Lei et al. BMC Systems Biology 2018, 12(Suppl 4):46
https://doi.org/10.1186/s12918-018-0573-y

BMC Systems Biology

RESEARCH Open Access

Improved flower pollination algorithm for identifying essential proteins

Xiujuan Lei[1*], Ming Fang[1], Fang-Xiang Wu[2] and Luonan Chen[3]

From The 11th International Conference on Systems Biology (ISB 2017)
Shenzhen, China. 18-21 August 2017

Abstract

Background: Essential proteins are necessary for the survival and development of cells. The identification of essential proteins can help to understand the minimal requirements for cellular life and it also plays an important role in the disease genes study and drug design. With the development of high-throughput techniques, a large amount of protein-protein interactions data is available to predict essential proteins at the network level. Hitherto, even though a number of essential protein discovery methods have been proposed, the prediction precision still needs to be improved.

Methods: In this paper, we propose a new algorithm, improved Flower Pollination algorithm (FPA) for identifying Essential proteins, named FPE. Different from other existing essential protein discovery methods, we apply FPA which is a new intelligent algorithm imitating pollination behavior of flowering plants in nature to identify essential proteins. Analogous to flower pollination is to find optimal reproduction from the perspective of biological evolution, and the identification of essential proteins is to discover a candidate essential protein set by analyzing the corresponding relationships between FPA algorithm and the prediction of essential proteins, and redefining the positions of flowers and specific pollination process. Moreover, it has been proved that the integration of biological and topological properties can get improved precision for identifying essential proteins. Consequently, we develop a *GSC* measurement in order to judge the essentiality of proteins, which takes into account not only the Gene expression data, Subcellular localization and protein Complexes information, but also the network topology.

Results: The experimental results show that FPE performs better than the state-of-the-art methods (DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON) in terms of the prediction precision, precision-recall curve and jackknife curve for identifying essential proteins and also has high stability.

Conclusions: We confirm that FPE can be used to effectively identify essential proteins by the use of nature-inspired algorithm FPA and the combination of network topology with gene expression data, subcellular localization and protein complexes information. The experimental results have shown the superiority of FPE for the prediction of essential proteins.

* Correspondence: xjlei@snnu.edu.cn
[1]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
Full list of author information is available at the end of the article

Background

Essential proteins are indispensable in the cellular life for the survival or development of an organism. Even though the deletion of only one of these proteins will cause a lethal flaw on an organism [1]. Studies have shown that essential proteins are related to disease genes [2] and contribute to the prediction of drug targets [3]. Therefore, identifying essential proteins is not only conducive to the understanding of minimal requirements for cellular life, but also important for the study of disease genes [4].

The traditional methods of identifying essential proteins are biological experiments, such as gene knockouts [5], RNA interference [6] and conditional knockouts [7], these biological experiment discovery methods are accurate, but time-consuming, low efficiency and expensive. Up to now, many computational methods for predicting essential proteins have been proposed. Particularly with the rapid development of high-throughput technologies, such as yeast two-hybrid screens [8], tandem affinity purification [9] and mass spectrometric analysis [10], a large amount of protein interaction data is detected, which provide new possibilities for the identification of essential proteins. It is becoming increasingly important to predict essential proteins by computational methods based on protein interaction data.

The identification of essential proteins based on protein-protein interaction (PPI) networks by using various topological properties is a very hot topic. Until now, many essential protein discovery methods have been proposed, while most of these methods are based on proteins with highly connected neighbors tend to be essential, named the "centrality-lethality" rule [11], such as Degree Centrality (DC) [11], Betweenness Centrality (BC) [12, 13], Closeness Centrality (CC) [14], Subgraph Centrality (SC) [15], Eigenvector Centrality (EC) [16], Information Centrality (IC) [17]. Moreover, there are also two neighborhood-based methods: Neighborhood Centrality (NC) [18] and Local Average Connectivity-based method (LAC) [19].

The above methods depend on the PPI networks to identify essential proteins and have made great progresses in the essential protein discovery tasks. However, it is still a challenge to improve the prediction precision owing to the PPI networks obtained by high-throughput technologies contain many false positives which may greatly affect the precision of identification of essential proteins [20]. Furthermore, these methods neglect the inherent biological significance of essential proteins. Hence, to reduce the effect of noise in the PPI networks, the researchers have tried to achieve higher precision of identifying essential proteins by integrating other biological information. For example, ION [21] used the orthologous information with the PPI networks. A method PeC [22] integrated gene expressions and PPI networks, Peng et al. [23] proposed UDoNC by integrating domains and PPI networks. Zhong et al. [24] used a feature selection method by collecting 26 different biological and topological features to identify essential proteins, SON [25] integrated subcellular localization, orthology and PPI networks, United complex Centrality (UC) [26] utilized protein complexes information to predict essential proteins. Besides the methods mentioned above, some researchers integrated topological or biological information to construct dynamic networks. For example, Xiao et al. [27] constructed an active PPI network to predict essential proteins.

Flower pollination algorithm (FPA) [28] is a nature-inspired intelligent optimization algorithm that considers the characteristics of flower pollination, which proposed by Yang in 2012. There are two main patterns of the pollination process viz. abiotic pollination and biotic pollination. Pollinators can be biotic or abiotic depending on the type of pollination. For biotic pollination, pollinators are some animals such as insects and birds. This type of pollination is called as global pollination. About 90% of the pollination is biotic in nature. For abiotic pollination, pollinators are natural resources such as wind, water and soil. This type of pollination is local pollination. The global pollination and local pollination are two main steps of FPA, these two steps are regulated by the switch probability. FPA has been applied in various practical problems such as clustering [29], feature selection [30] and multi-objective optimization problem [31]. Consequently, the efficiency of FPA makes it possible for addressing the problem of predicting essential proteins.

