INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 13(3), 365-382, June 2018.

Identifying Essential Proteins in Dynamic PPI Network with Improved FOA

X. Lei, S. Wang, L. Pan

Xiujuan Lei, Siguo Wang
School of Computer Science
Shaanxi Normal University
Xian 710119,Shaanxi, China
xjlei@snnu.edu.cn, wangsiguo@snnu.edu.cn

Linqiang Pan*
1. Key Laboratory of Image Information Processing and
Intelligent Control of Education Ministry of China
School of Automation
Huazhong University of Science and Technology
Wuhan 430074, Hubei, China
2. School of Electric and Information Engineering
Zhengzhou University of Light Industry
Zhengzhou 450002, Henan, China
*Corresponding author: lqpan@mail.hust.edu.cn

Abstract: Identification of essential proteins plays an important role for understanding the cellular life activity and development in postgenomic era. Identification of essential proteins from the protein-protein interaction (PPI) networks has become a hot topic in recent years. In this work, fruit fly optimization algorithm (FOA) is extended for identifying essential proteins, the extended algorithm is called EPFOA, which merges FOA with topological properties and biological information for essential proteins identification. The algorithm EPFOA has the advantage of identifying multiple essential proteins simultaneously rather than completely relying on ranking score identification individually. The performance of EPFOA is analyzed on dynamic PPI networks, which are constructed by combining the gene expression data. The experimental results demonstrate that EPFOA is more efficient in detecting essential proteins than the state-of-the-art essential proteins detection methods.
Keywords: essential proteins, protein-protein interaction (PPI), dynamic PPI networks, subcellular localization data, fruit fly optimization algorithm (FOA).

1 Introduction

Protein plays an important role in the cellular life activity, and essential proteins are critical for the growth and development of organisms under a variety of conditions [27]. The absence of a single essential protein is sufficient to cause lethality or infertility [50]. Some recent results suggest that a comprehensive analysis of essential proteins can provide a deeper understanding of the relationship between mutations and human diseases, revealing the general principles of human diseases [12, 15, 59]. Therefore, the identification of essential proteins is closely related to disease prediction and drug design [53].

With the development of high-throughput technologies, various biological data are available, e.g., yeast-two-hybrid, tandem affinity purification, and mass spectrometry. In [2], a greedy algorithm is proposed to optimize the detection of protein communities.

Existing methods for identifying essential proteins can be roughly divided into two types. The first type includes the biological experiment-based methods, e.g., gene knockouts [11], RNA interference [7], and conditional knockouts [35], which are expensive and time-consuming. The

other type includes the topology-based centrality method, e.g., Degree Centrality (DC) [14], Betweenness Centrality (BC) [24], Closeness Centrality (CC) [52], Subgraph Centrality (SC) [9], Eigenvector Centrality (EC) [3], Information Centrality (IC) [40], Neighborhood Centrality (NC) [20], and Local Average Connectivity-based method (LAC) [19]. By defining and computing the topologically potential value of each protein, these methods can obtain a precise ranking score reflecting the importance of proteins in the protein-protein interaction (PPI) network [18]. Some centrality analysis tools and RNA detection tools [54] have been developed. For example, CytoNCA [43], a Cytoscape plugin, integrated eight centrality measures, i.e., DC, BC, CC, EC, IC, SC, NC and LAC. Obviously, the topology-based centrality methods can improve the efficiency with less cost. However, these centrality methods also have their own shortcomings. It is well known that the performance of topology-based methods is closely related to the quality of the PPI networks, but there are many false positive and false negative in the PPI networks.

In order to deal with the drawbacks of these methods, some new methods are proposed to predict essential proteins by integrating their topological properties with their biological properties. Considering the interaction data and Gene Ontology (GO) annotations, Hsing et al. introduced a method to predict highly-interacting proteins [13]. Later, a new prediction method called PeC was proposed by Li et al. [21], and another method called WDC was proposed by Tang et al. [41], which integrate network topology with gene expression profiles. Afterwards, Tang introduced a new method to identify essential proteins in which topological features of PPI network is combined with subcellular localization information [42]. Next, a new centrality measure is proposed by Ren et al. to discover essential proteins, named harmonic centricity, which merges subgraph centrality with protein complexes to discover essential proteins [34]. Recently, a new prediction method, named UDoNC, that combine the domain features of proteins with their topological properties in PPI networks, was proposed by Peng et al. [30]. Some machine learning methods, e.g., Support Vector Machine, Naive Bayes, Bayes Network, and NBTree, were also adopted to predict essential proteins by using different features. For example, the random forest was adopted to predict essential proteins by Qin et al. [32]. These methods that combine the network topological features with biological data is capable of improving the accuracy and efficiency of prediction significantly. These existing methods regard the PPI networks as static networks that ignore the time-course of the networks. The real PPI networks in cell keeps changing over different stages of the cell cycle [31], and they can be classified into stable or transient PPI networks [46], which are usually described as dynamic PPI networks (DPIN). Thus it is important to construct dynamic PPI networks to investigate the temporal properties of individual proteins and protein interactions. Based on dynamic network topology and complex information, Luo and Kuang proposed a new method to predict essential proteins [22]. The results show that the identification of essential proteins in dynamic networks is more conducive than in static networks.

Fruit fly optimization algorithm (FOA) is a novel swarm intelligent algorithm that mimics the foraging behavior of fruit flies for global optimization [25]. FOA is easy to be understood and implemented, which has few parameters to be adjusted. Due to its simplicity and efficiency, FOA showed great success in solving some real-world complex problems like multidimensional knapsack problem [48]. Here, FOA will be used to find the essential proteins.

In this work, we present a new algorithm, called EPFOA, in which FOA is merged with topological properties and biological information for essential proteins identification. To the best of my knowledge, most of the methods of essential proteins identification focus on static PPI networks and ignore the intrinsic features of organisms.

In our method, we first integrate gene expression data with static PPI network to construct the dynamic network model. Then a new topological centrality method that combines GO annotation and edge aggregation coefficient (ECC) is proposed to measure the topological char-

acteristic of PPI networks with modular local average connectivity (LAC) in dynamic networks. Furthermore, the distribution of proteins in each compartment according to subcellular localization data is obtained, and the role of components in identifying essential proteins is analyzed.

