Hindawi
Complexity
Volume 2017, Article ID 4120506, 11 pages
https://doi.org/10.1155/2017/4120506

Research Article

Predicting Protein Complexes in Weighted Dynamic PPI Networks Based on ICSC

Jie Zhao,[1] Xiujuan Lei,[1] and Fang-Xiang Wu[2,3]

[1]*School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China*
[2]*School of Mathematical Sciences, Nankai University, Tianjin 300071, China*
[3]*Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*

Correspondence should be addressed to Xiujuan Lei; xjlei@snnu.edu.cn and Fang-Xiang Wu; faw341@mail.usask.ca

Received 31 March 2017; Accepted 2 July 2017; Published 28 August 2017

Academic Editor: Juan A. Almendral

Protein complexes play a critical role in understanding the biological processes and the functions of cellular mechanisms. Most existing protein complex detection algorithms cannot reflect dynamics of protein complexes. In this paper, a novel algorithm named Improved Cuckoo Search Clustering (ICSC) algorithm is proposed to detect protein complexes in weighted dynamic protein-protein interaction (PPI) networks. First, we constructed weighted dynamic PPI networks and detected protein complex cores in each dynamic subnetwork. Then, ICSC algorithm was used to cluster the protein attachments to the cores. The experimental results on both DIP dataset and Krogan dataset demonstrated that ICSC algorithm is more effective in identifying protein complexes than other competing methods.

1. Introduction

Proteins are indispensable to cellular life. Biological functions of cells are carried out by protein complexes rather than single proteins [1]. Detecting these protein complexes can help to predict protein functions and explain biological processes, which has great significance in biology, pathology, and proteomics [2]. Therefore, the study of protein complexes has become one of most important subjects. Many of experimental methods combined with computational strategies have been proposed to predict and identify protein complexes, such as affinity purification and mass spectrometry [3–5]. However, they are costly and have difficulty in capturing the protein complexes instantaneous and dynamic changes [6].

The high throughput techniques have generated a large amount of protein-protein interaction (PPI) data, gene expression data, and protein structure data, which enable scholars to find protein complexes based on the topological properties of PPI networks and structural information of proteins [7]. Bader and Hogue proposed MCODE [8] method to detect protein complexes based on the proteins' connectivity and density in PPI networks. Liu et al. [9] presented a method called CMC to identify protein complexes based on maximal cliques. Protein complexes integrate multiple gene products to perform cellular functions and may have overlapping. Nepusz et al. [10] developed a clustering algorithm ClusterONE to detect overlapping protein complexes. Gavin et al. [11] suggested that there are two types of proteins in complexes: core components and attachments [11]. According to the core-attachment structure of protein complexes, Leung et al. [12] designed CORE algorithm which calculated the p value to detect cores. Wu et al. [13] proposed COACH algorithm to detect dense subgraphs as core components. The biological processes are dynamic and PPIs are changing over time [14]. Therefore, it is necessary to shift the study of protein complexes from static PPI networks to the dynamic characteristics of PPI networks [15]. Wang et al. constructed dynamic PPI network based on time series gene expression data to detect protein complexes [16]. Zhang et al. proposed CSO [17] algorithm by constructing ontology attributed PPI networks based on GO annotation information. Some classical clustering algorithms such as Markov clustering (MCL) [18] and fuzzy clustering [19, 20] were also developed to detect protein complexes.

However, with the birth of the biological simulation technology, bioinspired algorithms provided a new perspective for solving protein complex detection problem [21]. In 2016, Lei et al. proposed F-MCL [22] clustering model based on Markov clustering and firefly algorithm which automatically adjusted the parameters by introducing the firefly algorithm. At the same year, Lei et al. proposed FOCA [6] clustering model which was based on the fruit flies' foraging behavior and protein complexes' core-attachment structure. The previous studies proved that the protein complex detection methods based on the bioinspired algorithms had shown a relatively better performance.

Cuckoo Search (CS) algorithm is a new intelligence optimization algorithm which has been successfully applied to the global optimization problem, clustering, and other fields [21]. In this study, according to the core-attachment structure of protein complexes and CS mechanism, a new clustering method named Improved Cuckoo Search Clustering (ICSC) algorithm was proposed to detect protein complexes in weighted dynamic PPI networks, in which the corresponding relationships between CS algorithm and clustering procedure of PPI data are established.

