Firefly Clustering Method for Mining Protein Complexes

Yuchen Zhang[1], Xiujuan Lei[1(✉)], and Ying Tan[2]

[1] School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
xjlei@snnu.edu.cn
[2] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

Abstract. It is a hot research to explore protein complexes which are closely related to biological processes from the biological network. As a novel swarm intelligence optimization algorithm, the firefly algorithm (FA) has been verified to solve many optimization problems. In this study, we transform the protein clustering problem into an optimization problem in protein-protein interaction (PPI) network. A new method for mining protein complexes based on the firefly algorithm was proposed, called FC. A new objective function was proposed to find the high cohesion and low coupling clusters. A thorough comparison completed for different protein clustering methods has been carried out. The clustering results show that FC method outperforms the other state-of-the-art methods in accuracy of detecting complexes from PPI network.

Keywords: Firefly clustering · Protein complexes · Dynamic PPI network · Clustering objective function

1 Introduction

Proteins play an important role in biological processes. Studying proteins helps us understand genes, disease mechanisms, and so on. However, proteins usually work by interacting with each other. The study of a single protein does not reflect its significance in Biology. The proteins and interactions of proteins form a biological network, protein-protein interaction network (PPI) [1]. At the same time, a group of proteins that works in same space can be used as a protein complex. Protein complexes often have specific functions that can reflect some of the protein properties. So, it is important to explore and research protein complexes. In recent years, a large amount of protein-protein interactions were generated by high-throughput experimental techniques such as yeast two-hybrid and mass spectrometry [2, 3]. These techniques provide a basis for the identification of protein complexes.

Many scholars have proposed a lot of methods to identify the protein complexes. Most of the methods are based on graph theory and dense region discovery. Bader and Hogue proposed the molecular complex detection (MCODE) [4]. MCL [5] was also used to identify protein complexes. There are two main operations, called expansion and inflation. Wang *et al.* [6] used the gene expression data to establish the sequential

Y. Tan et al. (Eds.): ICSI 2017, Part I, LNCS 10385, pp. 601–610, 2017.
DOI: 10.1007/978-3-319-61824-1_65

dynamic PPI network. The clustering result of MCL were optimized. Because of the core-attachment characteristics of protein complexes, CORE [7] and COACH [8] were proposed to predict protein complexes. In identifying overlapping protein complexes, Nepusz *et al.* introduced ClusterONE [9]. In order to make clustering method take into account the biological characteristics of protein complexes, some scholars considered other biological information. Such as CSO [10], it used the gene ontology (GO) annotation data to find complex cliques.

With the development of swarm intelligence algorithm, more and more scholars also begin to apply swarm intelligence algorithm to graph mining. Emad *et al.* try to detect protein complexes by using genetic algorithm [11]. And FA algorithm [12] is also applied to network clustering with significant performance. The algorithm simulates the behaviors of the fireflies that the darker fireflies move to the bright fireflies to solve the optimal solution. A community detecting algorithm [13] was proposed by Amiri *et al.* based on a multi-objective enhanced firefly algorithm. In our previous research, we also used FA to improve the parameters of the MCL algorithm [14].

In the paper, we used the FA algorithm to detect protein complexes from PPI network. To find a corresponding relationship between the behaviors of fireflies and the clustering process. And a new objective function is proposed to transform the PPI clustering into an optimization problem. Finally, in order to verify the performance of the proposed method, we compared it with other clustering methods on different PPI datasets.

2 PPI Network Preprocessing

In most studies, the PPI network is used as an undirected graph $G = (V, E)$, where V is a set of proteins and E represents all interactions. Protein complexes are a set of dense subgraphs with high cohesion and low coupling. Construction of protein network is very important for identifying protein complexes. Wang *et al.* have proved that the dynamic PPI network based on biological characteristics is better than the static PPI network. Therefore, we also used protein gene expression data to construct a dynamic PPI network. Using the 3-sigma method [6] to identify the activity of protein at the different time points.