Finally, EPFOA is designed to identify essential proteins. To assess the performance of our method, EPFOA is compared with some existing methods including DC, EC, IC, SC, NC, LAC, PeC and UDoNC, and the experimental results indicate that our method significantly outperforms with the existing methods.

2 Method

2.1 Fruit fly optimization algorithm

Fruit fly optimization algorithm (FOA) is a novel method for global optimization, which is inspired by the foraging behavior of fruit flies. In sensory perception, the fruit fly is superior to the other species, especially in olfactory and vision. The olfactory organs of fruit flies can collect all kinds of scents floating in the air, even smell the food source from 40 kilometers away. After the fruit fly gets close to the food, it can also use the sensitive vision to find food and the companyĄŹs flocking location, and fly to the direction [25]. The procedure of FOA is presented in pseudo code as follows.

Step 1. Randomly initialize the location of the fruit flies (X_{axis}, Y_{axis}).

Step 2. Give the random direction and distance for the search of food using osphresis by an individual fruit fly.

$$\begin{cases} X_i = X_{axis} + RandomValue \\ Y_i = Y_{axis} + RandomValue \end{cases} \tag{1}$$

Step 3. The distance $(Dist_i)$ to the origin is estimated, then the smell concentration judgment value (S_i) is calculated, which is the reciprocal of the distance.

$$\begin{cases} Dist_i = \sqrt{x^2 + y^2} \\ S_i = \frac{1}{Dist_i} \end{cases} \tag{2}$$

Step 4. Substitute smell concentration judgment value (S_i) into smell concentration judgment function (or called fitness function) to find the smell concentration $(Smell_i)$ of the individual location of the fruit fly.

$$Smell_i = Function(S_i) \tag{3}$$

Step 5. Find the individual with the maximal smell concentration among the fruit fly swarm according to the smell concentration value.

$$[bestSmell\ bestInedx] = max(Smell) \tag{4}$$

Step 6. Maintain the best smell concentration value x and y, where the fruit fly swarm will use vision to fly towards that location.

$$\begin{cases} Smell = bestSmell \\ X_{axis} = X(bestIndex) \\ Y_{axis} = Y(bestIndex) \end{cases} \tag{5}$$

Step 7. Repeat steps 2-5 until the smell concentration is superior to the previous smell concentration; otherwise, go to step 6.

2.2 Dynamic PPI network model construction

Gene expression data is valuable for revealing the dynamic properties of proteins and PPI. We integrate gene expression data with high-throughput PPI data to construct a dynamic PPI network. Note that protein does not always become active at a cell cycle, a protein is active at the highest gene expression level. In order to mark the active time of each gene, the active threshold of each gene should be calculated, and the gene is active if its expression value is greater than the active threshold. The calculation of active threshold is proceeded on the 3-sigma model [45].

$$AT(p) = \mu(p) + 3 \times \sigma(p) \times (1 - \frac{1}{1 + \sigma(p)^2}), \tag{6}$$

where $\mu(i)$ is the mean gene expression value of protein i and $\sigma(i)$ is the algorithm standard deviation of the expression values over time 1 to T for protein i. Since the gene expression data has three cycles and each cycle has 12 times tamps, the final gene expression at each time point is the average of the three cycles, which is defined as follows [16]:

$$FT(i) = \frac{T(i) + T(i+12) + T(i+24)}{3}, (i \in [1, 12]), \tag{7}$$

where $T(i)$ denotes the gene expression value at time point i. At a certain times tamp, if both proteins are active with an interaction, the interaction of the two proteins is also active. Eventually the entire PPI network was divided into 12 sub-networks, the dynamic PPI network was constructed.

2.3 Topological characteristics of dynamic networks

A PPI network is not only an important biological network but also a typical complex network, which meets the topological characteristics of complex network, such as small-world [49], scale-free [51], and modularity [10]. In this part, the role of the topological characteristics in the process of essential proteins identification is investigated, and a new topological centrality method based on the ECC and GO annotation is proposed. Furthermore, the modularity of the network that applied LAC is also considered.

Dynamic network topology centrality strategy

A PPI network can usually be expressed as an undirected graph $G = (V, E)$, where the set of vertices V represents protein, and E represents all of interactions between pairs of proteins. In order to assess the centrality of dynamic network topology, we introduce the GO annotation (since the ECC cannot fully reflect the characteristics). GO annotation provides valuable information and a convenient method to study the gene function similarity, some researches have shown that the adoption of GO semantic similarity term can improve the prediction accuracy of protein complexes gene and disease [36, 56, 57].

Weighting the networks via ECC

In order to measure the tightness of the two nodes, we use the ECC [41], which is defined as follows:

$$ECC(u, v) = \frac{|N_u \cap N_v| + 1}{min\{d_u, d_v\}}, \tag{8}$$

where N_u (or N_v) refers to the set of neighbours of node u (or v) in PPI networks, $|N_u \cap N_v|$ is the number of common neighbor nodes of u and v, which is consistent with the number of triangles which edge (u, v) belongs to d_u (or d_v) indicates the degrees of node u (or v).

Weighting the networks using the Gene Ontology

The GO information consists of three sub-ontologies: Biological Process (BP), Cellular Component (CC) and Molecular function (MF) [6]. In order to measure the semantic similarity between the GO terms to protein annotations in an interaction network, we applied the method developed by Wang et al. [47]:

$$GO\_sim(u,v) = \frac{\sum_{t \in T_u \bigcap T_v}(S_u(t) + S_v(t))}{\sum_{t \in T_u} S_u(t) + \sum_{t \in T_v} S_v(t)}, \tag{9}$$

where where T_u and T_v are the annotations of protein u and v; $S_u(t)$ is the S-value of GO term t related to term u and $S_v(t)$ is the S-value of GO term t related to term v.

Generating new weighted networks

Based on the definition of the ECC and gene functional similarity, a new centrality measure, named EG, is proposed. For a protein u, the essentiality $EG(u)$ is defined as the probability between the ECC and GO information:

$$EG(u) = \sum_{v \in N_u} ECC(u,v) \times GO\_sim(u,v), \tag{10}$$

where $N(u)$ denotes the set of all neighbors of node u. When computing dynamic $EG(u)$, we should consider the number of times that each node appears in a dynamic PPI network, since some nodes are not included in all the time networks. Dynamic $EG(u)$ can be defined as the follows:

$$D_{EG}(u) = \frac{\sum_{i=1}^{N} EG^i(u)}{tim(u)}, \tag{11}$$

where N is the number of temporal networks in the dynamic network, $EG^i(u)$ the EG of node u in the ith time point, $tim(u)$ the number of time networks that contain node u. If node u does not appear at time point i, $EG^i(u)$ is equal to zero.