2. Methods

2.1. Constructing Weighted Dynamic PPI Network. The static PPI networks data produced by high throughput experiments generally contain a high rate of false positive and false negative interactions [9], which makes it inaccurate to predict protein complexes and impossible to reflect the real dynamic changes of PPIs in a cell. To address this problem, some scholars used the computational methods to evaluate the interactions [23]. On the other hand, the protein dynamic information such as gene expression data, subcellular localization data, and transcription regulation data were integrated to reveal the dynamics of PPIs [24–26]. Tang et al. [27] constructed time course PPI network (TC-PIN) by using gene expression data over three successive metabolic cycles. The expression values of genes were compared with a single-threshold to determine whether a gene was expressed. Some essential genes were filtered out by the single-threshold for their low expression levels. Wang et al. [28] developed a three-sigma method to define an active threshold for each gene and then constructed dynamic PPI network (DPIN) by using active proteins based on the static PPI network in combination with gene expression data. Many previous studies have revealed that the three-sigma principle had better prediction performance. In this study, we use three-sigma principle to construct the DPIN. The gene expression data includes three successive metabolic cycles; each cycle has 12 timestamps, so the DIPN includes 12 subnetworks.

A protein p is considered to be active in a dynamic PPI subnetwork only if its gene expression value is greater than or equal to the active threshold Active_Th(p) [28]:

$$\text{Active\_Th}(p) = \mu(p) + 3\sigma(p)(1 - F(p)), \quad (1)$$

where $\mu(p)$ is the algorithmic mean of gene expression values of protein p over timestamps 1 to n and $\sigma(p)$ is the standard deviation of its gene expression values. $F(p)$ is defined as follows:

$$F(p) = \frac{1}{1 + \sigma^2(p)}. \quad (2)$$

A static PPI network is usually described as an undirected graph $G(V, E)$ which consists of a set of nodes V and a set of edges E, the nodes in V represent the proteins and the edges in $E = \{e(v_i, v_j)\}$ represent the connections between pairs of proteins v_i and v_j. $G_t(V_t, E_t)$ is denoted as the dynamic PPI subnetwork at timestamp t ($t = 1, 2, \ldots, n$). Protein v_i interacts with protein v_j in a dynamic PPI subnetwork G_t only if they are active in the same timestamp t and connect with each other in the static PPI network.

As shown in Figure 1, three-sigma principle was applied to calculate the active threshold Active_Th(p) for each protein and to determine the active timestamps. After that, 12 dynamic subnetworks were constructed.