In this study, we develop a new algorithm, named FPE, based on improved FPA to identify essential proteins by integrating gene expression data, subcellular localization and protein complexes information with the topological properties of PPI networks. Different from other essential protein discovery methods that already exist, we take advantage of the improved version of FPA to provide a new perspective for the identification of essential proteins. Also, our algorithm FPE integrates biological properties and topological properties of PPI networks to assess the essentiality of proteins and further improve the performance of prediction results. In order to evaluate the effectiveness of the proposed algorithm FPE, we apply it to the PPI networks and compare with ten previous essential protein discovery methods: DC [11], SC [15], IC [17], EC [16], LAC [19], NC [18], PeC [22], WDC [32], UDoNC [23] and SON [25]. The experimental results on the identification of yeast essential proteins show that FPE outperforms the ten previously proposed methods in terms of the prediction precision, as well as the precision-recall curve and the jackknife curve. The

modularity of proteins and the effect of the parameter p on the prediction results is also discussed.

The rest of this paper is organized as follows. We first introduce the basic knowledge of FPA. Then we present how to combine FPA with the identification of essential proteins, as well as the measurement for evaluating the essentiality of proteins. Next, the performance of FPE is validated by using a series of comparison experiments and the analysis of experimental results are also described. We conclude this study at the end.

Methods

The FPE algorithm is used to identify essential proteins on the basis of the combination of improved flower pollination algorithm with the gene expression data, subcellular localization and protein complexes information.

Flower pollination algorithm (FPA)

FPA [28] is a population-based global optimization technique, which is motivated by the pollination process of flowers. Pollination can be divided into two types, i.e., self-pollination and cross-pollination. Self-pollination takes place between the flowers of the same plant species while cross-pollination can occur from the flowers of different plant species. Biotic pollinators such as insects and birds can fly long distances causing cross-pollination, which thus can be considered as global pollination. Abiotic pollinators are the natural resources such as wind and water that are unable to take away pollens to long distances causing self-pollination, which is local pollination. The local pollination and global pollination interchange is controlled by a parameter $p \in [0, 1]$ defined by so called switch probability. Table 1 shows the basic knowledge of FPA.

In FPA, it is assumed that each plant only has one flower and each flower only has one pollen gamete for simplicity. Therefore, a flower or pollen gamete represented by a position vector that denotes as a candidate solution of the optimization problem. Flower pollens will be transferred in global pollination and local pollination.

In the global pollination, pollens are carried to long distances by pollinators, such as insects, because these pollinators can fly and move in a longer distance.

$$x_i^{t+1} = x_i^t + F\left(x_i^t - gbest\right) \tag{1}$$

where x_i^t is the pollen i at iteration t, and $gbest$ is the current best solution which is found among all solutions at the current iteration. The parameter F is the strength of the pollination, namely a step size, we use a Lévy flight to represent that insects move over a long distance with various distance steps. That is, $F > 0$ and follows Lévy distribution:

$$F \sim \frac{\lambda\Gamma(\lambda)\,\sin(\pi\lambda/2)}{\pi}\frac{1}{s^{1+\lambda}}, (s \gg s_0 > 0) \tag{2}$$

where $\Gamma(\lambda)$ is the standard gamma function, and this distribution is valid for large steps s > 0.

The local pollination occurs within a limited range thanks to pollinators like wind or water, which can be defined as:

$$x_i^{t+1} = x_i^t + \phi\left(x_j^t - x_k^t\right) \tag{3}$$

where x_j^t and x_k^t are pollen from the different flowers of the same plant species. This substantially models the flower constancy in a limited neighborhood. Mathematically, if x_j^t and x_k^t come from the same plant species or select from the same population, this can be a local random walk if ϕ follows the uniform distribution in [0, 1].

From the biological evolution point of view, it is a fact that the aim of the flower pollination is achieving the optimal reproduction of the flowering plants. The pollinator's movement towards the optimal solution is represented by the global optimum found by FPA, namely, the most suitable reproduction and pollination are found, which is represented by *gbest*.

Taking the basic principle of FPA algorithm into consideration, we design a new FPE algorithm, which is an improved version of FPA algorithm to identify essential proteins. In the FPE algorithm, the position of a pollen is represented as a set of candidate essential proteins contained Q proteins.

Table 1 The basic knowledge of FPA

	Local pollination	Global pollination
Types	Self-pollination (Abiotic)	Cross-pollination (Biotic)
Flowers	Same plant species	Different plant species
Pollinators	Wind, water	Insects, birds

Improved flower pollination algorithm for essential proteins identification (FPE)

In this section, we will use an improved FPA algorithm to develop a new algorithm, named FPE. Table 2 illustrates the corresponding relationships between FPA algorithm and the identification of essential proteins.

Pollen's position

A PPI network is described by an undirected graph $G(V, E)$, where V denotes a set of nodes that are proteins and E denotes a set of edges of PPI network.

In the basic FPA, the position of a flower is viewed as a candidate solution for the optimization problem that needs to be solved. Nevertheless, in our FPE algorithm, the positions of flowers are redefined as the candidate sets of essential proteins and each candidate set consists of Q proteins. A pollen can be encoded as a candidate

Table 2 The corresponding relationships between FPA algorithm and the identification of essential proteins

FPA algorithm	The identification of essential proteins
Pollen	A candidate essential protein set
Pollen's position	The serial numbers of Q candidate essential proteins
Fitness function	The measurement of proteins' essentiality
Pollination process	The process of identifying essential proteins

essential protein set $H = \{h_1, h_2,...h_Q\}$, where each of these elements represents the serial number of a protein.

Pollination process

We redesign the formulas for updating the pollen's positions considering that the basic FPA algorithm is continuous form while our proposed algorithm is discrete form.

In the global pollination of FPA algorithm represented by formula (1), on the one hand, pollen constantly move to the global optimal solution, on the other hand, Lévy distribution is used to make the pollen move in a longer distance.