Dynamic local average connectivity

The LAC of a node indicates its closeness [49], and the LAC of a node v is defined as:

$$LAC(u) = \frac{\sum_{v \in N_u} deg^{C_u}(v)}{|N_u|}, \tag{12}$$

where N_u is the neighbors of node v, $|N_u|$ the number of nodes in N_u, and C_u the subgraph induced by N_u. For a node u in C_v, its local connectivity in C_u is represented as $deg^{(C_u)}(v)$. Similar to $D_{EG}(u)$, we define Dynamic LAC as follows [30]:

$$D_{LAC}(u) = \frac{\sum_{i=1}^{N} LAC^i(u)}{tim(u)}, \tag{13}$$

where N is the number of temporal networks in the dynamic network, $LAC^i(u)$ the LAC of node v in the ith time point, and $tim(u)$ the number of time networks that contain node u.

2.4 Subcellular localization score

Subcellular location is divided into different compartments, different compartments play different roles in cell activities. In order to understand the relationship between subcellular localization and essential proteins, we analyze the number of essential proteins in each subcellular location and propose a method to evaluate subcellular localization in previous research. Assume

that in the Nucleus, the wider the distribution of the proteins is, the greater the possibility of essential protein becomes [17].

Let C_{max} denote the protein with the largest number of times appearing in subcellular localization of the nucleus, $|u|$ represents the number of times of the protein u appearing in the nucleus. The importance of protein u, denoted as $NSL(u)$, is calculated by the ratio of its size to the largest size of the nucleus. The value of $NSL(u)$ is in the range of $(0, 1]$.

$$NSL(u) = \frac{|u|}{|C_{max}|} \tag{14}$$

2.5 EPFOA algorithm

In order to make up for the shortcomings of traditional identification of essential proteins one by one, we propose the algorithm EPFOA. The algorithm can identify p candidate essential proteins simultaneously, which greatly improves the recognition efficiency. In what follows, we introduce the algorithm EPFOA. First, initialize the position of fruit fly and set the rules of location updating. Then find p candidate essential proteins according to the characteristic of FOA. Finally, the identified p essential proteins are compared with known essential proteins to verify the number of essential proteins identified correctly.

The initialization and update of the location of fruit flies

The initialization and location update rules of fruit fly play an important role in the performance of EPFOA. The position of the fruit fly is encoded as an integer set of p-dimensional set $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $(i = 1, 2, \ldots, n)$ which denotes a candidate essential protein set. Each element x_{ij} $(x_{ij} \leq |N|, j = 1, 2, \ldots, p)$ in X_i is the sequence number of a protein. First we randomly selected p proteins to initialize a fruit fly position X_i. Then we compare the selected p proteins with the known essential proteins and keep the proteins that are successfully matched. After that the remaining positions that represent proteins are updated. In order to speed up the convergence of the proposed EPFOA algorithm, we sort all the proteins based on degree except selected p proteins. A random value is assigned to the individual that is not essential protein in the X_i and update the position in a sequence that is ranked by degree.

Encoding and decoding of EPFOA

The framework of EPFOA is shown in Fig.2. We set every fruit fly as essential protein candidate set, and the location of fruit fly is the serial number of the candidate proteins. For the purpose of evaluating the topological characteristics of the network comprehensively, we combine LAC that represents the network modularity with the new network centrality. Thus, when a fruit fly is in a certain position, we suppose its smell concentration judgment value $S(i)$ can be calculated as following equation:

$$S(i) = \sum_{j=1}^{p} (D_{LAC}(\mu_j) + D_{EG}(\mu_j)), \tag{15}$$

where $D_{LAC}(u_j)$ denotes the dynamic local average connectivity of the jth protein among the p candidate essential proteins and $D_{EG}(u_j)$ denotes dynamic network topology centrality of the jth protein among the p candidate essential proteins.

The topological characteristics and biological data are both indispensable in the process of identifying essential proteins and subcellular localization data plays an important role in essential proteins identification. We set the following smell concentration judgement function to measure the possibility of essential proteins represented by a fruit fly individual:

$$Fit(i) = \alpha \times S(i) + (1 - \alpha) \times \sum_{j=1}^{p} NSL(\mu_j), \tag{16}$$