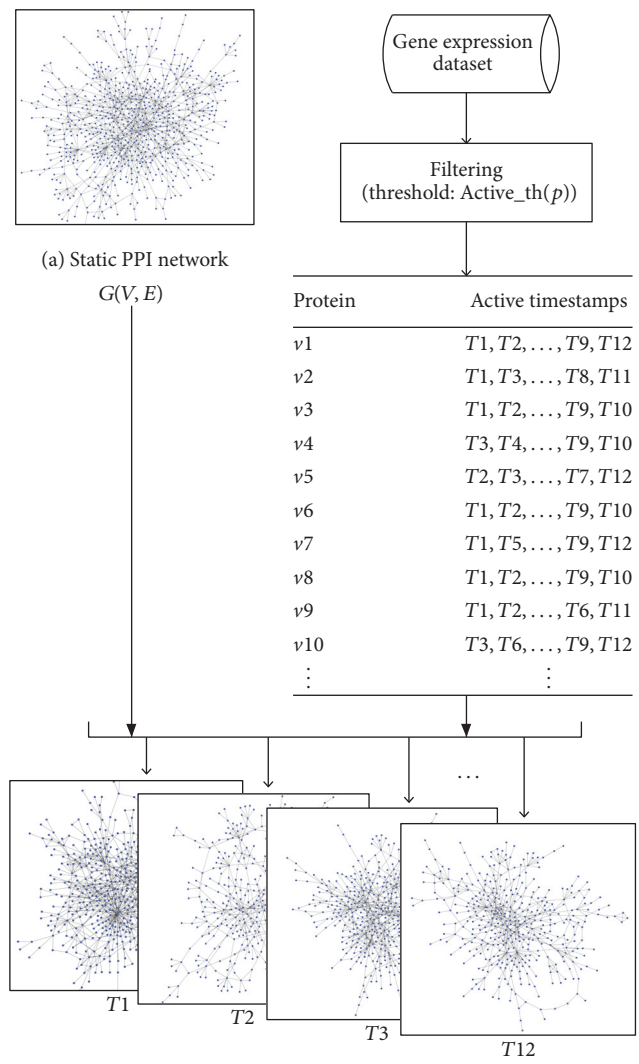


Figure 1: DPIN construction. (a) The static PPI network. (b) The subnetworks contained in the DPIN.

Clustering coefficient has been used as an effective tool to analyze the topology of PPI networks [29]. Radicchi et al. proposed the edge clustering coefficient (ECC) [30]. In PPI network, the ECC of an edge connecting proteins v_i and v_j can be expressed as follows:

$$\mathrm{ECC}_{ij} = \frac{Z_{ij}}{\min\left(\left|N_i\right| - 1, \left|N_j\right| - 1\right)}, \tag{3}$$

where Z_{ij} is the number of triangles built on edge (v_i, v_j); $|N_i|$ and $|N_j|$ are the degrees of protein v_i and v_j, respectively. Edge clustering coefficient is a local variable which characterizes the closeness of two proteins v_i and v_j.

The Pearson correlation coefficient (PCC) was calculated to evaluate how strong two interacting proteins are coexpressed [31]. The PCC value of a pair of genes $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$, which encode the corresponding paired proteins v_i and v_j interacting in the PPI network, is defined as

$$\mathrm{PCC}(x, y) = \frac{\sum_{k=1}^{n} (x_k - \mu(x))(y_k - \mu(y))}{\sqrt{\sum_{k=1}^{n} (x_k - \mu(x))^2} \sqrt{\sum_{k=1}^{n} (y_k - \mu(y))^2}}, \tag{4}$$

where $\mu(x)$ and $\mu(y)$ are the mean gene expression value of proteins v_i and v_j, respectively. The value of PCC ranges from −1 to 1; if $\mathrm{PCC}(x, y)$ is a positive value, there is a positive correlation between proteins v_i and v_j.

The protein complex is a group of proteins which show high coexpression patterns and share high degree of functional similarity, so we integrate GO-slims data from the point of view of protein functions. If two interacted proteins v_i and v_j have some common GO terms, their functions are more similar. Let GSM_{ij} denote this correlation which can be computed as follows:

$$\mathrm{GSM}_{ij} = \frac{\left|\mathrm{GSM}_i \cap \mathrm{GSM}_j\right|^2}{\left|\mathrm{GSM}_i\right| \times \left|\mathrm{GSM}_j\right|}, \tag{5}$$

where $|\mathrm{GSM}_i|$ and $|\mathrm{GSM}_j|$ represent the number of GO terms for proteins v_i and v_j, respectively. In the dynamic PPI subnetwork G_t, the weight between proteins v_i and v_j is defined as follows:

$$W_{ij} = \frac{\mathrm{PCC}_{ij} + \mathrm{ECC}_{ij} + \mathrm{GSM}_{ij}}{3}. \tag{6}$$

Up to now, the weighted dynamic PPI network was constructed.

2.2. Cuckoo Search Algorithm. CS algorithm was a novel bioinspired metaheuristic optimization algorithm proposed in 2009 [32], which was based on the obligatory brood parasitic behaviors of some cuckoo species in combination with the Lévy flight behaviors.

During the breeding period, some certain species of cuckoos lay their eggs in host nests. The cuckoos usually look for host birds which have similar incubation period and brood period. Moreover, their eggs are similar to each other in many aspects of color, shape, size, and cicatricle. The cuckoo flight strategy demonstrates the typical characteristics of Lévy flights. Lévy flights comprise sequences of randomly orientated straight-line movements. Actually, the strategies of frequently occurring but relatively short straight-line movements, as well as randomly alternating with more occasionally occurring longer movements, can maximize the efficiency of resource search [33].