However, there are a lot of false positives and false negatives in high throughput protein interaction data. So not all interaction relationships are reliable. In order to optimize the network, we weighted the edges of the dynamic network to distinguish their contribution for the task of detecting protein complexes. First, the topology score of edge e_{ij} [15] is defined

$$topology\_score_{ij} = \frac{|N_i \cap N_j| + 1}{max\{avg(G), |N_i|\} + max\{avg(G), |N_j|\}} \tag{1}$$

where N_i and N_j denote the neighbors of v_i and v_j. $|N_i \cap N_j|$ denotes the number of common neighbors of v_i and v_j. And avg(G) is the average degree of the network G.

On the other hand, Gene Ontology annotations are also considered [10]. If there are some common GO annotations between interacting proteins, the interaction is believed to be reliable. They are expressed as follows:

$$GSM_{ij} = \frac{|GSM_i \cap GSM_j|}{|GSM_i| \times |GSM_j|} \tag{2}$$

where the $|GSM_i|$, $|GSM_j|$ represent the number of GO annotations for v_i and v_j, respectively. $|GSM_i \cap GSM_j|$ denotes the number of common GO annotations for both v_i and v_j. Based on the topology score and GO annotations, the weight of an edge e_{ij} is given as:

$$W_{ij} = topology\_score_{ij} \cdot GSM_{ij} \tag{3}$$

The value range of W_{ij} is [0, 1]. If weight of an edge is 0, it will be considered to be false data and deleted from the dynamic network.

3 Firefly Clustering Method

3.1 Firefly Representation

In the process of clustering protein complexes, we first defined the representation of a firefly. A firefly corresponds to a clustering result. In other words, a firefly contains a set of clusters. Obviously, if a firefly is directly represented by a set of clusters, the subsequent steps will be not east to operate. Therefore, the locus-based adjacency representation [16] was used. In the graphic representation, a set of clusters (a firefly) are considered to be a N-dimensional vector. N is the number of nodes in a timestamp network. For a firefly $X = \{x_1, x_2, \ldots, x_N\}$, there is a set of possible range of values based on the adjacency matrix of PPI network. For example in Fig. 1. the node v_3 connect the nodes v_2, v_4, v_5, so the possible range of values $\vec{r}\ v_3$ for v_3 is {2, 4, 5}.

A firefly represents a group of protein complexes. For a firefly, if the value of ith element is j, node v_i and node v_j will be contained in same protein complex. In addition, the existence of independent nodes as clusters in the process of clustering. We added 0 into the value range of each element. So the finally range $\vec{r}\ v_3$ is {0, 2, 4, 5}. If the value of an element is 0 and no other values of elements are equal to its corresponding node, the corresponding node of the element is an independent cluster. After the clustering results are expressed by fireflies, a decoding operation is used to identify all the components in the timestamp network. We can build an adjacency matrix that contained at most N edges according to a firefly. We used the breadth first traversal method to find out connected subgraphs. This representation method does not need to be given a number of clusters in advance. In Fig. 1, a firefly is decoded into two clusters {v_1, v_2, v_3, v_4} and {v_5, v_6, v_7, v_8, v_9}, v_{10} is an independent cluster and is excluded. In the initial stage, we randomly generate m fireflies as initial population according to the range of each element.