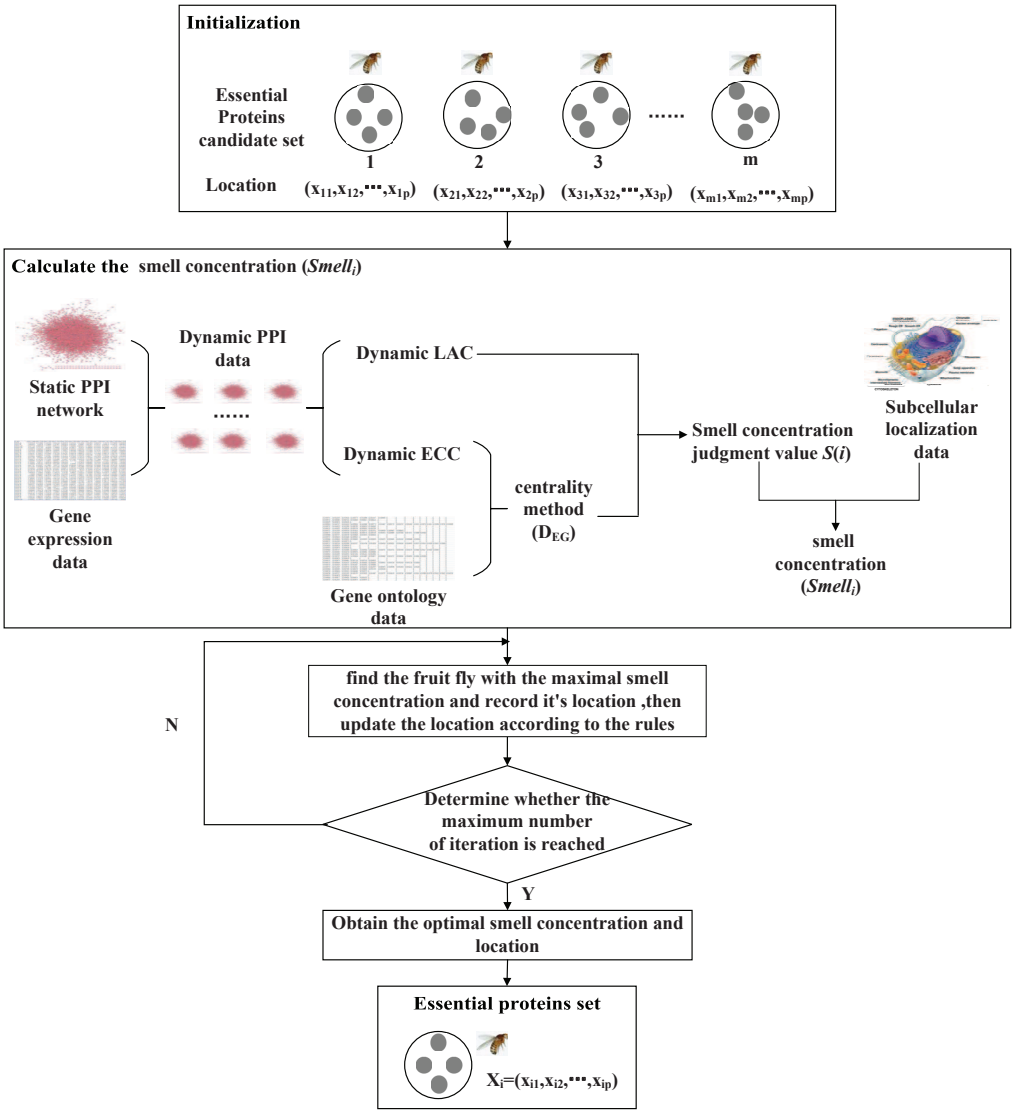

Figure 1: The framework of the algorithm EPFOA.

where $NSL(u_j)$ denotes subcellular localization score of the jth protein among the p candidate essential proteins and $\alpha \in [0, 1]$, α is a parameter that regulates the proportion of the network topology and biological information in the process of identifying essential proteins. If $\alpha = 0$, only subcellular location information works; else if $\alpha = 1$, only network topology works.

Pseudo code of EPFOA

The process of EPFOA can be divided into two steps. The first step calculates the topological and biological characteristics of protein nodes. The second step applies the process of FOA algorithm to seek the optimal to find the essential proteins. The pseudo code of EPFOA is shown in Algorithm 1.

Algorithm 1 The pseudo code of EPFOA

Ensure: $G = (V, E)$ (the PPI network), Gene expression data, Gene Ontology GO, Subcellular location data.
Require: Essential protein set.
Construct the dynamic PPI network
for each interacting protein pair (a, b) in PPI **do**
    Calculate ECC /*The closeness of the two nodes*/
    Calculate GO /*The functional similarity of the two nodes based on GO annotation*/
end for
for each node in G **do**
    Update the centrality $D_{EG}(u)$
    Calculate $D_{LAC}(u)$
    Calculate subcellular location score $NSL(u)$
end for
for fruit fly i **do**
    Initialize location $x(i)$ and its best location $b\_x(i)$
    Calculate the smell concentration smell(i)= Fit($S(i)$)
end for
for m in $[1, maxiter]$ **do**
    for fruit fly i **do**
        Update location $X(i) = X(i) + random$
        if smell(i)<Fit(S(i)) **then**
            $b\_x(i) = X(i)$
        end if
    end for
end for

3 Results and discussion

In this section, we first introduce the experimental data. Then we analyze the parameter A towards the performance of EPFOA. Next, in order to evaluate the performance of EPFOA more synthetically, we not only compare EPFOA with some topology-based centrality methods (DC, EC, IC, SC, NC, LAC) but also with some methods that integrate their topological properties with their biological properties (PeC and UDoNC). In order to assess the essentiality of proteins in PPI networks, these methods are ranked in descending order based on their ranking scores including eight existing centrality methods (DC, EC, IC, SC, NC, LAC, PeC and UDoNC). After

that, top 1%,5%, 10%, 15%, 20% and 25% of the ranked proteins are selected as candidates for essential proteins. In this paper, the size of the set of essential proteins candidate is 1274. Taking into account of the random optimization process of FOA, we conduct ten experiments and then use the average of ten experiments as the final result to to analyze the parameter towards the performance of EPFOA. The ten experiments are listed in the attachment 1. To further evaluate the EPFOA performance, we randomly choose a candidate essential proteins from ten experiments to compare with other methods. The performance is presented in the form of histograms of the number of essential proteins predicted by each algorithm and also use six statistical measures to evaluate them. And precision-recall curves and jackknife curves are also used to evaluate the performance of the proposed EPFOA method and the other eight methods. Finally,we analysis the modularity of identified essential proteins.

3.1 Experimental data

To evaluate the performance of our proposed algorithm EPFOA, we adopt PPI networks of S.cerevisiae which has been well characterized by knockout experiments and widely used in the evaluation of methods for essential proteins discovery. The PPI data of S.cerevisiae was downloaded from DIP database [58], which contains 5093 proteins and 24743 interactions after removing the repeated interactions and the self-interactions. The known essential proteins data of S.cerevisiae contains 1285 essential proteins among which 1167 essential proteins present in the DIP network, which are collected from four databases: MIPS [23], SGD [4], DEG [55], and SGDP (http://www-sequence.stanford.edu/group). The gene expression data of S.cerevisiae are downloaded from GEO database [44] that contains 7074 gene expression products. The Gene ontology annotation data of S.cerevisiae is obtained from GO Consortium [5]. Subcellular localization dataset of S. cerevisiae is downloaded from knowledge channel of COMPARTMENTS database [1], which includes 5095 yeast proteins and 206,831 subcellular localization records.

3.2 The effect of parameter α on performance

In our proposed algorithm EPFOA, evaluation function of proteins is changed with different values of α. To study the effect of parameter α on performance of EPFOA, we evaluate the prediction accuracy by setting different values of α, ranging from 0 to 1. The detailed results are listed in Table 1. As shown in Table 1, the results are similar with α, ranging from 0.4 to 1. Synthetically, we consider the optimal values to be $\alpha = 0.1$.

