Journal of Computers Vol. 31 No. 4, 2020, pp. 91-106
doi:10.3966/199115992020083104008

Functional Module Mining in Uncertain PPI Network Based on Fuzzy Spectral Clustering

Yi-min Mao, Yin-ping Liu*

Department of Information Engineering and Computer Science, Jiangxi University of Science and Technology, Ganzhou, Jiangxi, China
2304676691@qq.com

Received 4 January 2019; Revised 31 May 2019; Accepted 15 July 2019

Abstract. Aiming at the lack of accuracy, sensitivity and low time efficiency of protein function module mining methods based on spectral clustering and fuzzy C-means (FCM) clustering, a new algorithm named FSC-FM (functional module mining in uncertain PPI network based on fuzzy spectral clustering) was proposed. In the clustering process, in order to overcome the effect of false positives on the experimental results, the uncertain protein-protein interaction (PPI) network was constructed, in which each protein-protein interaction was assigned with an existence probability by using edge aggregation coefficient. At the same time, FEC (flow distance of edge clustering coefficient) measure was proposed to solve the problem that the spectral clustering is sensitive to the scale parameters in similarity matrix. Furthermore, based on density theory, a probability clustering center strategy was used to design an optimal selection method (DPCS) to improve accuracy and time efficiency of the algorithm. Finally, an improved EED (edge-expected density) metric was studied to filter out the functional modules to improve the precision of the algorithm. We compared our FSC-FM approach on yeast PPI data to the state-of-the-art functional modules prediction algorithms: CDUN, DCU, EA and MGPPA. The experimental results show the superiority of the FSC-FM algorithm in accuracy, sensitivity and time efficiency.

Keywords: FCM, functional module, expected density, uncertain data, protein-protein interaction (PPI), spectral clustering algorithm

1 Introduction

Most cellular processes are performed not by individual proteins, but by functional modules consisting of multiple proteins [1]. Identifying protein functional modules is crucial in understanding the cellular organizations and functional mechanisms; therefore, the mining algorithm for functional modules has become an important issue in academic research [2].

In the PPI network, the functional modules are the proteins in different time and environment, by binding with each other to participate in a particular cellular processes, such as Yeast pheromone response path and so on [3]. Over the past decade, the prediction and discovery of functional modules have been performed by biological experimental procedures. However, these techniques require a large investment of time and resources. Considering these experimental constraints, a variety of computational approaches have been designed which also become a useful supplement to the experimental methods [4]. Due to the advantages of having good accuracy and efficiency in the spectral clustering algorithm, an interesting line of research has focused on clustering of functional modules based on spectral clustering algorithm [5-8]. However, these algorithms are generally hard clustering methods which means do not allow data points belong to more than one cluster, and the experimental results are easily affected by the scale parameters in calculating the similarity matrix. To circumvent the above problems associated with the hard partition of spectral clustering algorithm, combining fuzzy C-means (FCM) algorithm with

* Corresponding Author

spectral clustering, fuzzy spectral clustering algorithm was proposed to mine protein functional modules [9-10]. For the weighting fuzziness parameter in the FCM algorithm, which is used to overcome the hard partition of spectral clustering, while FCM is sensitive to the initial clustering centers, and it is easy to fall into a local optimum in the mining process.

Most functional module mining methods based on determination graphs, whose edges are either present or absent, neighboring information is neglected in these methods. Furthermore, it is known that PPI networks obtained from high-throughput biological experiments have been found to contain false positives as well as false negatives. In other words, the existence of protein-protein interactions is uncertain [11], which presents a challenge for modules discovery from PPI data. To assess the reliability of high-throughput protein interactions, some studies have been proposed to improve the reliability of PPI networks. Using uncertain graph model to deal with such PPI networks is more reasonable than existing graph model. Recently, a great deal of attention has been paid to clustering issues for functional modules based on uncertain grapy model. In particular, Zhang et al. [12] constructed the dynamic uncertain PPI network (DUPN) by integrating gene expression and PPI data, and designed a clustering algorithm to identify protein complexes, named CDUN. Because of the demonstrated significance of the structure in predicting protein complexes, based on the core-attachment concept, a new method called DCU (detecting complex based on uncertain graph model) for predicting complexes from PPI networks was developed by Zhao et al. [13]. In the method, the expected density combined with the relative degree was used to determine whether a subgraph represents a complex with high cohesion and low coupling. To deal with uncertain PPI data, Halim et al. [14] focused on solving the problem of clustering probabilistic graphs using an evolutionary algorithm (EA). In another work, Bano et al. [15] designed a medical gene or protein prediction algorithm (MGPPA) to generate efficient gene or protein clusters over uncertain and noisy data. Even for these methods overcome the influence of false positives on the experimental results, the sensitivity and accuracy of clustering results are poor. Furthermore, these methods are not sufficient to deduce satisfactory conclusions when a large amount of protein interaction data appears.

Though the detection of functional modules in uncertain PPI networks has aroused widespread attention over the past few years, how to design correct and effective functional module detection methods is still a challenging and important scientific problem in computational biology. In this study, we took into account the reliability of PPIs and constructed an uncertain PPI network, in which the reliability of each interaction was represented as a probability. To test the effectiveness of the uncertain PPI network, a novel functional module prediction method named FSC-FM (Functional module mining in uncertain PPI network based on fuzzy spectral clustering) was proposed. The remainder of this paper is organized as follows. In section 2, the FSC-FM algorithmwas proposed. In section 3, the FSC-FM algorithm was described in details. The framework of the proposed algorithm FSC-FM includes: The uncertain PPI network was constructed by edge aggregation coefficient to overcome the effect of false positives. In order to overcome the sensitivity of spectral clustering algorithm to the scaling parameter, FEC measure was designed to calculate similarity matrix between nodes. Furthermore, DPCS (density-based probability center selection) method was proposed to solve the problem that the FCM algorithm is sensitive to the initial cluster center and the cluster number. Finally, the EED (edge-expected density) metric was proposed to filter out the functional modules. Experimental results and discussion were shown and analyzed in Section 4. Section 5 was with the concluding remarks.

2 The Proposed Algorithm

FSC-FM algorithm combines ideas of FCM clustering and spectral clustering in a way that uses the strength of each method and avoids their weakness. Spectral clustering algorithm [16] derives from spectral graph theory, and solves eigenvalue decomposition of matrix to get the low dimensional embedding of data for later clustering. It is not limited to the distribution shape of the original data and can converge to the global optimal solution. FCM [17] is a method of clustering which allows one piece of data to belong to two or more clusters, and fuzzy partitioning is carried out through an iterative optimization of the objective function. Thus, fuzzy spectral clustering has been proposed to mine functional modules [9-10] at present.