Inspired by the global pollination of basic FPA, we consider its two aspects mentioned above in a comprehensive way to update the position of pollen, hence, in the global pollination of our FPE algorithm, the position of pollen is defined as follows:

$$L_i^{t+1} = cat\left(dim,\ L_i^{t'},\ RANDOM \right) \tag{4}$$

where the *cat* operation is the function that forms the position vector of a pollen. The value of *dim* is 1, which indicates that two position vectors obtained by $L_i^{t'}$ and *RANDOM* are concatenated in a column in our FPE algorithm. Then the *RANDOM* indicates that a global search in V is performed to update the position of pollen. Finally, the pollen's new position is obtained by using *cat* function which can connect the position vectors obtained by $L_i^{t'}$ and *RANDOM* to guarantee that the pollen's new position L_i^{t+1} not only keeps moving towards the global optimal solution, but also searches in a global scope. $L_i^{t'}$ can be represented as follows:

$$L_i^{t'} = intersect\left(L_i^t, Gbest\right) \tag{5}$$

where the *intersect* function denotes that the elements in L_i^t intersect with the elements in *Gbest*. Here, the elements in *Gbest* are those of a certain proportion in *Gbest*. This can mimic the update of pollen's position in the basic FPA algorithm so that the pollen is constantly approaching the global optimum *Gbest*.

In the local pollination, the pollen's position remains unchanged, which is represented as:

$$L_i^{t+1} = L_i^t \tag{6}$$

The overall flow of the FPE algorithm is shown in Fig. 1.

The measurement of the essentiality of proteins (GSC)

According to the aforementioned analysis and the conclusions from previous studies, we have known that the position of a flower can be viewed as a candidate solution set of essential proteins and each candidate set consists of Q proteins. Then the measurement of the proteins' essentiality should be needed. Accordingly, we define a new measurement to determine the essentiality of a pollen represented by the Q candidate essential proteins, called *GSC*, which consists of three types of information, gene expression data, subcellular localization and protein complexes information. *GSC* can be used to assess the quality of each candidate solution, which corresponds to the fitness function of FPA.

The *GSC* is a measure of combining the centrality measure *PeC*, subcellular localization *SL* and protein complexes *PC*. Subsequently, we will introduce them in detail.

For a candidate set $H = \{h_1, h_2,...h_Q\}$, where each element h_i denotes a candidate essential protein, its essentiality is evaluated by $GSC(H)$:

$$GSC(H) = \sum_{i=1}^{Q} \{SL \times [\alpha \times PeC + (1-\alpha) \times PC]\} \tag{7}$$

where the parameter α is a constant between [0, 1], which is used to adjust the proportions of three types of information. When $\alpha = 0$, only the information about protein complexes and subcellular localization is considered, and when $\alpha = 1$, only the information about subcellular localization and gene expression data with the PPI network is considered. To start with, we use $\alpha = 0.5$ as an initial value and then it has been certified that $\alpha = 0.6$ works better for most applications from our parametric analysis.

Next, we will introduce how to integrate gene expression data, subcellular localization and protein complexes information with the topological properties of PPI networks to determine the proteins' essentiality.

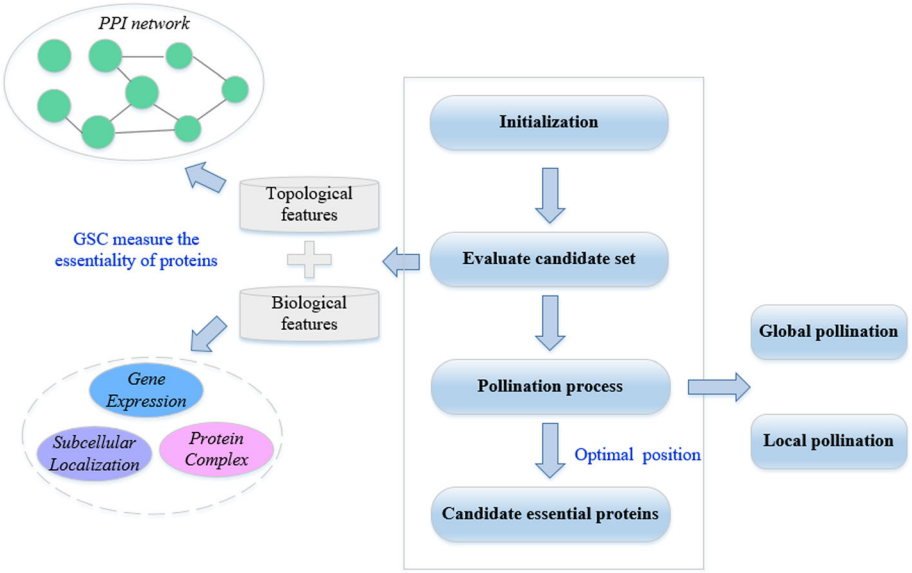

Fig. 1 The overall flow of FPE algorithm

PeC

As we know the edge clustering coefficient (*ECC*) can describe the closeness of two connected nodes in a PPI network. The *ECC* is an important measure to represent the topological properties of PPI networks and it has been proved that *ECC* has a good performance in identifying protein complexes and essential proteins. Furthermore, Pearson correlation coefficient (*PCC*) is a measure that is used to evaluate how likely two interacting proteins are co-expressed. Based on gene expression data and protein-protein interaction data, the centrality method PeC [22] using *ECC* and *PCC* is a very effective essential protein discovery method.

Given a PPI network $G(V, E)$, where a node $i \in V$ denotes a protein and an edge $(i, j) \in E$ connecting node i and node j, its edge clustering coefficient $ECC(i, j)$ can be defined by the following formula:

$$ECC(i,j) = \frac{|N_i \cap N_j|}{min\{d_i, d_j\}} \tag{8}$$

where N_i and N_j denote the set of all neighbors of protein i and j, respectively, d_i and d_j denote the degree of protein i and j, respectively.