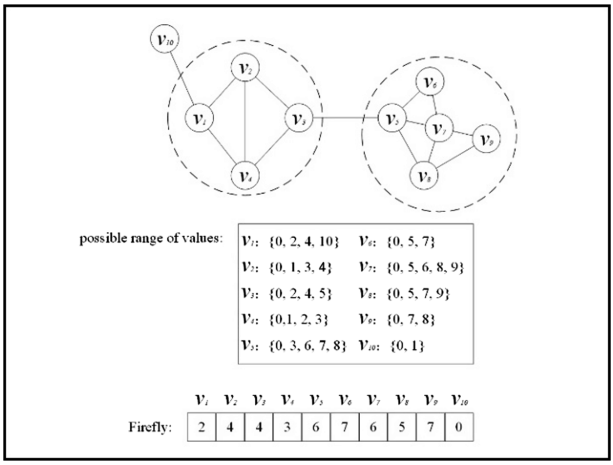

Fig. 1. Correspondence between a firefly and a set of clusters

3.2 Clustering Objective Function

In order to translate the protein clustering problem into an optimization problem, the method needs a reasonable objective function. In FA algorithm, the brightness of firefly is the value of objection function. The objective function need to reflect the properties of protein complexes. Therefore, the objective function should be able to distinguish between high cohesion—low occasional clusters and generic clusters. For objective clusters, there are many edges inside, and the edges between them are less. In this paper, we given the following objective function by combing definition of the density of the cluster and considering the appropriate number of clusters. We also used weight sum of edges to replace the number of edges in clusters.

$$F(\{C^1, C^2, \ldots, C^k\}) = \frac{\sum_{i=1}^{k} \frac{C_{in}^i}{C_{in}^i + C_{out}^i + W_{ave} \times |C^i| \times (|C^i|-1)/2} \times |C^i|}{\sum_{i=1}^{k} |C^i|} \tag{4}$$

$$C_{in}^i = \sum_{p,q \in C^i} W_{pq} \tag{5}$$

$$C_{out}^i = \sum_{p \in C^i, q \notin C^i} W_{pq} \tag{6}$$

where $\{C^1, C^2, \ldots, C^k\}$ represents a set of clusters determined by a firefly. $|C^i|$ represents the number of proteins in a cluster. $\sum_{i=1}^{k} |C^i|$ is the total number of proteins found in the protein complexes detected. W_{ave} is the average weight in the network G. And $|C^i| \times (|C^i|-1)/2$ is the maximum possible number of edges in the cluster C^i. C_{in}^i is the

weight sum of all edges in the cluster C^i. C^i_{out} is the weight sum of edges whose one endpoint is in C^i and another endpoint is not. The goal of proposed method is to find the maximum value of F.

3.3 Firefly Movement Strategy

After generating the initial firefly population, fireflies will find adaptively the optimal solution and generate a set of clusters. However, due to the complexity of the PPI network and dimension of the problem, FC method is easy to fall into the local optimal solution. In order to avoid such problems, we introduced some random search fireflies to improve the method. The FC randomly selected r fireflies from the population, and mutated randomly their some element value with mutation probability mp in each iteration. After mutating, the fitness values of all new fireflies are calculated by clustering objective function F. If the fitness values of new fireflies are greater the fitness values of original fireflies, the new fireflies will replace the original fireflies. The mutation probability mp_i of firefly i is defined as follows:

$$mp_i = \frac{F_{max} - F_i + \alpha}{F_{max}} \tag{7}$$

where F_{i} is fitness value of firefly i. F_{max} is the fitness of the brightest firefly. α is a constant to avoid that the probability is 0.

In each iteration, the fireflies will automatically move to the better solution through the exchange of information between them. If the brightness of a firefly is greater, it will attract the lesser brightness fireflies in the surrounding. In function optimization problem, a firefly moves to all the higher brightness fireflies. However, in the processes of mining protein complexes, the value range of each element is not continuous. So the firefly can only move into one direction. Therefore, we can estimate the probabilities that the firefly will move to the next positions to make the firefly close to the brightest firefly (the optimum solution). Where Firefly$i$_kth is kth element of firefly i. The corresponding probability of movement position is $\vec{p}_{ik}$ in the next generation. We used roulette to determine the direction of movement of fireflies in next generation. For example, in Fig. 2, the brightness of firefly 2, 3, 4 are greater than the brightness of firefly 1. Firefly 1 will move to one of firefly 2, 3, 4. It can be found that is better clustering results, when node v_3 and v_5 is not connected. The number of values in $\vec{r}\ v_3 = \{0, 2, 4, 5\}$ that appear in Firefly2_3th, Firefly3_3th, Firefly4_3th are 0, 1, 2, 0, respectively. FC used that the occurrences number of each value divided by number of the brighter fireflies as occurrences probability of each value in next generation. It can be found that 2 appeared one times, and 4 appeared two times. So the probability of movement position $\vec{p}_{13}$ is $(0, \frac{1}{3}, \frac{2}{3}, 0)$. The firefly will move in the brighter firefly with the probability in each iteration.