Table 1: Effect of parameter α on the performance of EPFOA

α / TOP	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1%	37	**45**	42	40	40	40	39	40	40	39	39
5%	181	183	183	184	184	**185**	177	178	177	176	178
10%	**350**	341	334	317	304	295	287	289	288	284	284
15%	444	**451**	445	429	423	417	415	409	409	400	415
20%	**563**	542	537	532	531	528	529	530	530	529	544
25%	610	**628**	624	621	618	612	616	617	617	616	616

3.3 Comparison with other prediction measures

In order to demonstrate the advantage of our proposed EPFOA, we compare EPFOA with eight existing methods including DC, EC, IC, SC, NC, LAC, PeC and UDoNC. The essential proteins candidate population size p is set to 1274 (5093*25%=1274). The top 1, 5, 10, 15, 20 and 25% proteins are selected as candidate essential proteins, respectively. Then the prediction results are compared with the known essential proteins, and the experimental results are shown in Fig. 3. It can be observed that the percentage of essential proteins predicted by EPFOA is consistently higher than that achieved by the eight compared methods. Taking top 1% (top 51) predicted essential proteins as an example, 46 essential proteins are correctly identified by EPFOA while SC and EC have correctly predicted 24 essential proteins.

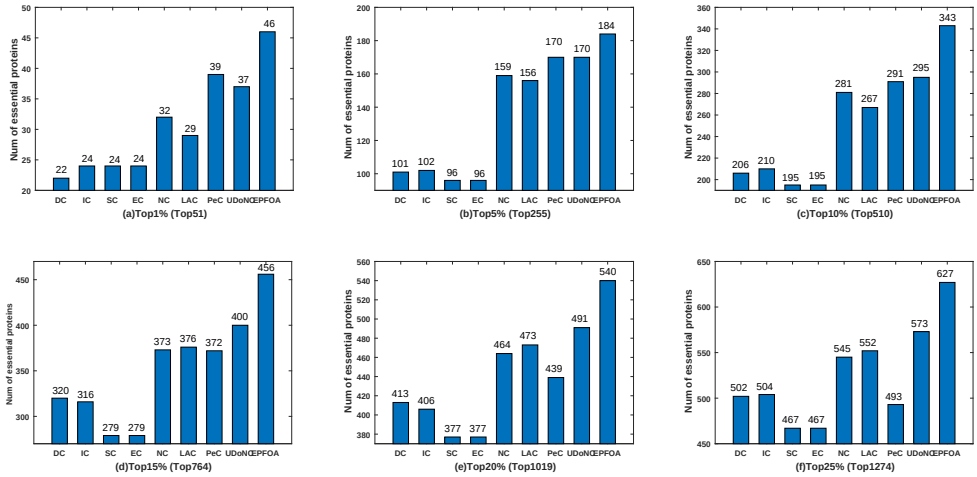

Figure 2: EPFOA compared with several existing methods.(a) Top 1% (Top 51), (b) Top 5% (Top 255), (c) Top 10% (Top 510), (d) Top 15% (Top 764), (e) Top 20% (Top 1019), (f) Top 25% (Top 1274).

3.4 Validation using six statistical measures

In order to evaluate the performance of EPFOA, we compare EPFOA with the other methods using six statistical measures: sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure, and accuracy (ACC). Each statistical measure is defined as follows:

$$SN = \frac{TP}{TP + FN}, \tag{17}$$

$$SP = \frac{TN}{TN + FP}, \tag{18}$$

$$PPV = \frac{TP}{TP + FP}, \tag{19}$$

$$NPV = \frac{TN}{TN + FN}, \tag{20}$$

$$F - measure = \frac{2 \times SN \times PPV}{SN + PPV}, \tag{21}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{22}$$

where TP is the number of essential proteins correctly identified as essential proteins, FP is the number of nonessential proteins mistakenly identified as essential proteins, TN is the number of nonessential proteins correctly identified as nonessential proteins, and FN is the number of essential proteins mistakenly identified as nonessential proteins. The comparison results between EPFOA and the other predicted essential proteins methods by six statistical measures performed on DIP are shown in Table 2. Obviously, we can see that EPFOA significantly outperforms all the compared methods.

Table 2: Comparison of EPFOA and the other methods in terms of SN, SP, PPV, NPV, F-measure, and ACC on the PPI networks.

Method	SN	SP	PPV	NPV	F-measure	ACC
DC	0.4302	0.8033	0.394	0.8258	0.4113	0.7178
EC	0.4002	0.7944	0.3666	0.8167	0.3826	0.704
IC	0.4319	0.8038	0.3956	0.8263	0.4129	0.7186
SC	0.4002	0.7944	0.3666	0.8167	0.3826	0.704
NC	0.467	0.8143	0.4278	0.8371	0.4465	0.7347
LAC	0.473	0.8161	0.4333	0.8389	0.4523	0.7374
PeC	0.4225	0.801	0.387	0.8235	0.4039	0.7143
UDoNC	0.491	0.8214	0.4498	0.8444	0.4695	0.7457
EPFOA	**0.5373**	**0.8352**	**0.4922**	**0.8586**	**0.5137**	**0.7669**

3.5 Comparison of the experimental results based on precision-recall curves

To further validate the performance of EPFOA, we study the Precision-Recall (PR) of EPFOA on the PPI networks and compare with the other methods. The precision and recall of the top n ranked proteins are defined as follow:

$$Precision(n) = \frac{TP(n)}{TP(n) + FP(n)}, \tag{23}$$

$$Recall(n) = \frac{TP(n)}{P}, \tag{24}$$

where $TP(n)$ is the number of true predicted essential proteins among the top n ranked proteins, $FP(n)$ is the number of false predicted essential proteins among the top n ranked proteins, P is the total number of essential proteins under consideration. Fig. 4 shows the PR curves of EPFOA and the other eight methods on the PPI networks. Obviously, EPFOA obtains the best performance, which demonstrates that the algorithm EPFOA works well in identifying essential proteins.

3.6 Validation using jackknife curves

A more general comparison between the proposed algorithm EPFOA and the eight previously proposed methods is tested by using a jackknife curves. The experimental results validated by Jackknife curves are shown in Fig. 5 the X-axis represents the proteins ranked in descending order from left to right according to the values computed using the corresponding methods, and the Y-axis represents the number of true essential proteins among the top n proteins, where n is the number along the X-axis. The area under the curve is always used to measure the generality of a method. As shown in Fig. 5, EPFOA clearly performs better than the other methods.