Unfortunately, a significant proportion of PPI networks obtained from these high-throughput biological experiments have been found to contain false positives, due to the limitations of the associated

experimental techniques and the dynamic nature of protein interaction graphs, which will have negative effects on the further study of PPI networks. In the fuzzy spectral clustering algorithm, althoughspectral algorithm can deal with arbitrary distribution dataset, it is sensitive to the scaling parameter in calculating the similarity matrix. Furthermore, in the clustering problems based on FCM algorithm for functional modules, clustering results are sensitive to the initial clustering centers and the cluster number. Regarding this situation, in order to improve time efficiency, accuracy, sensitivity and avoid the influence of false positives, we proposed an effective algorithm for mining protein function module named FSC-FM. The framework of the proposed algorithm FSC-FM includes: constructing uncertain PPI network, FEC measure for calculatingsimilaritymatrix, DPCS strategy for selecting clustering centers and EED metric for filtering modules.

3 Research Method

3.1 Constructing Uncertain PPI Network

It is known that the PPI network and other biological data generally bear uncertainties attributed to noise, incompleteness and inaccuracy in practice, and the PPI data contains false positive and false negative rates, which impact the correctness of predicting functional modules. With regard to this problem, in order to improve the prediction accuracy, the PPI network was modeled as an uncertain grapy, in which each protein-protein interaction was endowed with a measure using edge clustering coefficient.

Suppose u and v is two connecting proteins in a graph G, $z(u,v)$ denotes the number of triangles which include the edge (u, v), d_u and d_v are degrees of nodes u and v, respectively. The edge clustering coefficient (ECC) of edge (u,v) is defined as follows [18]:

$$ECC = \frac{z(u,v)}{\min(d_u - 1, d_v - 1)}. \quad \textbf{(1)}$$

Algorithm 1 illustrates the procedure of constructing uncertain PPI network.

Algorithm 1. Constructing uncertain PPI network

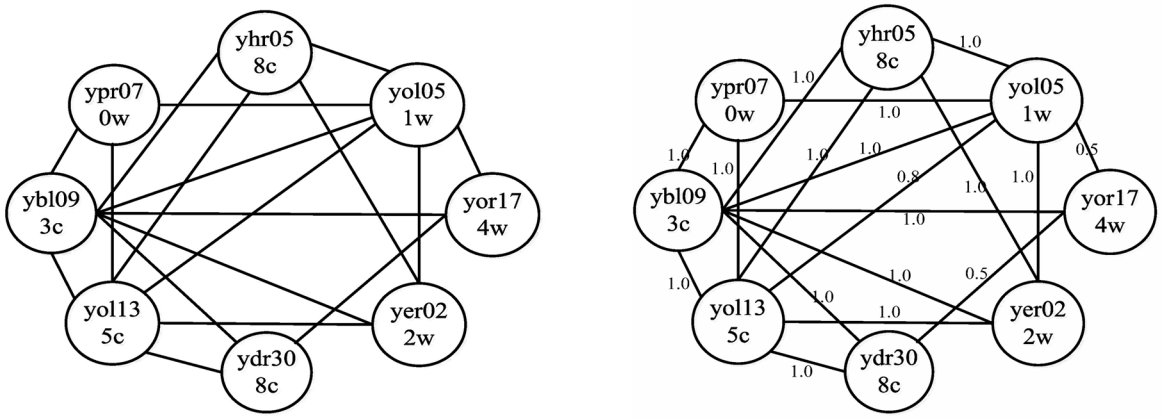
Input: the PPI network $IG=(V, E)$;
Output: the uncertain network $UG(V, E, P)$

1. For each edge $(u,v) \in E$,
   compute its probability value $P(u,v)$ by edge clustering coefficient;
2. Generate UG and the set of possible PPI network $PG = \{g_1, g_2, ..., g_n\}$

Fig. 1 shows an illustration example of the uncertain PPI network construction. In Fig. 1(a), we constructed a static PPI network based on high-throughput PPI data, which contains 8 proteins and 18 interactions. In Fig. 1(b), to construct the uncertain PPI network, we used Eq. 1 to calculate the existence probability of each interaction in the uncertain PPI network.

3.2 Calculating Similarity Matrix

For the spectral clustering algorithm, the traditional Gaussian kernel function is used to measure the similarity between protein nodes. It can only reflect the local consistency characteristics of the cluster structure, and it is sensitive to the scale parameters. Furthermore, the efficiency of the traditional spectral algorithm for detecting modules is low. In order to solve this problem, in the uncertain PPI network, the similarity measure named FEC was proposed by integrating the topological characteristics of high-throughput PPI data and the flow distance.



(a) PPI network (b) uncertain PPI network

Fig. 1. Uncertain network construction change graph

Suppose b_0 and b_s are two vertices in graph G, the vertex sequence $r=(b_0,b_1,...,b_s)$ represents the path connecting the two vertices, where $b_k \in V(0 \le k \le s),(b_k,b_{k+1}) \in E(0 \le k \le s)$. R_{0s} represents the set of paths that may be reached between the vertices b_0 and b_s on G, and the flow distance between two nodes is defined as follows [19]:

$$FD(b_0,b_s)=\frac{1}{\min\limits_{b \in R_{os}} \sum\limits_{k=1}^{|s|-1}\left(e^{\rho d^2(b_k,b_{k+1})}-1\right)+1}. \tag{2}$$

Where $d(b_k,b_{k+1})$ is the Euclidean distance between b_0 and b_s, and the scaling factor $\rho(\rho>1)$ is an adjustable parameter.

Theorem 1 (FEC measure) given an uncertain PPI network $G=(V,E,P)$, where $V=(v_1,v_2,v_3,v_4,v_4,...,v_n)$ is a set of proteins, $E=(e_1,e_2,e_3,e_4,e_5,...e_m)$ is a set of interactions. The similarity between b_i and b_j is defined as follows:

$$FEC(b_i,b_j)=ECC(b_i,b_j)\times FD(b_i,b_j). \tag{3}$$

Where $ECC(b_i,b_j)$ is the existence probability of (b_i,b_j), and $FD(b_i,b_j)$ is the flow distance between two nodes.

Proof:

(1) For $\forall b_i,b_j$, $FEC(b_i,b_j)=FEC(b_j,b_i)$, Symmetry satisfaction;

(2) For $\forall b_i,b_j$, $ECC \ge 0$ and $FD(b_i,b_j) \ge 0$, so $FEC(b_i,b_j) \ge 0$, Positivity satisfaction;

(3) For $\forall b_i,b_j, b_z$, $FEC(b_i,b_j)+\mathrm{FEC}(b_j,b_z) \ge FEC(b_i,b_z)$, Triangle inequality satisfaction.