For the brightest firefly, it will be randomly perturbed to jump out of the local optimum. The process stops until the algorithm convergence or the maximum number of iterations is reached. Finally, the set of clusters decided by the brightest firefly are the predicted protein complexes. Since the method is run on the dynamic network,

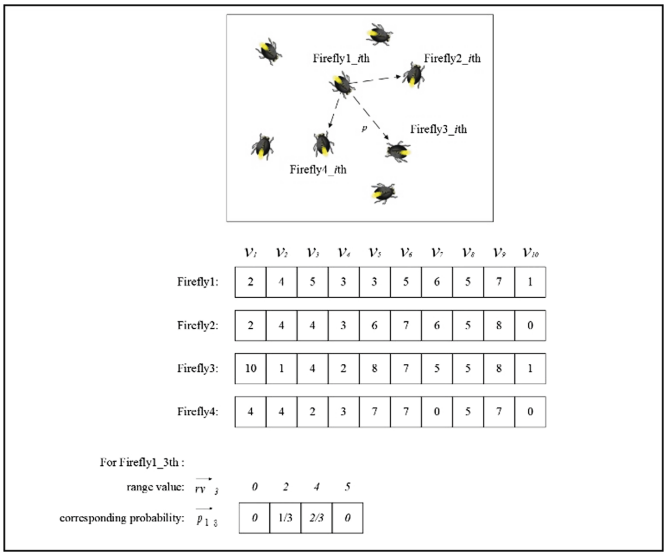

Fig. 2. Firefly movement strategy

there are overlapping protein complexes in results. We removed the protein complexes that are contained by other complexes. Table 1 shows the corresponding relation between firefly biological characteristics and FC method.

Table 1. The corresponding relation of firefly biological characteristics in FC

Firefly characteristics and behavior	Firefly clustering (FC) method
Firefly	A group of protein complexes
Position of firefly (element)	Two proteins in a same complex
Firefly brightness	Clustering objective function value
Movement	Detecting protein complexes
The brightest firefly	Protein complexes (optimal solution)

4 Experiment Results

In order to test the performance of the algorithm, we used three Saccharomyces cerevisiae PPI datasets, DIP [17], Krogan [18], MIPS [19]. And the gene expression data are provided by Gene Expression Omnibus (GEO) [20], accession number of the data is GSE3431. Gene ontology data is the most commonly used data can represent the functions of proteins. In this paper, we used GO-slims data. This data is cut-down version of the GO ontology data [10], which can be acquired at (http://www.yeastgenome.org/download-data/curation). And we used CYC2008 [21] as a known protein complexes set. There are 408 protein complexes.

In the evaluations of predicted protein complexes, the algorithm used several commonly evaluation methods. The Overlapping Score (OS), Sensitivity (Sn), Specificity (Sp), *f-measure* [4] and *p-value* [22] both are used commonly. According to Literature [4], when OS is greater than 0.2, we considered that the predicted protein complexes is matched. If OS is equal to 1, the predicted complexes is perfectly matched. The *p-value* denotes the probability that a predicted protein complex is enriched by a given functional group with random chance. It is generally believed that the smaller the *p-value* (less than 0.01) is, the more significant the predicted protein complex is.