$X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are two sequences of gene expressions, *PCC* is calculated by:

$$PCC(i,j) = \frac{\sum_{i=1}^{T}(x_i - \mu(x))(y_i - \mu(y))}{\sqrt{\sum_{i=1}^{T}(x_i - \mu(x))^2 \cdot \sum_{i=1}^{T}(y_i - \mu(y))^2}} \tag{9}$$

The value of *PCC* is between − 1 and 1. The probability that protein i and j are co-clustered can be calculated as follows:

$$p_c(i,j) = ECC(i,j) \times PCC(i,j) \tag{10}$$

Given a protein i, its $PeC(i)$ is defined as follows:

$$PeC(i) = \sum_{v \in n_i} p_c(i,v) \tag{11}$$

where n_i denotes the set of all neighbors of protein i. It is obvious that a protein gets higher values of *ECC* and *PCC*, it will obtain a relatively higher value of *PeC* and thus tends to be an essential protein.

Subcellular localization

It is well known that subcellular localization is a significant property of essential proteins and a protein must be appeared in an appropriate subcellular location. The basic idea that we use subcellular localization information to identify essential proteins is that essential proteins appear more often in certain subcellular locations [25]. Consequently, we hypothesize that the proteins, which are in the same subcellular location as the known essential proteins are tend to be essential.

In order to prove our hypothesis, we analyze the relationship between the final subcellular localization dataset R and the known essential protein dataset, the relationship dataset is defined as S, then each of the 11 subcellular locations is called S_i, as shown in Fig. 2.

From Fig. 2 we can see that the known essential proteins appear most frequently in the Nucleus and it shows that the proteins in the Nucleus are more likely to be essential proteins. However, few of the known essential proteins appear in the Peroxisome, indicating that the proteins appear in the Peroxisome are essential proteins with a small probability.

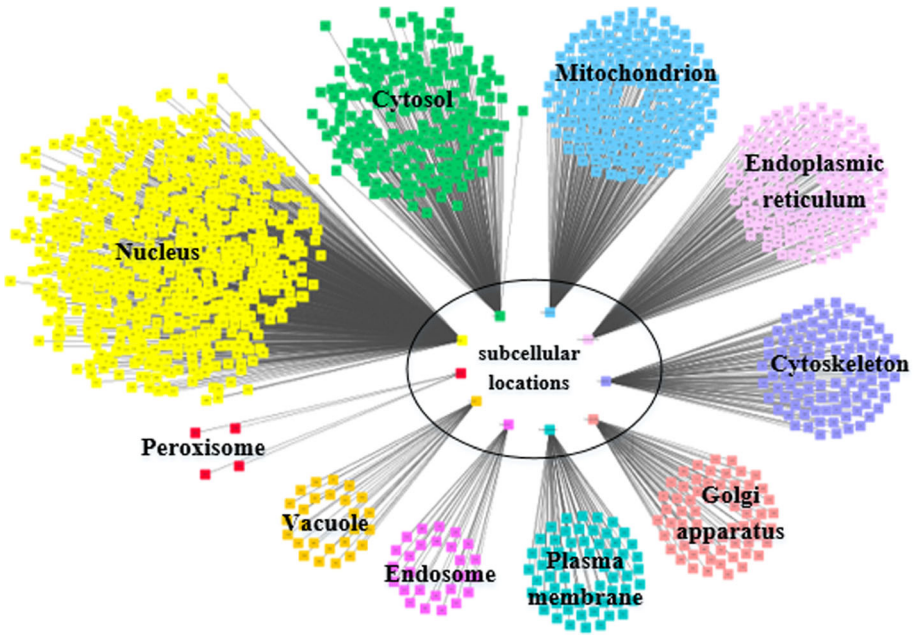

Fig. 2 The relationship dataset. The nodes with different colors represent the distribution of the known essential proteins in different subcellular locations. There should have been 11 subcellular locations, but the known essential proteins only appear in the 10 subcellular locations other than the Extracellular space

If protein i exists in R, we calculate the frequency where each of the 11 subcellular locations appears, the corresponding score for each location is denoted as $F_i(r)$ by the following formula:

$$F_i(r) = \begin{cases} \dfrac{S_r}{length(S)}, if\ i \cap R \\ 0 \quad , otherwise \end{cases} \tag{12}$$

where $length(S)$ is the number of subcellular location records of the known essential proteins in the dataset R.

An efficient computational method is obtained to determine the subcellular localization score of a protein from the above-mentioned analysis. For this reason, we use the subcellular localization information to devise subcellular localization scores of proteins. For a given protein i, its subcellular localization score $SL(i)$ is defined as the sum of scores of all the subcellular locations in which it appears.

$$SL(i) = \sum_{C(i)} F_i(r) \tag{13}$$

where $C(i)$ denotes the set of corresponding subcellular locations in which protein i in the dataset R. Note that a protein may appear in multiple subcellular locations.

Protein complexes

Proteins often bind together to constitute protein complexes for carrying out their functions [33]. Based on the observation that essentiality is more likely to be the product of a protein complex rather than an individual protein [34] and proteins existed in complexes are tend to be essential compared to the proteins not appeared in complexes [26], in this subsection, we use two different protein complex datasets obtained from [35] that contain 270 and 425 complexes, respectively. After removing the repeated protein complexes, we collect 538 known protein complexes into a dataset, named P, denotes as $P = \{P_1, P_2, \ldots, P_k\}$.

A protein's complex score is evaluated by the number of times it appears in the known protein complexes. For a given protein i, its protein complex score $PC(i)$ is defined as follows:

$$PC(i) = \sum_{k=1}^{M} T_i(k) \tag{14}$$

$$T_i(k) = \begin{cases} 1, & if\ i \in P_k \\ 0, & otherwise \end{cases} \tag{15}$$

where M is the number of the known protein complexes. If a protein exists in the known protein complexes, the value of its PC is the number of times it appears in the known protein complexes. If a protein does not appear in any protein complexes, the value of its PC is 0. We can clearly find that for a given protein i, it appears in more protein complexes and can get a higher value of PC.

Pseudocode of FPE

Our proposed new algorithm FPE adopts the improved version of FPA algorithm by simulating the pollination process of flowers to identify essential proteins.