In this way, Eq. 3 is one distance measure, which satisfies the above-mentioned conditions of metric space.

In this paper, the spectral clustering algorithm with FEC measure was used to preprocess PPI data. The algorithm consists of four steps as shown in Algorithm 2 below.

Algorithm 2. Preprocessing uncertain PPI network.

Input: the uncertain network *UG(V, E, P)*; *N* nodes

Output: Y_{ij}=Q_{ij} / $\left(\sum_j Q_{ij}^2 \right)^{1/2}$

1. For all vertexes in *UG(V, E, P)*, compute similarity matrix W_{ij} by using FEC;
2. Construct Laplace matrix $L = D^{-1/2} W D^{-1/2}$, where $D_{ii} = \sum_{j=1}^{n} W_{ij}$ is defined as a diagonal matrix with diagonal elements;
3. Calculate the eigenvalue vector corresponding to k maximum eigenvalues of L, construct matrix $Q = [q_1, q_2, q, ..., q_k] \in R^{n \times k}$;
4. Normalize the row vectors of Q to get the matrix Y_{ij}=Q_{ij} / $\left(\sum_j Q_{ij}^2 \right)^{1/2}$, take each line of Y as a point in R^k space, and cluster it into k clusters using FCM clustering.

3.3 Selecting Clustering Centers

The weighting fuzziness parameter in the FCM algorithm is used to improve hard partition in spectral clustering. However, clustering results are sensitive to the initial clustering centers and cluster numbers. If the initial center is biased, clustering results generally are not consistent with the actual situation; it is easy to fall into a local optimum in the process of mining function modules. In order to deal with this problem, a probability DPCS strategy was used to design an optimal selection method of initial clustering centers based on density theory. The probability density center is obtained according to the closeness between the protein data. It is used to approximately simulate the initial clustering center of the global data in the FSC-FM algorithm. DPCS method can not only get a better initial clustering center but also avoid falling into a local optimum. The DPCS method can be summarized as follows:

Algorithm 3: Selecting clustering centers

Step 1. Fix the uncertain network *UG* (*V, E, P*), ($\varepsilon > 0$), (m>0) and Y_{ij}=Q_{ij} / $\left(\sum_j Q_{ij}^2 \right)^{1/2}$

Step 2. Calculate the similarity between data nodes by using Eq. 4.

Step 3. Calculate the density function of x_i by using Eq. 5:

$$D_i^{(0)} = \sum_{j=1}^{N} \frac{1}{1 + \frac{4FEC(x_i, x_j)}{r_d^2}} . \tag{4}$$

$$r_d = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} FEC(x_i, x_j)}{N(N-1)}} . \tag{5}$$

Step 4. Get an initial clustering center by using Eq. 6:

$$D_i^{(k)} = D_i^{(k-1)} - D_k^* \frac{1}{1 + \frac{4FEC(x_i, c_k)}{r_d^2}}, \ D_k^* = \max\left\{ D_i^{(k-1)}, i = 1, 2, ..., N \right\} \ k = 1, 2, ..., N . \tag{6}$$

Step 5. Calculate the objective function of FCM algorithm by using Eq. 7:

$$J(u,c)=\sum_{i=1}^{N}\sum_{j=1}^{K}u_{ij}{}^{m}d(x_i,c_j)\,. \tag{7}$$

Step 6. Update the membership u_{ij} and the cluster center c_j by using Eq. 8 and Eq. 9:

$$u_{ij}=\left[\sum_{k=1}^{K}\left(\frac{d\left(x_i,c_j\right)}{d\left(x_i,c_k\right)}\right)^{\frac{1}{m-1}}\right]^{-1}\quad i=1,2,...,N;\ j=1,2,...,K\,. \tag{8}$$

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N}u_{ij}^{m}},\ j=1,2,...,K\,. \tag{9}$$

Where m is the weighting fuzziness parameter, $d\left(x_i,c_j\right)$ is the Euclidean distance between x_i and c_j, the number of clusters is K.

Step 7. Compute $\| J\left(u,c\right)^{(t)}-J\left(u,c\right)^{(t-1)}\|$

If $\| J\left(u,c\right)^{(t)}-J\left(u,c\right)^{(t-1)}\|<\varepsilon$, Stop

Else t=t+1 and return step 6.

In this way, we can get K elements with large density when $D_k^*<\delta D_1^*$, which can approximately be the global optimal initial clustering centers of the FCM algorithm. DPCS method can not only get a better initial clustering center but also avoid falling into a local optimum.

3.4 Filtering Modules

The last stage is redundancy-filter. Although some redundancy may have biological significances, modules overlapped a lot compared to their expected density should be discarded. With quantifying the extent of overlap between each pair of modules by using the overlap score *NA*; module with smaller expected density is discarded.

However, with the increase of the number of interactions in PPI networks, the number of possible PPI networks would grow exponentially. It will cause high computational consumption to get the expected density of protein subgraph from the existing definition. In order to deal with this problem, based on the uncertain PPI network, the expected density optimization EED metric was proposed, which fully considers the neighborhood information of the node and the internal cohesion degree of the PPI network. In this paper, the EED metric was used to filter the function modules. If the value of *EED* is less than the threshold *T*, the module would be filtered out to avoid repeated partition, which can improve the prediction of the algorithm. The following theorem gives a simple formula to compute the expected density.

Theorem 2 (EED metric) suppose the possible $S=\left(V^{"},E^{"},P^{"}\right)$ is a subgraph instantiation of an uncertain graph $G=\left(V,E,P\right)$, where $V^{"}\subseteq V$, $E^{"}\subseteq E$. $P\left(e\right)=ECC\left(e\right)$ denotes the probability of the interaction between nodes, the expected density of S in G can be represented as:

$$EED=\frac{2}{|V^{"}|\times(|V^{"}|-1)}\sum_{e\in E^{"}}ECC\left(e\right). \tag{10}$$

Proof: Assume that the subgraph S consists of M vertices and I edges, the existence probability of edge i is measured by the edge aggregation coefficient, so:

$$EED=\frac{2}{M\times\left(M-1\right)}\sum_{i=1}^{I}ECC_i\,. \tag{11}$$

For subgraph S with 2^n deterministic graphs, which are marked as $S_{fj}\left(f=1,2,3,...n; j=1,2,3,...C_n^f\right)$ in increasing order of the number of edges. In the $S_{fj}\left(f=1,2,...,I; j=1,2,...,C_I^f\right)$, f represents the number of edges, and j represents the number of the determined graphs by the same number of edges. The corresponding density is defined as follows:

$$den\left(S_{fj}\right)=\frac{2f}{M\left(M-1\right)}. \tag{12}$$

In this paper, we assume that the existence probabilities of different interactions in an uncertain PPI network are independent to each other, so:

$$EED=\sum_{f=1}^{I}\sum_{j=1}^{C_I^f} den\left(S_{fj}\right)ECC\left(S_{fj}\right). \tag{13}$$

