. DOI: 10.1371/journal.pcbi.1000140