Algorithm 1: FPE essential proteins identification

Input: A PPI network $G = (V, E)$, A population of n flowers or pollen and the length of each pollen is Q, Switch probability $p \in [0, 1]$, *Maxiter*: the maximal iterations of external loop, Gene expression data, Subcellular localization information, Protein complexes information, Parameter α.

Output: set (Top Q essential proteins)

Initialization: The Q proteins with the highest degree in the PPI network G are selected as the pollen's position $H = \{h_{i1}, h_{i2}, \ldots h_{iQ}\}$ and each element is denoted as h_{ij}.

Find the best solution *Gbest* in the initial population

for each i do

  for each j do

    if rand > a perturbance factor

      h_{ij}= Substitute(h_{ij});

    end if

  end for

end for

Calculate the essentiality of proteins via $GSC(H) = \sum_{i=1}^{Q}\{SL \times [\alpha \times PeC + (1-\alpha) \times PC]\}$

while ($t <$ *Maxiter*)

  for $i = 1 : n$ (all n pollen in the population)

    if rand > p

      Global pollination via $L_i^{t+1} = cat(dim,\ L_i^{t'},\ RANDOM)$

    else

      Local pollination via formula $L_i^{t+1} = L_i^t$

    end if

    Evaluate new solutions via $GSC(H) = \sum_{i=1}^{Q}\{SL \times [\alpha \times PeC + (1-\alpha) \times PC]\}$

    If new solutions are better, update them in the population

    end for

      Find the current best solution *Gbest*

    end while

    Output the global optimum (A candidate essential protein set)

In FPE, first the Q proteins with the highest degree in the PPI network are selected as initial position of pollen to improve efficiency of FPE algorithm and using a perturbance factor that is a constant between [0, 1] to make sure that each pollen is different. Then, the measurement *GSC* is used to assess the quality of each candidate set. We redefine the update rules of the pollen' s position and each pollen is updated by tailing the global optimal solution in each iteration since the global optimum can be viewed as a reliable guide for pollen to search better solution. The pseudo code of improved flower pollination algorithm for identifying essential proteins is described shown in Algorithm 1.

Switch probability p can be used to switch between global pollination and local pollination. The effect of p on the results will be discussed in experimental section and our experimental results demonstrate that the better result can be obtained when the value of p is 0.3.

Results and discussion

In order to test whether our proposed algorithm FPE is effective for identifying essential proteins, we apply it to identify essential proteins of *S. cerevisiae*. First, we use the FPE algorithm to identify essential proteins and compare with ten other essential protein discovery methods: DC [11], SC [15], IC [17], EC [16], LAC [19], NC [18], PeC [22], WDC [32], UDoNC [23] and SON [25]. Then, the performance of FPE is evaluated in terms of the PR curve and the jackknife curve. After that, the modularity of proteins is used to confirm the performance of FPE. Finally, the effect of parameter p on the experimental results of proposed algorithm FPE is discussed.

Experimental data

All the experiments in this study are based on the PPI network data of *S. cerevisiae* to identify essential proteins because it is the most complete data and has widely been used in the study of predicting essential proteins. The PPI network dataset of *S. cerevisiae* is downloaded from the DIP database [36]. The final yeast PPI network includes 5093 proteins and 24,743 interactions after the repeated interactions and the self-connecting interactions are removed. Other types of biological information used in this study are described as follows:

Gene expression data: The yeast gene expression data, GSE3431, are obtained from the Gene Expression Omnibus (GEO) database [37]. A total of 7074 gene products are used in our experiment.

Subcellular localization data: The protein subcellular localization dataset of *S. cerevisiae* is obtained from the subcellular localization database of COMPARTMENTS [38]. The yeast proteins have a total of 11 subcellular localizations as follows: Cytoskeleton, Golgi apparatus, Peroxisome, Cytosol, Endosome, Mitochondrion, Plasma membrane, Nucleus, Extracellular space, Vacuole, Endoplasmic reticulum. After preprocessing, it still includes 6892 subcellular localization records.

Protein complexes data: We integrate two real protein complex sets into one protein complex set. These two sets from [35] contain 270 and 425 complexes, respectively. The final known protein complex dataset contains 538 complexes, gathered from these two complex sets by removing the repeated protein complexes, named *P*.

Standard essential protein set: A list of the known essential proteins of *S. cerevisiae* is collected from the following databases: MIPS (Mammalian Protein-Protein Interaction Database) [39], SGD (Saccharomyces Genome Database) [40], DEG (Database of Essential Genes) [41], and SGDP (Saccharomyces Genome Deletion Project) [42]. There are 1285 essential proteins are collected in this dataset.

Comparison of FPE with other essential protein discovery methods

To compare the performance of FPE with other previous essential protein discovery methods DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON, we first apply these ten methods on the yeast PPI network. Then, similar to most methods of predicting essential proteins, we rank all the proteins in descending order in the PPI network and select the top 100, top 200, top 300, top 400, top 500, and top 600 proteins as essential candidates. Finally, according to the standard essential protein dataset, the number of true essential proteins is detected by ten competing methods DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON in the yeast PPI network.

For our proposed FPE algorithm, first the global optimum, i.e., a candidate essential protein set is obtained based on the improved FPA, then we rank Q proteins in descending order from the obtained candidate set by using our redefined measurement GSC and select the top 100, top 200, top 300, top 400, top 500, and top 600 proteins as essential candidates, finally, we achieve the number of true essential proteins predicted by FPE.

The comparison results are shown in Fig. 3. From Fig. 3 we can see that FPE has a better performance compared with the other ten essential protein discovery methods for predicting essential proteins from the yeast PPI networks. The number of true essential proteins identified by FPE is consistently higher than those generated by the ten previously proposed methods: DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON from top 100 to top 600 proteins. By choosing top 100 proteins, FPE can obtain a prediction precision of 89%. Especially compared to LAC, the improvements of FPE are 50.85, 36.67, 28.41, 27.19, 30.08 and 29.08% from top 100 to top 600 proteins, respectively.