The Eq. 12 is substituted into the Eq. 13 to obtain:

$$EED=\sum_{f=1}^{I}\sum_{j=1}^{C_I^f}\frac{2f}{M\times\left(M-1\right)}ECC\left(S_{fj}\right)=\frac{2}{M\times\left(M-1\right)}\sum_{f=1}^{I}\sum_{j=1}^{C_I^f} fECC\left(S_{fj}\right)=.$$

$$\frac{2}{M\times\left(M-1\right)}[ECC_1\left(1-ECC_2\right)...\left(1-ECC_n\right)+\left(1-ECC_1\right)ECC_2...\left(1-ECC_I\right)+...+$$

$$\left(1-ECC_1\right)\left(1-ECC_2\right)...ECC_n+2\times ECC_1ECC_2\left(1-ECC_3\right)...\left(1-ECC_I\right)+$$

$$2\times\left(1-ECC_1\right)ECC_2ECC_3...\left(1-ECC_I\right)+...+I\times ECC_1ECC_2...ECC_I$$

The right expansion of Eq. 11 is

$$\frac{2\times\sum_{i=1}^{I}ECC_i}{M\times(M-1)}=\frac{2}{M\times\left(M-1\right)}\left(ECC_1+ECC_2+ECC_3+...ECC_I\right)$$

$$=\frac{2}{M\times\left(M-1\right)}[ECC_1[ECC_2+\left(1-ECC_2\right)][ECC_3+\left(1-ECC_3\right)...[ECC_I+\left(1-ECC_I\right)]+$$

$$[ECC_1+\left(1-ECC_1\right)][ECC_2+\left(1-ECC_2\right)][ECC_3+\left(1-ECC_3\right)]...ECC_I]=$$

$$\frac{2}{M\times\left(M-1\right)}[ECC_1\left(1-ECC_2\right)...\left(1-ECC_I\right)+\left(1-ECC_1\right)ECC_2...\left(1-ECC_I\right)+...+$$

$$\left(1-ECC_1\right)\left(1-ECC_2\right)...ECC_I+2\times ECC_1ECC_2\left(1-ECC_3\right)...\left(1-ECC_I\right)+$$

$$2\times\left(1-ECC_1\right)ECC_2ECC_3...\left(1-ECC_I\right)+...+I\times ECC_1ECC_2...ECC_I]$$

The conclusion is proved.

The exponential calculation complexity of expected density is reduced to the linear level in this paper.

In this paper, the fuzzy spectral clustering algorithm with EED metric was used to filter modules. The algorithm consists of two steps as shown in Algorithm 4 below.

Algorithm 4. Filtering modules

Input: *FM*: the set of protein modules; expected density EED threshold *T*

Output: *PM*: the set of filtered protein modules

1. For each module $A\in FM$
   insert A into *PM*;
2. If $EED\left(A\right)<T$ then remove *A* from *PM*; label *A* with *DISCARDED*.

3.5 Clustering Process of FSC-FM Algorithm

The special steps of FSC-FM algorithm are shown as follows:
Step 1. Calculate the probability between each group of interactions by using Eq. 1, then construct an uncertain PPI network.
Step 2. Calculate the similarity matrix between protein interactions in the PPI network according to Eq. 3, and preprocess the PPI data by using the spectral clustering algorithm with improved similarity measure.
Step 3. Obtain K initial cluster centers through DPCS strategy, update the membershipmatrix and clustering center by using Eq. 8 and Eq. 9, respectively. And calculate the value of the function J by using Eq. 7. If $\| J(u,c)^{(t)} - J(u,c)^{(t-1)} \| < \varepsilon$ then stop, otherwise, iteratively repeat step 3.
Step 4. Calculate the mined module density according to Eq. 10, filter out modules whose EED value is less than the threshold T. This paper sets $T = 0.1$.

3.6 Analysis of FSC-FM Algorithm

The time complexity of FSC-FM algorithm is composed of the following steps. The time complexity of constructing the uncertain PPI network is $O(|E|)$. The time complexity of preprocessing uncertain PPI network using spectral clustering algorithm with FEC measure is dependent on the similarity matrix computation and the eigenvalue decomposition. The time complexity of computing the similarity matrix is $O(N^2)$, and the time complexity of calculating the eigenvalue decomposition is $O(N^3)$, so the overall time complexity of the spectral clustering algorithm is $O(N^3)$. The time complexity of the FCM algorithm using the DPCS strategy to select the initial center depends mainly on calculating the probability density function and the maximum value of the objective function. The time complexity of calculating the probability density function is $O(N^2)$, and the time complexity of searching for the maximum value is $O(N)$, so the overall time complexity of the FCM algorithm is $O(N^2+N)$, as $O(N^2)$. Filtering the protein function modules by using EED metric's complexity is $O(K)$. Therefore, the time complexity of the FSC-FM algorithm is $O(|E|+N^3+N^2+\mathrm{K})$, as $O(N^3)$.

The above symbols and notations are defined as shown in Table 1.

Table 1. Description

symbols and notations	description
u or v	protein
G	graph
$z(u,v)$	the number of triangles
d_u and d_v	degrees of nodes u and v
(u,v)	edge
$P(u,v)$	edge clustering coefficient
$UG(V, E, P)$	the uncertain network
$IG=(V, E)$	the PPI network
$PG=\{g_1, g_2, ..., g_n\}$	the set of possible PPI network
b_0 or b_s	vertices
$r=(b_0, b_1, ..., b_s)$	the vertex sequence
R_{0s}	the set of paths
$d(b_k, b_{k+1})$	Euclidean distance between b_0 and b_s
$ECC(b_i, b_j)$	existence probability of (b_i, b_j)
$FD(b_i, b_j)$	the flow distance

4 Results and Discussion

An experimental computer was configured with the windows 7 ultimate operating system, an Intel i5 dual-core processor, 2.5-GHz frequency and 6.0GB of memory. The algorithm is programmed in python.

4.1 Experimental Data

In order to investigate the performance of our algorithm, the relatively complete and reliable network yeast PPI network is selected as the experimental data. The specific experimental data are shown as follows:

(1) The yeast PPI network data is derived from the DIP database [20], removing self-interactions and repeated ones, which consists of 21554 interactions among 4995 proteins.

(2) To evaluate the protein functional modules predicted by our method, a benchmark set is derived from CYC2008 [21], which consists of 408 functional modules.