In order to verify the superiority of FC method, we compared with other method such as MCODE [4], MCL [5], CORE [7], CSO [10], ClusterONE [9] and COACH [8] in a same dynamic PPI network. The comparison results are showed in the Table 2. *PC* denotes the total number of protein complexes by predicted. *MPC* represents the

Table 2. Performance comparsion with other methods

Dataset	Method	*Sn*	*Sp*	*f-measure*	*PC*	*MPC*	*MKC*	*Perfect*	*AS*
DIP	MCODE	0.2318	0.6182	0.3372	165	102	70	6	6.7212
	MCL	0.7031	0.2505	0.3694	1541	386	245	14	4.4361
	CORE	0.7381	0.2769	0.4027	1517	420	259	39	2.443
	CSO	0.4403	0.6257	0.5169	342	214	136	11	4.652
	ClusterONE	0.6093	0.3385	0.4352	972	329	197	15	3.5422
	COACH	0.5009	0.5591	0.5284	474	265	144	13	4.9789
	FC-*best*	**0.655**	**0.4612**	**0.5413**	**774**	**357**	**220**	**39**	**3.4406**
	FC-*worst*	**0.6345**	**0.4329**	**0.5147**	**790**	**342**	**211**	**31**	**3.4278**
	FC-*ave*	**0.6359**	**0.4422**	**0.5217**	**778**	**344**	**211**	**36**	**3.4511**
Krogan	MCODE	0.2749	0.7937	0.4084	160	127	73	10	5.125
	MCL	0.566	0.4559	0.5051	658	300	178	40	3.9544
	CORE	0.5417	0.4121	0.4681	677	279	172	39	2.6041
	CSO	0.3284	0.8254	0.4699	189	156	89	10	5.2646
	ClusterONE	0.5232	0.4632	0.4914	585	271	161	28	3.935
	COACH	0.3566	0.81	0.4952	221	179	85	11	5.3575
	FC-*best*	**0.4271**	**0.7537**	**0.5452**	**272**	**205**	**133**	**36**	**3.6765**
	FC-*worst*	**0.4008**	**0.7559**	**0.5239**	**254**	**192**	**121**	**34**	**3.8307**
	FC-*ave*	**0.4131**	**0.7493**	**0.5325**	**265**	**199**	**126**	**39**	**3.7542**
MIPS	MCODE	0.1714	0.5333	0.2595	135	72	60	4	5.437
	MCL	0.5451	0.2017	0.2945	1259	254	196	17	4.7434
	CORE	0.6235	0.249	0.3558	1217	303	225	29	2.5859
	CSO	0.2835	0.5163	0.366	246	127	87	6	4.5528
	ClusterONE	0.4483	0.2796	0.3444	744	208	152	17	3.1317
	COACH	0.3145	0.3662	0.3384	396	145	92	5	6.5253
	FC-*best*	**0.51**	**0.4205**	**0.461**	**604**	**254**	**164**	**32**	**3.2897**
	FC-*worst*	**0.4896**	**0.3865**	**0.432**	**608**	**235**	**163**	**27**	**3.2599**
	FC-*ave*	**0.4989**	**0.4147**	**0.453**	**590**	**245**	**162**	**30**	**3.2989**

number of matched predicted complexes. *MKC* is the number of matched known protein complexes. *AS* denotes average size of predicted protein complexes. Since the FC method has random characteristics, we run the FC method 10 times to analyze the results. FC-*best*, FC-*worst*, and FC-*ave* represent the best result, the worst result, the average result according to *f-measure*, respectively. On DIP data, the *f-measure* of FC-*best* is the highest. And the *f-measure* of FC-*ave* is slightly lower than the value of CSO. And the *f-measure* of FC is the highest on Krogan and MIPS data. In addition, the number of perfect matched complexes is the largest.