It should be noted that the identification result of each time in our algorithm FPE with randomness due to the characteristics of the intelligent algorithm itself, but the result of each time is basically maintained within a stable range.

Validation in terms of the precision-recall curve

In this subsection, we use precision-recall (PR) curve that is a common methodology to evaluate the performance of the proposed algorithm FPE. A comparison of FPE with the ten methods DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON for predicting essential proteins from the yeast PPI networks by using the PR curve is shown in Fig. 4.

From Fig. 4 we can see that the PR curve of FPE obtains the better result compared to the PR curves of ten other previously proposed essential protein discovery methods: DC, SC, IC, EC, LAC, NC, PeC, WDC,

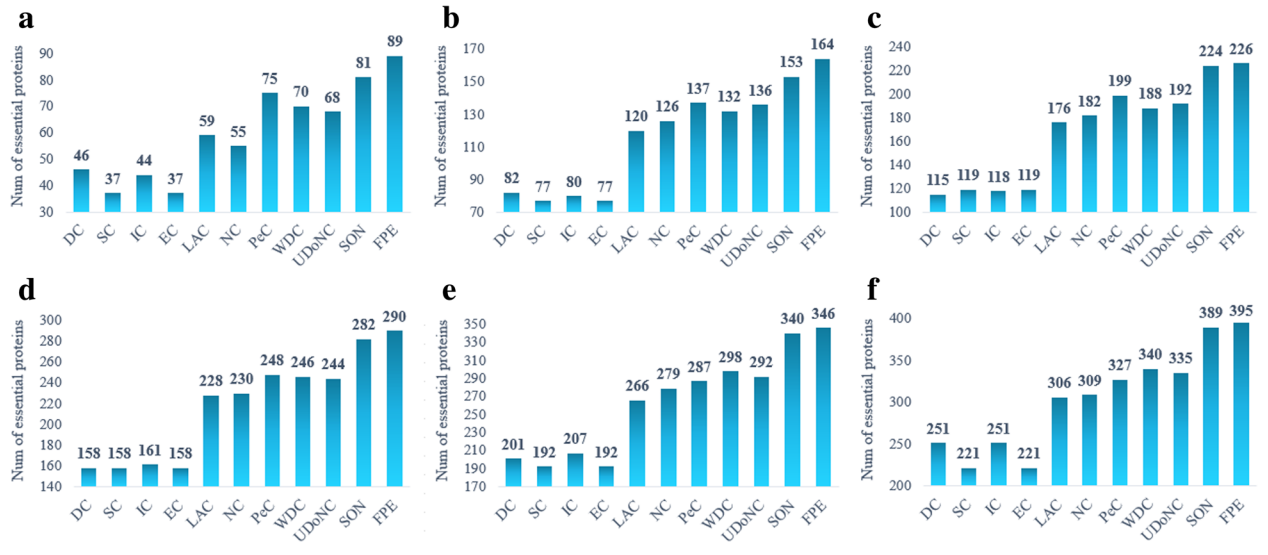

Fig. 3 Comparison of FPE with other essential protein discovery methods. The ten competing methods are DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON. (**a**), (**b**), (**c**), (**d**), (**e**), and (**f**) show the results of these methods when select top 100, 200, 300, 400, 500, and 600 proteins as essential candidates, respectively. Note that the number of identification presented here is the result of FPE algorithm running ten times and then averaging the number of ten times

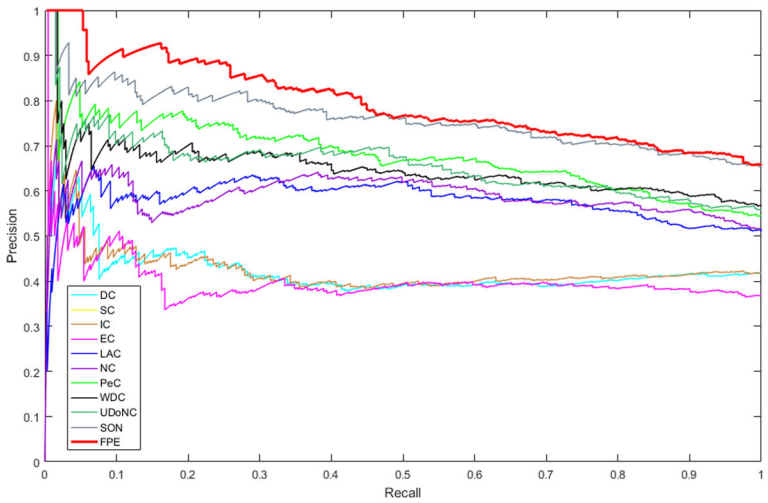

Fig. 4 Validation in terms of the precision-recall curve. Comparison of DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC, SON and FPE based on the validation of PR curve

UDoNC and SON. The PR curves of EC and SC are almost the same.

We have known our algorithm FPE with randomness, but the result of each time remains in a stable range, the PR curve of the FPE algorithm here is randomly selecting from the ten times running results.

Validation in terms of the jackknife curve

To evaluate the effectiveness of FPE more generally, we further use the jackknife curve to illustrate the prediction results of DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC, SON and our proposed algorithm FPE. The results are shown in Fig. 5, the x-axis represents the number of proteins are ranked by each essential protein discovery method and the y-axis is the cumulative count of true essential proteins.

The areas under the curves can measure the performances of the above-mentioned methods. As shown in Fig. 5, the jackknife curve of FPE is better than the other methods DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC and SON for identifying essential proteins from the yeast PPI networks. It demonstrates that FPE is more effective than other ten methods for identifying essential proteins. The jackknife curves of EC and SC are almost the same. The jackknife curve uses the same running results of the FPE algorithm as the PR curve.