4.2 Evaluation Criteria

To assess the quality of the produced functional modules, for any predicted module pc_n and known module bc_m , the overlap score NA is defined as: $NA(pc_n, bc_m) = |pc_n \cap bc_m|^2 / (|pc_n||bc_m|)$. The overlap threshold F is typically set as 0.2 [22], If $NA(pc_i, bc_j) = 1$, they are perfectly matched.

Specificity (Sp) and sensitivity (Sn) are the commonly used measures to evaluate the performance of protein functional module prediction methods. Specificity is the fraction of predicted modules that are true modules while sensitivity is the fraction of benchmark modules that are retrieved.

Given the predicted functional module set $PC = \{pc_1, pc_2, pc_3, ..., pc_n\}$ and the benchmark modules set $BC = \{bc_1, bc_2, ...bc_m\}$.

$$TP = |\{pc_i \in PC \mid \exists bc_j \in BC, NA(pc_i, bc_j) \geq F\}| . \quad \textbf{(14)}$$

$$FP = |\{pc_i \in PC \mid \forall bc_j \in BC, NA(pc_i, bc_j) < F\}| . \quad \textbf{(15)}$$

$$FN = |\{bc_i \in BC \mid \forall pc_j \in PC, NA(pc_i, bc_j) < F\}| . \quad \textbf{(16)}$$

From (14)-(16), TP is the number of correctly predicted modules, FP is the number of incorrectly predicted modules, while TN is the number of predicted benchmark modules and FN is the number of unpredicted benchmark modules.

$$Sp = \frac{TP}{TP + FP}, \ Sn = \frac{TP}{TP + FN} . \quad \textbf{(17)}$$

F-measure is a harmonic mean of specificity and sensitivity, so it can be used to evaluate the overall performance. It is defined as:

$$F - measure = \frac{2 \times Sn \times Sp}{Sn + Sp} . \quad \textbf{(18)}$$

4.3 *P-value* Measure

In PPI network, protein modules can be statistically evaluated using *P-value* from the hypergeometric distribution, which is defined as:

$$P-value=1-\sum_{i=0}^{l-1}\frac{\binom{|X|}{i}\binom{|V|-|X|}{|C|-i}}{\binom{|V|}{|C|}}. \tag{19}$$

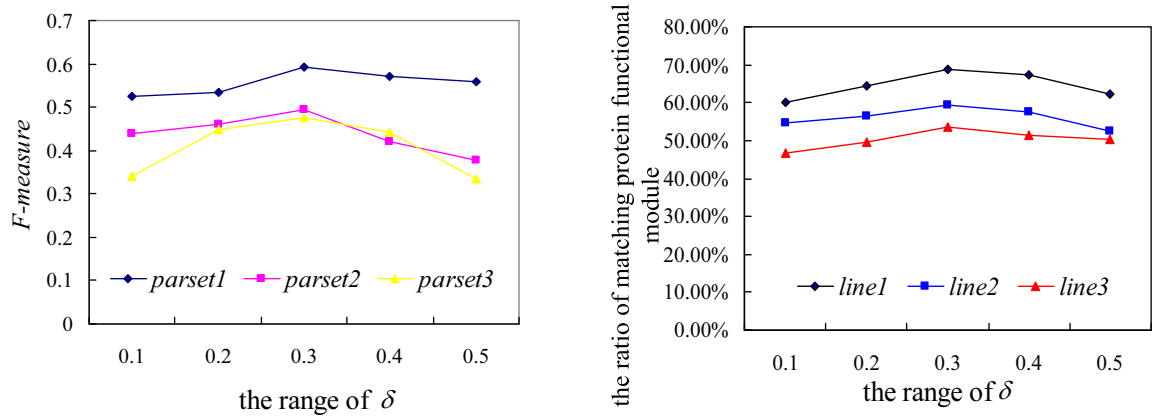
Where $|V|$ is the total number of proteins, X is a known protein module, the size is $|X|$, C is a protein module, $|C|$ denotes the size of the module, and l is the size of the intersection of C and X [23].

4.4 Parameter Selection

In order to evaluate the effect of parameters δ and ε for protein functional modules prediction, 20 experiments were performed on 15 groups of δ and ε. The specific parameter setting is shown in Table 2, where *seti* denotes the *ith* parameter. The average values of *F-measure* and the ratio of matching protein function module of FSC-FM algorithm are as shown in Fig. 2. The parameters in the experiment are set as follows: $m=2$, ρ=3, $T=0.1$.

Table 2. Setting experimental parameters

the range of ε	the range of δ				
	0.1	0.2	0.3	0.4	0.5
0.0015	set1	set2	set3	set4	set5
0.0045	set6	set7	set8	set9	set10
0.0075	set11	set12	set13	set14	set15



(a) *F-measure* value of experimental results (b) Matched protein function module ratio

Fig. 2. *F-measure* value and matched protein function module ratio change graph

From Fig. 2, the *parsetj* in Fig. 2(a) represents the values of *F-measure*, the *linej* in Fig. 2(b) denotes the ratio of matching protein function module under various value of δ, respectively. It can be seen that as δ increases from 0 to 0.3, the value of *F-measure* and the number of the matched functional modules also increase gradually under various value of ε. On the contrary, as δ increases from 0.3 to 0.5, the value of *F-measure* and the number of the matched functional modules decrease gradually under various values of ε. The reason is that if the selection of the initial cluster center is biased by using DPCS strategy, the functional modules that can be matched are more stringent, which result in thevalue of *F-measure* and the matching ratio of the algorithm firstly increase and then decrease with the different values of δ. Multiple experiments show that FSC-FM algorithm achieves the highest *F-measure* of 0.59 and the matching ratio of 68.8347% when ε=0.0015, δ=0.3. Based on these experimental results on the DIP data, it can be seen that our method can achieve high performance for protein module prediction by

setting ε=0.0015 and δ=0.3.

4.5 Effective Analysis of FEC Measure

In order to verify the superiority of the FEC measure, the functional module mining results of FSC-FM algorithm and ADMSC algorithm were compared on the DIP database. The comparative analysis of experimental results is shown in Fig. 3.