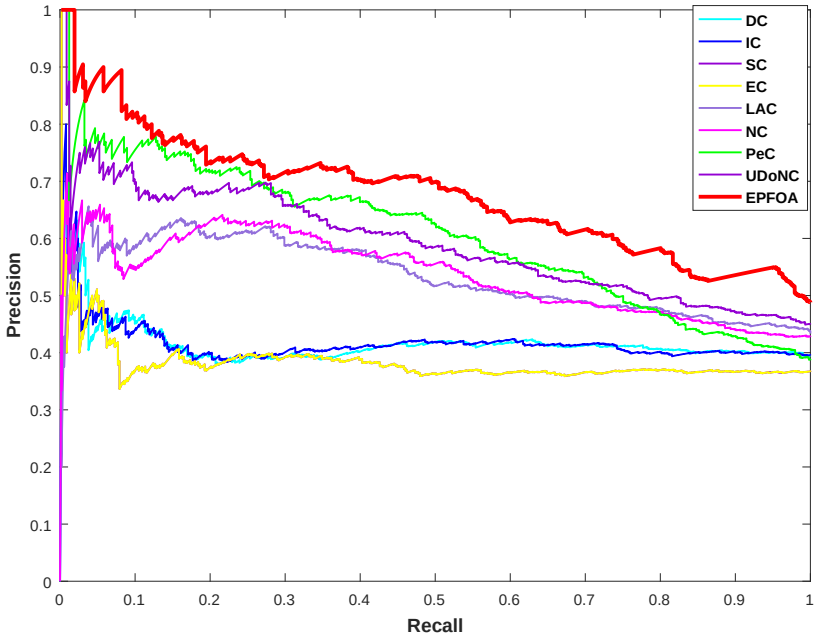

Figure 3: The PR curves of GSP and that of other methods.

3.7 The modularity of essential proteins predicted by EPFOA

Proteins usually perform tasks in biological system with protein complexes or functional modules and rarely act alone. Therefore, protein modularity may be an appropriate measurement to evaluate the significance of essential proteins identified by EPFOA. In order to examine the modularity of essential proteins identified by EPFOA, we compare EPFOA with DC that fully depend on network topology and PeC that combine network topology with biological information. We show the top 1% identified essential proteins of each method. As illustrated in Fig. 5, the number of essential proteins identified by EPFOA is higher than DC and PeC obviously. It also can be seen in Fig. 5, it is worthy to note that the essential proteins identified by EPFOA show more significant modularity than DC and PeC. It indicates that EPFOA is effective in identifying essential proteins.

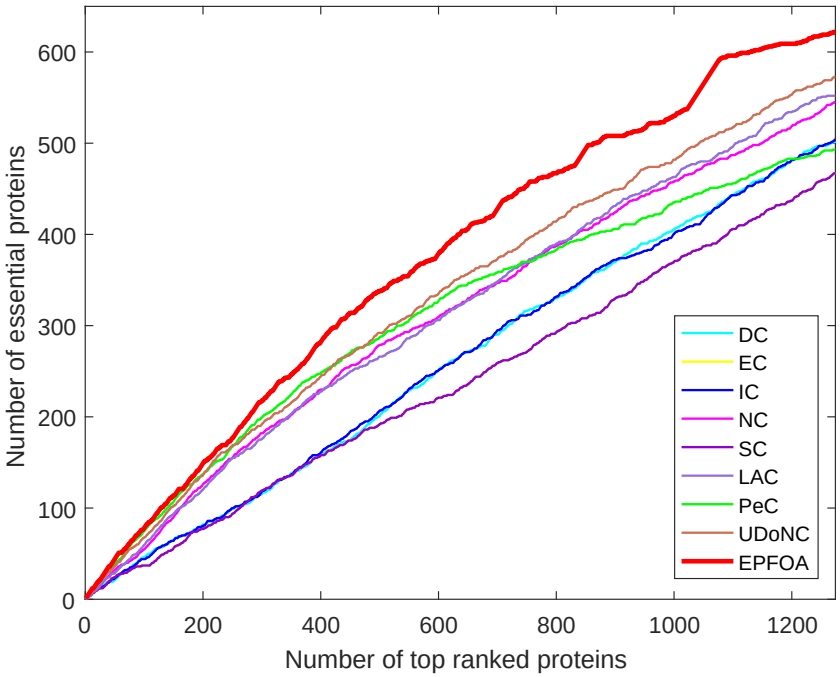

Figure 4: The jackknife curves of GSP and the other nine methods.

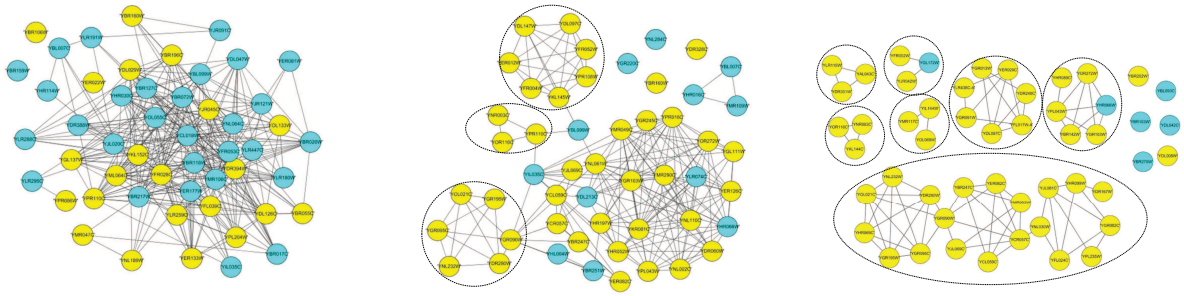

Figure 5: The modules formed by the top 1% identified essential proteins predicted by DC, PeC and EPFOA. Yellow circles are the essential proteins predicted by EPFOA, and blue circles are the non-essential proteins that incorrectly predicted.

4 Conclusion

It is believed that identification of essential proteins is very useful for understanding the minimal requirements for cellular life, and even the disease study and drug design. Although there are many methods have been proposed, it is still a challenge to improve the predicted precision. It is a strong potential way to use computational methods to identify essential proteins. In this study, we propose a novel algorithm EPFOA to boost the performance of essential proteins. We not only analyze the network topological characteristics in the dynamic PPI networks with GO annotation, but also analyze the biological characteristics with the subcellular location information. By comparing with other existing methods, FOCA can more effectively identify the essential proteins with the higher precision. As future work, it would be interesting to apply the EPFOA to other studies, such as gene and disease prediction.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (61672334, 91530320, 61502290, 61401263, and 61320106005) and the Innovation Scientists and Technicians Troop Construction Projects of Henan Province (154200510012).