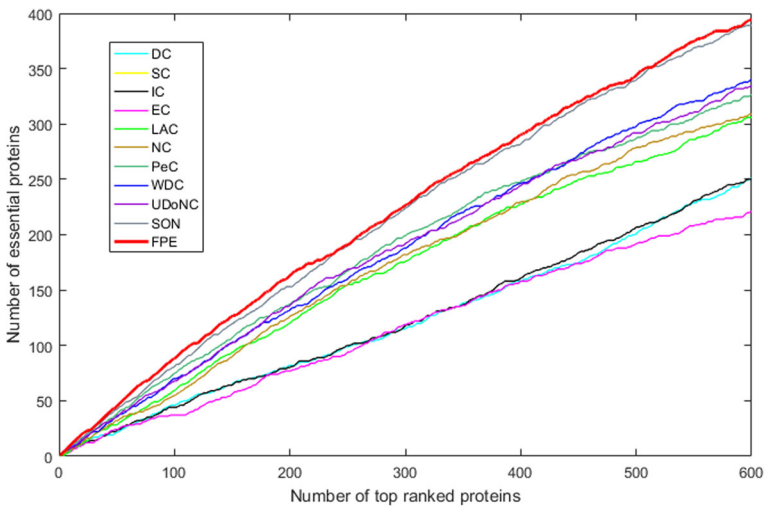

Fig. 5 Validation in terms of the jackknife curve. Comparison of DC, SC, IC, EC, LAC, NC, PeC, WDC, UDoNC, SON and FPE based on the validation of jackknife method

Evaluation of the modularity of proteins predicted by FPE, DC and PeC

Proteins often form protein complexes or functional modules to perform their biological functions. Therefore, we try to use protein modularity to assess the essential proteins predicted by FPE. To study the modularity of the proteins, we first choose the top 100 proteins identified by FPE, DC and PeC to construct three small PPI networks. Each small network consists of the top 100 proteins ranked by FPE, DC and PeC. Then, MCODE [43] is used to discover protein modules from the three small PPI networks. The results are shown in Fig. 6.

As shown in Fig. 6, the top 100 proteins ranked by FPE include 88 essential proteins (blue nodes in Fig. 6(a)), whereas DC only identifies 46. MCODE has detected eight modules in the PPI network of FPE, five modules in the PPI network of DC and six modules in the PPI network of PeC. From Fig. 6(b) and Fig. 6(c), there are modules that have not been discovered by MCODE. From the results, we can see that the essential proteins predicted by FPE show more obvious modularity than those identified by DC and PeC.

Effects of the parameter *p*

In this subsection, we discuss the influence of the parameter p that is the switch probability on the prediction results of FPE. With $p = 0$, flowers do not perform the local pollination while with $p = 1$, flowers do not perform the global pollination. For this purpose, we set the switch probability p vary from 0.1 to 0.9. Then the FPE algorithm is ran ten times from $p = 0.1$ to $p = 0.9$, respectively, and we calculate their average. Finally, the number of true essential proteins identified by FPE is shown in Table 3.

According to Table 3, we can see that the differences between the results of $p < 0.6$ and $0.6 \le p < 1.0$ are obvious, when $p < 0.6$, the number of true essential proteins identified by FPE is almost higher than $0.6 \le p < 1.0$, which implies that selecting $p < 0.6$ is a good choice. Moreover, when $p = 0.3$, more superior results can be obtained and further demonstrate that setting the value of switch probability p to be 0.3 is the best choice for predicting essential proteins of FPE. Hence, in this study, we determine the optimal value to be $p = 0.3$.

Table 3 The number of true essential proteins identified by FPE with different *p*

p	Top 100	Top 200	Top 300	Top 400	Top 500	Top 600
0.1	**89**	163	225	287	344	393
0.2	**89**	163	225	289	343	393
0.3	**89**	**164**	226	290	**346**	**395**
0.4	**89**	**164**	225	289	343	390
0.5	**89**	163	227	289	344	391
0.6	88	163	226	288	342	387
0.7	88	163	226	288	342	385
0.8	88	163	227	**291**	343	384
0.9	88	163	**228**	289	343	384

The data in boldface represents the maximum value in each column

Conclusions

The identification of essential proteins is very significant to understand the minimal requirements for cellular life and disease study. In this study, we propose a new algorithm FPE based on the improved flower pollination algorithm to identify essential proteins by integrating gene expression data, subcellular localization and protein complexes information with the topological properties of PPI networks.

To test whether the proposed algorithm is effective, we apply our proposed algorithm FPE on the PPI network of *S. cerevisiae*. First, the comparisons of FPE with ten previous proposed methods DC, SC, IC, EC, LAC,

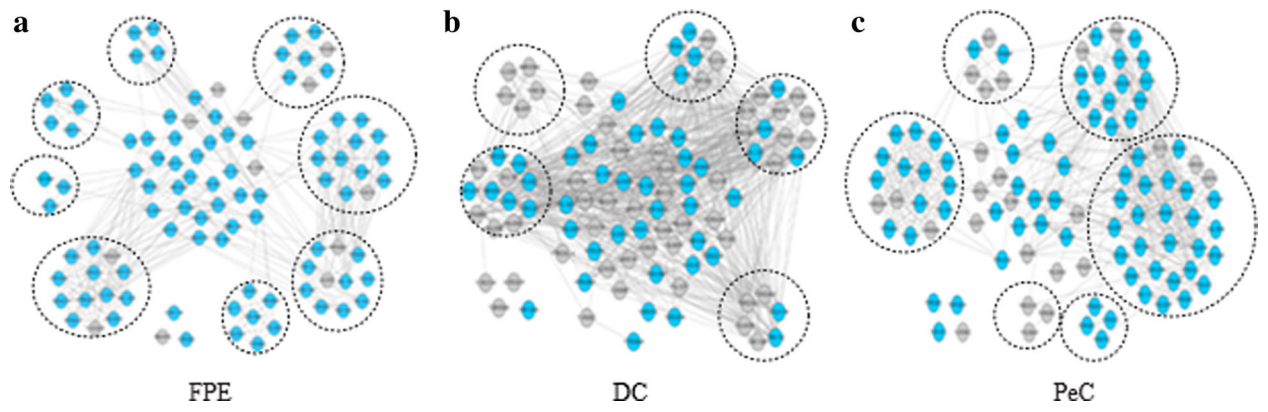

Fig. 6 Evaluation of the modularity of proteins predicted by FPE, DC and PeC. Top 100 proteins in the yeast PPI networks identified by FPE, DC and PeC, respectively. The blue nodes represent the essential proteins and the grey nodes represent the non-essential proteins. In (**a**), MCODE discovers 8 functional modules. (**b**) shows that MCODE identifies 5 functional modules. (**c**) shows that MCODE detects 6 functional modules

NC, PeC, WDC, UDoNC and SON have been made in terms of the number of predicted true essential proteins, as well as the PR curve and the jackknife curve. Then, we further analyze the modularity of proteins and the effect of the switch probability p on the identification results. Both the numerical and the graphical experiment results show that FPE is more competitive than other methods for the identification of essential proteins.