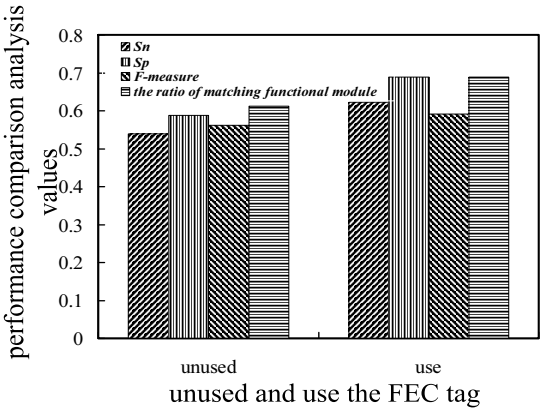


Fig. 3. Comparative analysis of FEC measure

As shown in Fig. 3, FSC-FM algorithm using FEC measure achieves the largest value of *Sn*, *Sp*, *F-measure* and the ratio of matched protein functional modules. The *Sn*, *Sp*, *F-measure* and the number of matched protein modules of FSC-FM algorithm are 15.29%, 17.27%, 5.12% and 12.41% higher than ADMSC algorithm, respectively. The experimental results show that the FSC-FM algorithm with FEC measure achieves an obviously better prediction performance than the ADMSC algorithm. The reason is that the ADMSC algorithm is generally hard clustering method which does not allow data points to belong to more than one cluster at the same time, and the experimental results are easily affected by the scale parameters in calculating the similarity matrix, which result the poor performance. While the FEC measure in FSC-FM algorithm is proposed by integrating the topological characteristics of high-throughput PPI data and the flow distance, which can overcome the sensitivity of spectral clustering algorithm to the scaling parameter. The experimental results show that our approach can effectively deal with the uncertain data in uncertain PPI networks.

4.6 Effective Analysis of DPCS and EED Strategies

In order to further investigate the superiority of the DPCS and EED strategies in FSC-FM algorithm, the functional module mining results of FSC-FM algorithm and the algorithm proposed in [10] were compared on the DIP database. The comparative analysis of experimental results are shown in Fig. 4.

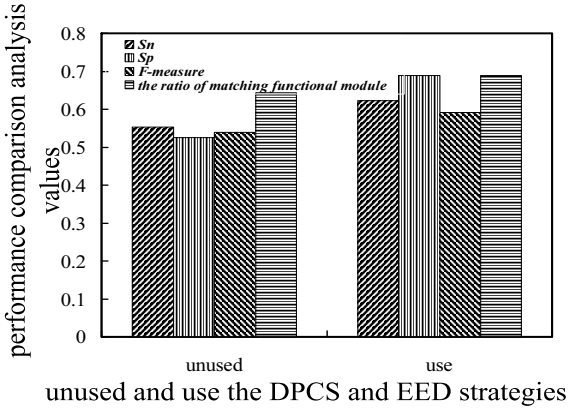


Fig. 4. Comparative analysis of DPCS and EED strategies

As can be seen from Fig. 4, the FSC-FM algorithm using DPCS and EED strategies achieves the largest value of *Sn*, *Sp*, *F-measure* and the ratio of matched protein functional modules, which are 12.5%, 30.86%, 9.63% and 7.05% higher than the algorithm proposed in [10], respectively. The results clearly show that the FSC-FM algorithm using DPCS and EED strategies achieves a better prediction performance than the algorithm proposed in [10]. Because FCM algorithm is proposed for functional modules mining in the literature [10], clustering results are sensitive to the initial clustering centers and clustering numbers, and it is easy to fall into a local optimum in the process of mining function modules, which result the poor performance. Furthermore, there is excessive overlap in the prediction results of the algorithm without DPCS and EED metrics. On the contrary, in FSC-FM algorithm, the DPCS is used to select the initial clustering center, and EED metric is used to filter the mined functional modules, which can improve the performance of the FSC-FM algorithm. Experimental results show that the proposed FSC-FM algorithm is effective in achieving good functional module detection results.

4.7 Performance Comparison

In order to verify the performance of the FSC-FM algorithm, we compared FSC-FM with other functional modules identification methods: CDUN [12], DCU [13], EA [14] and MGPPA [15]. All those algorithms are compared with each other on DIP database using CYC2008 as the benchmark. For all those competitive algorithms, the optimal parameters are set as recommended by their authors. The parameters in this paper are set as follows: $m = 2$, ρ=3, ε=0.0015, δ=0.3, ρ=3, $T = 0.1$.

(1) Comparative analysis of *Sp, Sn* and *F-measure*

In order to verify the performance of the algorithm in this paper, 20 experiments were performed on the DIP data sets. The average results are shown in Table 3 and Fig. 5.

Table 3. Basic information of functional modules by various algorithms

algorithms	*PM*	*Full*	*TP*
CDUN	126	6	54
DCU	254	9	153
EA	378	8	127
MGPPA	354	7	92
FSC-FM	369	18	254

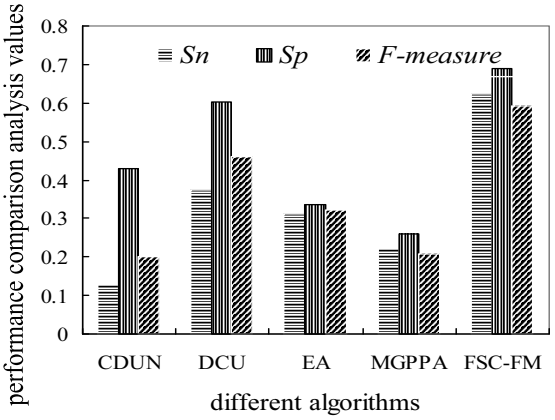


Fig. 5. Performance comparison of different algorithms

As can be seen from Table 3, *PM* is the total number of predicted functional modules, while *Full* is the number of functional modules perfectly matching the known functional modules. From Table 3, it can be seen that the *Full* and *TP* of FSC-FM algorithm achieves the highest *value* of18 and 254, respectively.

The basic performance comparison about predicted functional modules by various algorithms running on DIP data is presented in Fig. 5, including *Sn, Sp* and *F-measure*.

From Fig. 5, it can be seen that FSC-FM algorithm achieves the largest value of *F-measure*, *Sp* and *Sn.*

The *F-measure* of FSC-FM algorithm is 192.37%, 27.92%, 82.98% and 182.23% higher than CDUN, DCU, EA and MGPPA. Experimental results indicate that FSC-FM algorithm performs significantly better than the state-of-the-art methods. The reason is that CDUN and DCU algorithms identified modules from the uncertain PPI networks based on the core-attachment, the protein modules overlapping to a very high extent were predicted. In the clustering process of the EA algorithm, the cluster assignment is initialized randomly, which results that the performance and stability of the algorithm is poor. In MGPPA algorithm, the cluster is predicted only based on gene features, some important topology information in the uncertain PPI network may be lost, which results that the performance of the MGPPA algorithm is low. However, in the entire clustering process of the FSC-FM algorithm, the uncertain PPI network was constructed to overcome the effect of false positives using edge aggregation coefficient. At the same time, the spectral clustering algorithm with FEC measure was designed to preprocess the uncertain PPI network, which can reduce the dimension of the data. Furthermore, In FCM algorithm, the cluster centers are selected by DPCS method and the module was filtered out by EED metric. In return, based on spectral clustering and FCM algorithm, no matter what sample space structure it is, there are fewer iterations, faster convergence and higher clustering accuracy for the FSC-FM algorithm. Therefore, our method achieves the state-of-the-art performance for functional modules identification.