Bibliography

[1] Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., Jensen, L. J. (2014); COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence, *Database, bau012*, 2014.

[2] Bocu, R., Tabirca, S. (2011); The Flag-based Algorithm - A Novel Greedy Method that Optimizes Protein Communities Detection, *International Journal of Computers Communications & Control*, 6(1), 33-44, 2011.

[3] Bonacich, P. (1987); Power and Centrality: A Family of Measures, *American Journal of Sociology*, 92(5), 1170-1182, 1987.

[4] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Schroeder, M. (1998); SGD: Saccharomyces Genome Database, *Nucleic Acids Research*, 26(1), 73, 1998.

[5] Consortium, G. O. (2015); Gene Ontology Consortium: Going Forward, *Nucleic Acids Research*, 43 (Database issue), 1049-1056, 2015.

[6] Consortium, G. O., Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., Buza, T. (2013); Gene Ontology Annotations and Resources, *Nucleic Acids Research*, 41(D1), 530-535, 2013.

[7] Cullen, L. M., Arndt, G. M. (2005); Genome-Wide Screening for Gene Function Using RNAi in Mammalian Cells, *Immunology Cell Biology*, 83(3), 217-223, 2005.

[8] Dzitac, I. (2015); Impact of Membrane Computing and P Systems in ISI WoS. Celebrating the 65th Birthday of Gheorghe Păun, *International Journal of Computers Communications & Control*, 10(5), 617–626, 2015.

[9] Estrada, E., Rodriguez-Velázquez, J. A. (2005); Subgraph Centrality in Complex Networks, *Physical Review E Statistical Nonlinear Soft Matter Physics*, 71(2), 056103, 2005.

[10] Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Dampelfeld, B. (2006); Proteome Survey Reveals Modularity of The Yeast Cell Machinery, *Nature*, 440(7084), 631-636, 2006.

[11] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., André, B. (2002); Functional Profiling of the Saccharomyces Cerevisiae Genome, *Nature*, 418(6896), 387, 2002.

[12] Gill, N., Singh, S., Aseri, T. C. (2014); Computational Disease Gene Prioritization: An Appraisal, *Journal of Computational Biology A Journal of Computational Molecular Cell Biology*, 21(6), 456-465, 2014.

[13] Hsing, M., Byler, K. G.,Cherkasov, A. (2008); The Use of Gene Ontology Terms for Predicting Highly-Connected 'Hub' Nodes in Protein-Protein Interaction Networks, *BMC Systems Biology*, 2(1), 1-14, 2008.

[14] Jeong, H., Mason, S. P., Barabási, A. L., Oltvai, Z. N. (2001); Lethality and Centrality in Protein Networks, *Nature*, 411(6833), 41-42, 2001.

[15] Jimenezsanchez, G., Childs, B., Valle, D. (2001); Human Disease Genes, *Nature*, 409(6822), 853-855, 2001.

[16] Lei, X., Wang, F., Wu, F. X., Zhang, A., Pedrycz, W. (2016); Protein Complex Identification Through Markov Clustering with Firefly Algorithm on Dynamic Protein-Protein Interaction Networks, *Information Sciences*, 329(6), 303-316, 2016.

[17] Lei, X., Wang, S., Pan, L. (2017); Predicting Essential Proteins Based on Gene Expression Data, Subcellular Localization and PPI Data. *Bio-inspired Computing: Theories and Applications: 12th International Conference, Proceedings of*, 92-105, 2017.

[18] Li, M., Lu, Y., Wang, J., Wu, F. X., Pan, Y. (2015); A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks, *IEEE/ACM Transactions on Computational Biology Bioinformatics*, 12(2), 372, 2015.

[19] Li, M., Wang, J., Chen, X., Wang, H., Pan, Y. (2011); A Local Average Connectivity-Based Method for Identifying Essential Proteins from the Network Level, *Computational Biology Chemistry*, 35(3), 143-150, 2011.

[20] Li, M., Wang, J., Wang, H., Pan, Y. (2012); Identification of Essential Proteins Based on Edge Clustering Coefficient, *IEEE/ACM Transactions on Computational Biology Bioinformatics*, 9(4), 1070, 2012.

[21] Li, M., Zhang, H., Wang, J. X., Pan, Y. (2012); A New Essential Protein Discovery Method Based on the Integration of Protein-Protein Interaction and Gene Expression Data, *BMC Systems Biology*, 6(1), 15, 2012.

[22] Luo, J., Kuang, L. (2014); A New Method for Predicting Essential Proteins Based on Dynamic Network Topology and Complex Information, *Computational Biology Chemistry*, 52(C), 34, 2014.

[23] Mewes, H. W., Frishman, D., Mayer, K. F. X., Münsterkötter, M., Noubibou, O., Pagel, P., St¨šmpflen, V. (2006); MIPS: Analysis and Annotation of Proteins from Whole Genomes in 2005, *Nucleic Acids Research*, 34 (Database issue), 169-172, 2006.

[24] Newman, M. E. J. (2005); A Measure of Betweenness Centrality Based on Random Walks, *Social Networks*, 27(1), 39-54, 2005.

[25] Pan, W. T. (2012); A New Fruit Fly Optimization Algorithm: Taking the Financial Distress Model as an Example, *Knowledge-Based Systems*, 26(2), 69-74, 2012.

[26] Pan, L., Păun, Gh. (2009); Spiking Neural P Systems with Anti-Spikes. *International Journal of Computers Communications & Control*, 4(3), 273–282, 2009.

[27] Pál, C., Papp, B., Hurst, L. D. (2003); Genomic function: Rate of Evolution and Gene Dispensability, *Nature*, 421(6922), 496-497, 2003.

[28] Păun, Gh. (2000); Computing with Membranes, *Journal of Computer and System Sciences*, 61(1), 108–143, 2000.

[29] Păun, Gh. (2016); Membrane Computing and Economics: A General View, *International Journal of Computers Communications & Control*, 11(1), 105–112, 2016.