Acknowledgements
This paper is supported by the National Natural Science Foundation of China (61672334, 61502290, 61401263).

Funding
The publication cost of this article was funded by the National Natural Science Foundation of China (61672334).

Availability of data and materials
Yes, however it is currently in the research stage, sharing data may cause infringement, and the integrity of the data needs to be processed subsequently and may be shared later.

About this supplement
This article has been published as part of *BMC Systems Biology* Volume 12 Supplement 4, 2018: Selected papers from the 11th International Conference on Systems Biology (ISB 2017). The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-12-supplement-4.

Authors' contributions
XL and MF conceptualized the algorithm, designed the method and drafted the manuscript, MF performed the experiments and analyzed the data, FXW and LC polished the English expression and finalized the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have declared that no competing interests exist.

Publisher's Note

Author details
[1]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. [2]Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, Canada. [3]Key Laboratory of Systems Biology, CAS center for Excellence in Molecular Cell Science, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

Published: 24 April 2018

References

1. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H. Functional characterization of the S. Cerevisiae genome by gene deletion and parallel analysis. Science. 1999; 285(5429):901–6.
2. Furney SJ, Albà MM, López-Bigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics. 2006;7(1):165.
3. Lu Y, Deng J, Rhodes JC, Lu H, Lu LJ. Predicting essential genes for identifying potential drug targets in aspergillus fumigatus. Comput Biol Chem. 2014;50:29–40.
4. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics. 2009;10(1):290.
5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418(6896):387–91.
6. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature. 2003;421(6920):231–7.
7. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, et al. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50(1):167–81.
8. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 2000; 403(6770):623–7.
9. Gavin A-C, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002; 415(6868):141–7.
10. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki YA. Comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A. 2001;98(8):4569–74.
11. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2.
12. Newman ME. A measure of betweenness centrality based on random walks. Soc Networks. 2005;27(1):39–54.
13. 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.
14. Wuchty S, Stadler PF. Centers of complex networks. J Theor Biol. 2003; 223(1):45–53.
15. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Phys Rev E Stat Nonlinear Soft Matter Phys. 2005;71(5):056103.
16. Bonacich P. Power and centrality: a family of measures. Amer J Sociol. 1987; 92(5):1170–82.
17. Stephenson K, Zelen M. Rethinking centrality: methods and examples. Soc Networks. 1989;11(1):1–37.
18. 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.
19. Li M, Wang J, Chen X, Wang H, Pan Y. A local average connectivity-based method for identifying essential proteins from the network level. Comput Biol Chem. 2011;35(3):143–50.
20. Brohee S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics. 2006;7:488.
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. Li M, Zhang H, Wang J, 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.
23. 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 networks. IEEE/ACM Trans Comput Biol Bioinform. 2015;12(2): 276–88.
24. Zhong J, Wang J, Peng W, Zhang Z, Li M. A feature selection method for prediction essential protein. Tsinghua Sci Technol. 2015;20(5):491–9.
25. Li G, Li M, Wang J, Wu J, Wu F-X, Pan Y. Predicting essential proteins based on subcellular localization, orthology and PPI networks. BMC Bioinformatics. 2016;17(Suppl 8):279.
26. Li M, Lu Y, Niu Z, Wu F-X. United complex centrality for identification of essential proteins from PPI networks. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(2):370–80.
27. Xiao Q, Wang J, Peng X, Wu F-X, Pan Y. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. BMC Genomics. 2015;16

28. Yang XS. Flower pollination algorithm for global optimization. Int Conf on Unconventional Computation and Natural Computation. 2012:240–9.
29. Wang R, Zhou Y, Qiao S, Huang K. Flower pollination algorithm with bee pollinator for cluster analysis. Inf Process Lett. 2016;116(1):1–14.
30. Rodrigues D, Yang XS, Souza And, Papa JP. Binary flower pollination algorithm and its application to feature selection: Springer International Publishing. 2015.
31. Yang XS, Karamanoglu M, He X. Multi-objective flower algorithm for optimization. Procedia Computer Science. 2013;18(1):861–8.
32. Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinform. 2014; 11(2):407–18.
33. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003;100(21):12123–8.
34. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics. 2007;8:236.
35. Luo J, Qi Y. Identification of essential proteins based on a new combination of local interaction density and protein complexes. PLoS One. 2015;10(6): e0131418.
36. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000;28(1):289–91.
37. Tu BP, Kudlicki A, Rowicka M, McKnight SL. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes,. Science (New York NY) 2005; 310(5751): 1152–8.
38. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, Jensen LJ. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford). 2014;2014:bau012.
39. Mewes HW, Frishman D, Mayer KFX, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 2006; 34(Database issue):D169–72.
40. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. SGD: Saccharomyces genome database. Nucleic Acids Res. 1998;26(1):73–9.
41. Zhang R, Ou H-Y, Zhang C-T. DEG: a database of essential genes. Nucleic Acids Res. 2004;32(Database issue):D271–2.
42. Saccharomyces Genome Deletion Project, 1998. [http://sequence.stanford.edu/group/yeast_deletion_project].
43. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics. 2003;4:2.