Similarly, we compared the proposed method with other methods on function enrichment analysis. We calculated the *p-values* of the protein complexes mined by the algorithms in Biological Process (BP). The result are showed in Table 3. On the Krogan data, percentage of protein complexes whose *p-value* are greater than 0.01 in all complexes identified by FC is the smallest. And on DIP, MIPS data, the *p-value* of FC-*best* are the smallest. And the *p-value* of FC-*worst* are slightly high than the value of CSO and MCODE, respectively. Therefore, in the biological significance terms of predicted proteins complexes, the performance of the proposed method also can be accepted.

Table 3. Function enrichment analysis of predicted protein complexes from different methods

Dataset	Algorithms	PC (size $\geq$ 3)	<E-15	[E-15, E-10)	[E-10, E-5)	[E-5, 0.01)	$\geq$0.01
DIP	MCODE	165	12 (7.27%)	17 (10.30%)	80 (48.48%)	38 (23.03%)	18 (10.91%)
	MCL	1053	19 (1.80%)	47 (4.46%)	183 (17.38%)	362 (34.38%)	442 (41.98%)
	CORE	344	1 (0.29%)	3 (0.87%)	78 (22.67%)	114 (33.14%)	148 (43.02%)
	CSO	342	26 (7.6%)	42 (12.28%)	148 (43.27%)	90 (26.32%)	36 (10.53%)
	ClusterONE	574	21 (3.66%)	52 (9.06%)	177 (30.84%)	184 (32.06%)	140 (24.39%)
	COACH	474	33 (6.96%)	44 (9.28%)	205 (43.25%)	126 (26.58%)	66 (13.92%)
	FC-*best*	**393**	**23 (5.85%)**	**51 (12.98%)**	**179 (45.55%)**	**106 (26.97%)**	**34 (8.65%)**
	FC-*worst*	**404**	**21 (5.20%)**	**49 (12.13%)**	**172 (42.57%)**	**119 (29.46%)**	**43 (10.64%)**
Krogan	MCODE	160	8 (5.00%)	28 (17.50%)	68 (42.50%)	46 (28.75%)	10 (6.25%)
	MCL	403	16 (3.97%)	43 (10.67%)	103 (25.56%)	119 (29.53%)	122 (30.27%)
	CORE	255	3 (1.18%)	10 (3.92%)	60 (23.53%)	102 (40.00%)	80 (31.37%)
	CSO	189	20 (10.58%)	36 (19.05%)	79 (41.80%)	42 (22.22%)	12 (6.35%)
	ClusterONE	399	13 (3.26%)	43 (10.78%)	98 (24.56%)	120 (30.08%)	125 (31.33%)
	COACH	221	23 (10.41%)	37 (16.74%)	91 (41.18%)	54 (24.43%)	16 (7.24%)
	FC-*best*	**157**	**14 (8.92%)**	**27 (17.20%)**	**73 (46.50%)**	**36 (22.93%)**	**7 (4.46%)**
	FC-*worst*	**155**	**14 (9.03%)**	**25 (16.13%)**	**75 (48.39%)**	**34 (21.94%)**	**7 (4.52%)**
MIPS	MCODE	135	5 (3.70%)	10 (7.41%)	70 (51.58%)	39 (28.89%)	11 (8.15%)
	MCL	606	5 (0.83%)	13 (2.15%)	94 (15.51%)	220 (36.30%)	274 (45.21%)
	CORE	340	0 (0.00%)	4 (1.18%)	65 (19.12%)	107 (31.47%)	164 (48.24%)
	CSO	246	7 (2.85%)	27 (10.98)	110 (44.72%)	73 (29.67%)	29 (11.79%)
	ClusterONE	372	7 (1.88%)	16 (4.30%)	117 (31.45%)	126 (33.87%)	106 (28.49%)
	COACH	396	16 (4.04%)	46 (11.62%)	145 (36.62%)	149 (37.63%)	40 (10.10%)
	FC-*best*	**285**	**7 (2.46%)**	**25 (8.77%)**	**127 (44.56%)**	**106 (37.19%)**	**20 (7.02%)**
	FC-*worst*	**290**	**8 (2.76%)**	**25 (8.62%)**	**133 (45.86%)**	**97 (33.45%)**	**27 (9.31%)**