In order to show the clustering effect more clearly, we visualize the detected protein modules. In this paper, we use the software CytoScape to visualize a protein complex and generated FSC-FM (Fig. 6(b)), CDUN (Fig. 6(c)), DCU (Fig. 6(d)), EA (Fig. 6(e)), MGPPA (Fig. 6(f)) and the corresponding cluster in the standard dataset (Fig. 6(a)). In Fig. 6(b), FSC-FM completely identified the standard module perfectly. Fig. 6(c) is the functional module identified by the CDUN algorithm, in which the proteins YJR127C, YKL110C, YDR461W and YJL140W are the wrong proteins. The experimental result of detecting the standard module by DCU algorithm was presented in In Fig. 6(d), where the protein YJL140W is the wrong protein. Fig. 6(e) is the functional module identified by the EA algorithm, where the proteins YDR461W and YJL140W are the wrong proteins. In Fig. 6(f), the protein YJR127C is the wrong protein. From Fig. 6, it indicates that the algorithm FSC-FM is more advantageous in detecting functional modules.

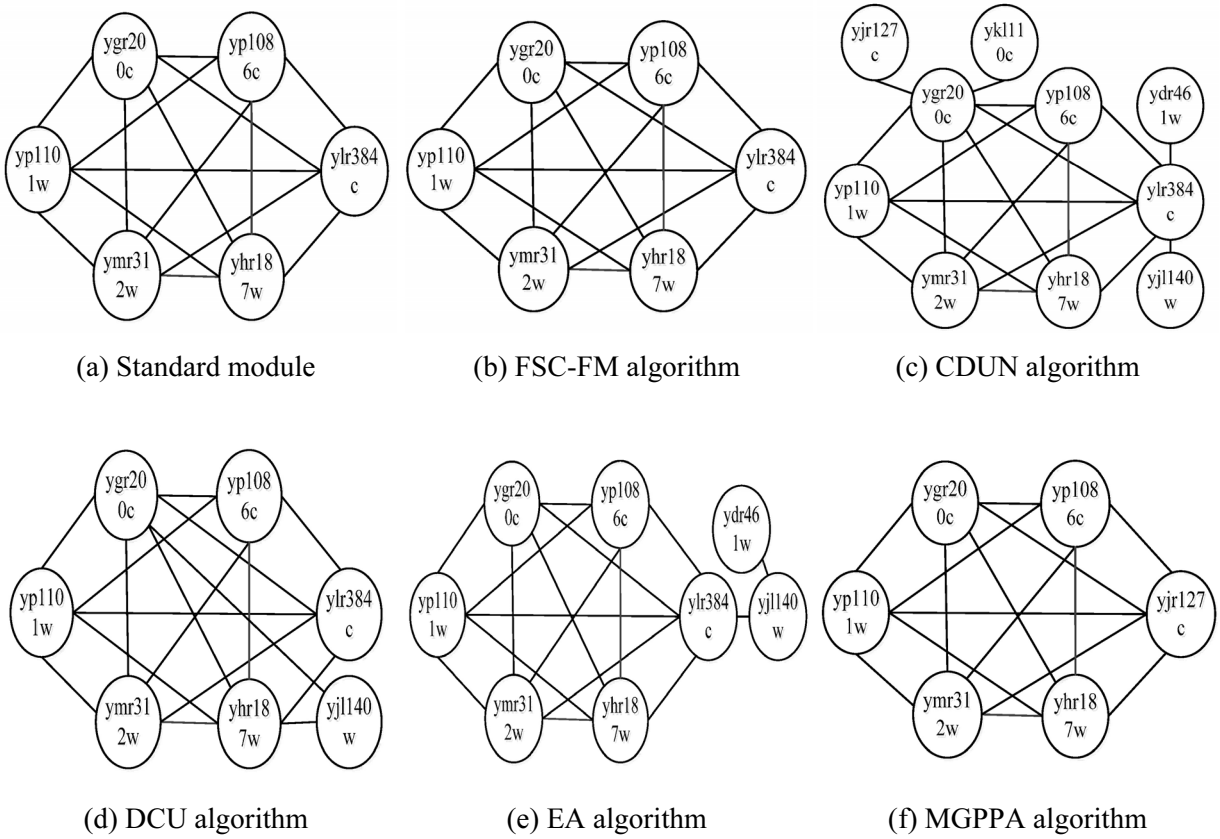


Fig. 6. Visualization comparison of functional modules of each algorithm

(2) Comparative analysis of GO terms

To evaluate the statistical and biological significance of the predicted modules, we also calculate the *P-value* of functional modules predicted by various algorithms.

A low *P-value* of a predicted module indicates that those proteins in the module do not happen merely by chance, so the module has high statistical significance. Generally, a module is considered to be significant with corrected *P-value*<0.01. Research shows that the proportion of significant modules over all predicted ones can be used to evaluate the overall performance of various algorithms. Table 4 lists the comparison results of various algorithms on the DIP data.

Table 4. Statistical significance of predicted functional modules mined by various algorithms

algorithms	*PM*	*SC*	Proportion
CDUN	126	63	50.00%
DCU	254	167	65.75%
EA	378	208	55.03%
MGPPA	354	180	50.85%
FSC-FM	369	307	83.20%

In Table 4, *PM* is the number of predicted modules, and *SC* is the number of significant modules. From Table 4, we can see that our method achieves the largest value of proportion, and the value of proportion of our method is 66.4%, 26.54%, 51.19% and 63.62% higher than CDUN, DCU, EA and MGPPA, respectively. The experimental results show that our FSC-FM approach outperforms these algorithms in terms of statistical significance.

(3) Efficiency analysis

To further assess the quality of the FSC-FM algorithm, Table 5 illustrates the technical comparison of different algorithms, Table 6 illustrates a comparison of the running time of FSC-FM algorithm and the other four methods for predicting modules on DIP data. From Table 5 and Table 6, we can see that FSC-FM algorithm's running time is 508.25s. Experimental results have shown that the FSC-FM algorithm in this paper has lower running time compared to others. The reason is that the spectral clustering algorithm with FEC measure was designed to preprocess the uncertain PPI network, which can reduce the dimension of the data. Furthermore, In the FCM algorithm, DPCS method can not only get a better initial clustering center but also avoid falling into a local optimum and the module was filtered out by EED metric to improve the performance of the algorithm. Therefore, experimental results indicate that our method outperforms other typical methods.