[30] Peng, W., Wang, J., Cheng, Y., Lu, Y., Wu, F., Pan, Y. (2015); UDoNC: An Algorithm for Identifying Essential Proteins Based on Protein Domains and Protein-Protein Interaction Networks, *Computational Biology Bioinformatics IEEE/ACM Transactions on*, 12(2), 276-288, 2015.

[31] Przytycka, T. M., Singh, M., Slonim, D. K. (2010); Toward the Dynamic Interactome: It's about Time, *Briefings in Bioinformatics*, 11(1), 15-29, 2010.

[32] Qin, C., Sun, Y., Dong, Y. (2017); A New Computational Strategy for Identifying Essential Proteins Based on Network Topological Properties and Biological Information, *PLoS ONE*, 12(7), e0182031, 2017.

[33] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004); Defining and Identifying Communities in Networks, *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2658-2663, 2004.

[34] Ren, J., Wang, J., Li, M., Wang, H., Liu, B. (2011); Prediction of Essential Proteins by Integration of PPI Network Topology and Protein Complexes. *Information Bioinformatics Research and Applications - International Symposium, Isbra 2011, Changsha, China, May 27-29, 2011. Proceedings of*, 12-24, 2011.

[35] Roemer, T., Jiang, B., Davison, J., Ketela, T., Veillette, K., Breton, A., Marta, C. (2003); Large-Scale Essential Gene Identification in Candida Albicans and Applications to Antifungal Drug Discovery, *Molecular Microbiology*, 50(1), 167-181, 2003.

[36] Schlicker, A., Lengauer, T., Albrecht, M. (2010); Improving Disease Gene Prioritization Using the Semantic Similarity of Gene Ontology Terms, *Bioinformatics*, 26(18), i561, 2010.

[37] Song, B., Pan, L., Pérez-Jiménez, M. J. (2016); Cell-Like P Systems with Channel States and Symport/Antiport Rules, *IEEE Transactions on NanoBioscience*, 15(6), 555–566, 2016.

[38] Song, B., Song, T., Pan, L. (2017); A Time-Free Uniform Solution to Subset Sum Problem by Tissue P Systems with Cell Division, *Mathematical Structures in Computer Science*, 27(1), 17–32, 2017.

[39] Song, B., Zhang, C., Pan, L. (2017); Tissue-Like P Systems with Evolutional Symport/Antiport Rules, *Information Sciences*, 378, 177–193, 2017.

[40] Stephenson, K., Zelen, M. (1989); Rethinking centrality: Methods and Examples, *Social Networks*, 11(1), 1-37, 1989.

[41] Tang, X., Wang, J., Zhong, J., Pan, Y. (2014); Predicting Essential Proteins Based on Weighted Degree Centrality, *IEEE/ACM Transactions on Computational Biology Bioinformatics*, 11(2), 407-418, 2014.

[42] Tang, X. W. (2017); Predicting Essential Proteins Using a New Method, *Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, Proceedings of*, Part II, 301-308, 2017.

[43] Tang, Y., Li, M., Wang, J., Pan, Y., Wu, F. X. (2015); CytoNCA: A Cytoscape Plugin for Centrality Analysis and Evaluation of Protein Interaction Networks, *BioSystems*, 127, 67-72, 2015.

[44] Tu, B. P., Mcknight, S. L. (2005); Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes, *Science*, 310(5751), 115, 2005.

[45] Wang, J., Peng, X., Li, M., Luo, Y., Pan, Y. (2011); Active Protein Interaction Network and Its Application on Protein Complex Detection, *IEEE International Conference on Bioinformatics and Biomedicine*, 37-42, 2011.

[46] Wang, J., Peng, X., Peng, W., Wu, F. X. (2014); Dynamic Protein Interaction Network Construction and Applications, *Proteomics*, 14(4-5), 338-352, 2014.

[47] Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., Chen, C. F. (2007); A New Method to Measure the Semantic Similarity of GO Terms, *Bioinformatics*, 23(10), 1274, 2007.

[48] Wang, L., Zheng, X. L., Wang, S. Y. (2013); A Novel Binary Fruit Fly Optimization Algorithm for Solving The Multidimensional Knapsack Problem, *Knowledge-Based Systems*, 48(2), 17-23, 2013.

[49] Watts, D. J., Strogatz, S. H. (1998); Collective Dynamics of 'Small-World' Networks, *Nature*, 393(6684), 440, 1998.

[50] Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bussey, H. (1999); Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis, *Science*, 285(5429), 901-906, 1999.

[51] Wuchty, S. (2001); Scale-Free Behavior in Protein Domain Networks, *Molecular Biology Evolution*, 18(9), 1694, 2001.

[52] Wuchty, S., Stadler, P. F. (2003); Centers of Complex Networks, *Journal of Theoretical Biology*, 223(1), 45, 2003.

[53] Yan, W., Sun, H., Wei, D., Enrico, B., Gabriella, V., Ying, X., Liang, Y. (2014); Identification of Essential Proteins Based on Ranking Edge-Weights in Protein-Protein Interaction Networks, *PLoS ONE*, 9(9), e108716, 2014.

[54] Zeng, X., Lin, W., Guo, M., Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools, *PLoS Computational Biology*, 13(6), e1005420, 2017.

[55] Zhang, R., Lin, Y. (2009); DEG 5.0, A Database of Essential Genes in both Prokaryotes and Eukaryotes, *Nucleic Acids Research*, 37 (Database issue), D455, 2009.

[56] Zhang, X. F., Dai, D. Q., Ouyang, L., Yan, H. (2014); Detecting Overlapping Protein Complexes Based on a Generative Model with Functional and Topological Properties, *BMC Bioinformatics*, 15(1), 186, 2014.

[57] Zhang, Y., Lin, H., Yang, Z., Wang, J. (2013); Construction of Ontology Augmented Networks for Protein Complex Prediction, *PLoS ONE*, 8(5), : e62077, 2013.

[58] Zhao, B., Wang, J., Li, M., Wu, F. X., Pan, Y. (2014); Detecting Protein Complexes Based on Uncertain Graph Model, *IEEE/ACM Transactions on Computational Biology Bioinformatics*, 11(3), 486-497, 2014.

[59] Zhu, C., Wu, C., Aronow, B. J., Jegga, A. G. (2014); Computational Approaches for Human Disease Gene Prediction and Ranking, *Advances in Experimental Medicine Biology*, 799, 69, 2014.