5 Conclusion

In this paper, a novel protein complexes clustering method based on firefly algorithm was proposed. The proposed method has high accuracy in predicting protein complexes. Combined with the FA algorithm, the clustering problem is abstracted as an optimization problem. Because of the adaptability of the algorithm, it is not necessary to set the number of clusters in advance. In the clustering process, through the firefly searching, the algorithm does not need to consider other aspects and its implementation is simple. And the experiment results show that FC method outperforms the other method for mining protein complexes.

Acknowledgments. This paper is supported by the National Natural Science Foundation of China (61672334, 61502290, 61401263), Industrial Research Project of Science and Technology in Shaanxi Province (2015GY016).

References

1. Zhang, A.: Protein Interaction Networks. Cambridge University Press, New York (2009)
2. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.A.: Comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature **403**(6770), 623–627 (2000)
3. Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**(6868), 141–147 (2002)
4. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform. **4**(1), 2 (2003)
5. Dongen, S.M.V.: Graph clustering by flow simulation. Ph.D. thesis. University of Utrecht, The Netherlands (2000)
6. Wang, J., Peng, X., Li, M., Luo, Y., Pan, Y.: Active protein interaction network and its application on protein complex detection. In: IEEE International Conference on Bioinformatics and Biomedicine, Atlanta, pp. 37–42, November 2011
7. Leung, H.C., Xiang, Q., Yiu, S.M., Chin, F.Y.: Predicting protein complexes from PPI data: a core-attachment approach. J. Comput. Biol.: J. Comput. Mol. Cell Biol. **16**(2), 133–144 (2009)
8. Wu, M., Li, M., Kwoh, C.K., Ng, C.K.: A core-attachment based method to detect protein complexes in PPI networks, BMC Bioinform., **10**(1, article 169), 1–16 (2009)
9. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods **9**(5), 471–472 (2012)
10. Zhang, Y., Lin, H., Yang, Z., Wang, J, Li, Y., Xu, B.: Protein complex prediction in large ontology attributed protein-protein interaction networks. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **10**(3), 729–741 (2013)
11. Emad, R., Ahmed, N., Moataz, A.: Protein complexes predictions within protein interaction networks using genetic algorithms. BMC Bioinform. **17**(7), 269 (2016)
12. Yang, X.S.: Firefly Algorithm. Nature-Inspired Metaheuristic Algorithms, vol. 2. Luniver Press, Bristol (2010)
13. Amiri, B., Hossain, L., Crawford, J.W., Wigand, R.T.: Community detection in complex networks: multi–objective enhanced firefly algorithm. Knowl.-Based Syst. **46**, 1–11 (2013)

14. Lei, X., Wang, F., Wu, F.X., Zhang, A., Pedrycz, W.: Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. Inf. Sci. **329**(6), 303–316 (2016)
15. Liu, G., Wong, L., Chua, H.N.: Complex discovery from weighted PPI networks. Bioinformatics **25**(15), 1891–1897 (2009)
16. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Trans. Evol. Comput. **11**, 56–76 (2007)
17. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D.: DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**(1), 303–305 (2002)
18. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Igatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440**(7084), 637–643 (2006)
19. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, M., Mewes, H.W., Stumpflen, V.: MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. **34**(Suppl. 1), 436–441 (2006)
20. Tu, B.P., Kudlicki, A., Rowicka, M., McKnight, S.L.: Logic of the yeast metabolic cycle: temporal compart mentalization of cellular processes. Science **310**, 1152–1158 (2005)
21. Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S.J.: Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. **37**(3), 825–831 (2009)
22. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S.: Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinform. **7**, 207–219 (2006)