Table 5. Technical comparison

algorithm	technology
CDUN	uncertain weighted network
DCU	uncertain weighted network
EA	uncertain weighted network
MGPPA	uncertain weighted network
FSC-FM	uncertain weighted network, FEC measure, DPCS method, EED metric

Table 6. Efficiency analysis of FSC-FM algorithm and other four algorithms

algorithms	The total number of predicted modules	the size of the modules	matching ratio	running time (s)
CDUN	126	4~38	42.8750%	1485.9490
DCU	254	4~167	60.2362%	650.1639
EA	378	4~117	33.5979%	929.9832
MGPPA	354	4~97	25.9887%	1434.4330
FSC-FM	369	4~295	68.8347%	508.2500

5 Conclusions

In this paper, the uncertain FSC-FM algorithm was obtained by modifying the FCM and spectral clustering algorithm to apply to PPI network. In the clustering process, the uncertain PPI network was constructed by edge aggregation coefficient to overcome the effect of false positives. In order to overcome the sensitivity of spectral clustering algorithm to the scaling parameter, FEC measure was designed to calculate similarity matrix between nodes. Furthermore, DPCS method was proposed to solve the problem that the FCM algorithm is sensitive to the initial cluster center and the cluster number. Finally, the EED metric was proposed to filter out the functional modules. The comparative experiments on DIP datasets show that FSC-FM algorithm outperforms the state-of-th e-art methods for protein modules detection in terms of several criteria such as *specificity*, *sensitivity*, *F-measure*, running efficiency and statistical significance on yeast PPI network. Although the protein functional module mining algorithm based on fuzzy spectral clustering in uncertain PPI network outperforms the existing module mining algorithm, there are still two problems. Firstly, how to cluster with a large amount of uncertain data is a challenging task. Secondly, how to combine multiple biological information to construct dynamic protein network. The above problems still need to study further.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (41562019), the National Key Research and Development Projects of China (2018YFC1504705), Natural Science Foundation of Jiangxi (GJJ161566) and Jiangxi Provincial Technology of Education Department (GJJ181504).

References

[1] J.-Z. Ji, G.-X. Gao, Detecting functional module method based on cultural algorithm in protein-protein interaction networks, Journal of Beijing University of Technology 43(1)(2017) 0013-0021.

[2] G.-M. Liu, B.-F. Chai, K. Yang, Overlapping functional modules detection in PPI network with pair-wise constrained nonnegative matrix tri-factorisation, IET Systems Biology 12(2)(2018) 45-54.

[3] X.-H. Yao, J.-W. Yan, K.-F. Liu, Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules, Bioinformatics 33(20)(2017) 3250-3257.

[4] X.-J. Lei, Y.-L. Ding, F.-X. Wu, Detecting protein complexes from DPINs by density based clustering with pigeon-inspired optimization algorithm, SCIENCE CHINA-Information Sciences 59 (7)(2016) 103-116.

[5] Z.-J. Fan, Z. Luo, Y.-Z. Ma, A spectral clustering algorithm based on fuzzy kernel clustering, Computer Engineering 43(11)(2017) 161-165.

[6] J. Luo, C. Pan, G. Xiang, Identifying functional modules in Co-Regulatory networks through overlapping spectral clustering, IEEE Trans Nano bioscience 17(2)(2018) 134-144.

[7] G.-M. Qin, L. Gao, Spectral clustering for protein complexes in protein-protein interaction (PPI) networks, Mathematical and Computer Modelling 52(11-12)(2010) 2066-2074.

[8] K. Inoue, W.-J. Li, H. Kurata, Diffusion model based spectral clustering for protein-protein interaction networks, Public Library of Science ONE 5(9)(2010) 12623-12632.

[9] D.-E. Na, Exploiting fuzzy spectral clustering in protein-complex detection, Chang Sha: Central South University of Computer Science (2012) 22-34.

[10] K. Trivodaliev, I. Cingovska, S. Kalajdziski, Protein function prediction by spectral clustering of protein interaction network, in: T.-h. Kim, H. Adeli, A. Cuzzocrea, T. Arslan, Y. Zhang, J. Ma, K.-i. Chung, S. Mariyam, X. Song (Eds.),

Database Theory and Application, Bio-Science and Bio-Technology, Springer-Verlag Berlin Heidelberg, 2011, pp. 108-117.

[11] Z.-N. Zou, J.-Z. Li, H. Gao, Mining frequent subgraph patterns from uncertain graph data, IEEE Trans on Knowledge and Data Engineering 22(9)(2010) 1203-1218.

[12] Y.-J Zhang, H.-F. Lin, Z.-H. Yang, An uncertain model-based approach for identifying protein complexes in uncertain protein-protein interaction networks, BMC Genomics 18(7)(2017) 743-752.

[13] B.-H. Zhao, J.-X. Wang, M. Li, Detecting protein complexes based on uncertain graph model, IEEE/ACM Transactions on Computational Biology & Bioinformatics 11(3)(2014) 486-497.

[14] Z. Halim, M. Waqas, S.-F. Hussain, Clustering large probabilistic graphs using multi-population evolutionary algorithm, Information Sciences 317(1)(2015) 78-95.

[15] R. Bano, K. Rao, Graph based gene/protein prediction and clustering over uncertain medical databases, Journal of Theoretical and Applied Information Technology 82(3)(2015) 347-352.

[16] D. Rafailidis, E. Constantinou, Y. Manolopoulos, Landmark selection for spectral clustering based on weighted PageRank, Future Generation Computer Systems 68(3)(2017) 465-472.

[17] O. Kesemen, O Tezel, E. Ozkul, Fuzzy c-means clustering algorithm for directional data (FCM4DD), Expert Systems with Applications 58(C)(2016) 76-82.

[18] X.-J. Lei, X.-Q. Yang, A new method for predicting essential proteins based on participation degree in protein complex and subgraph density, PLOS ONE 13(6)(2018) 0198998-0199016.

[19] L. Wang, L.-F. Bo, L.-C. Jiao, Density-sensitive spectral clustering, ACTA Electronica SINICA 35(8)(2007) 1577-1581.

[20] I. Xenarios, L. Salwinski, X.-J. Duan, DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Research 30(1)(2002) 303-305.

[21] S. Pu, J. Wong, B. Turner, Up-to-date catalogues of yeast protein complexes, Nucleic Acids Research 37(3)(2009) 825-831.

[22] S. Hu, H.-J Xiong, X.-Y. Li, Construction of multi-relation protein networks and its application, ACTA Automatic a SINICA 41(12)(2015) 2155-2163.

[23] W. Liu, L.-Y. Ma, B. Jeon, A network hierarchy-based method for functional module detection in protein-protein interaction networks, Journal of Theoretical Biology 4(455)(2018) 26-38.