Specifically, for a cuckoo i when generating new solutions $x(t + 1)$, a Lévy flight is performed by using the following equation:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\beta), \quad (i = 1, 2, \ldots, n), \tag{7}$$

where $\alpha > 0$ is the step size which should be related to the scales of the problem of interests. In most cases, we can use $\alpha = 1$; $\oplus$ means the Hadamard product operator. The Lévy flight is a type of random walk which has a power law step length distribution with a heavy tail and the value of β between 1 and 3.

2.3. The ICSC Algorithm. Our ICSC is developed to detect protein complexes in weighted dynamic PPI network through the use of improved CS algorithm. It has been widely accepted that protein complexes are organized in the core-attachment structure.

The core is a small subgraph in a PPI network with high density. As shown in Figure 2(a), four highly connected subgraphs constitute cores, denoted by *core1*, *core2*, *core3*, and *core4* (red round proteins in the dashed circle). Several peripheral connection protein nodes are attachments (blue square proteins) in this PPI network. The blue square proteins and black diamond proteins are all noncore proteins.

In ICSC algorithm, each cuckoo was viewed as a noncore protein (marked with black round in Figure 2(b)), and the nest was viewed as the core proteins (marked with black circles in Figure 2(b)), while the cuckoo population is denoted as a group of clustering results. The noncore proteins become attachments if a cuckoo finds an appropriate nest to lay eggs. Figure 2 illustrates the corresponding relationships between ICSC algorithm and the clustering procedure of a PPI network. Algorithm 1 indicates the function of the proposed algorithm ICSC. The ICSC method operates in three phases. In the first step, some dense subgraphs were selected as initial nests. Then the cuckoos are generated based on these nests. Last the improved Cuckoo Search strategy was applied to generate protein complexes. The complexes in different dynamic subnetworks may have a high level of similarity, so a refinement procedure is applied in order to filter out redundancies and generate the final set of protein complexes.

"Initial nest" subfunction (Algorithm 1) tries to generate initial nests. The initial nests can be seen as the core proteins for each protein complex. The weight of dynamic PPI subnetwork $G_t(V_t, E_t)$ has considered the PCC, ECC, and GSM, so the weight threshold wth can be used to find some protein pairs which have highly functional similarity

Input. The weighted PPI sub-network: $G_t(V_t, E_t)$, $t = 1, 2, \ldots, 12$;
Output. The detected protein complexes: *Complex*
Begin
(1) **for** each G_t **do**
(2) **Initialization**: (1) maximum iterations: maxiter; cuckoo populations' size: *np*;
(3) (2) weight threshold: wth;
(4) (3) Initial nest *nest*: **for** each $e(v_i, v_j) \in E_t$ **do**
(5) **if** $w_{ij} \geq$ (mean(w)/wth) **then** insert (v_i, v_j) into nest **end if**
(6) **end for**
(7) Merge operation;
(8) (4) Initial solutions *Nest*: Nest(:, :, i) = nest, $i = 1, 2, \ldots, np$;
(9) **while** iter ≤ maxiter **do**
(10) **for** $i = 1$ to *np*
(11) Generation cuckoos $Cuckoo_i$: each $v \in V_t$, **if** $v \notin$ Nest(:, :, i) **then** insert v into Cuckoo_i **end if**
(12) **for** each $\text{cuckoo}_j \in \text{Cuckoo}_i$ **do**
(13) **for** each $\text{nest}_k \in$ Nest(:, :, i) **do**
(14) Calculate closeness($\text{cuckoo}_j, \text{nest}_k$);
(15) **if** closeness($\text{cuckoo}_j, \text{nest}_k$) > 0 **then**
(16) Roulette wheel selection $cuckoo_j$, set nest_t = union($\text{nest}_k, \text{cuckoo}_j$);
(17) Calculate objective function $F(\text{nest}_t)$;
(18) **if** $F(\text{nest}_t) > F(\text{nest}_k)$ **then**
(19) insert cuckoo_j into nest_k;
(20) **end if**
(21) **end if**
(22) **end for**
(23) **end for**
(24) Calculate the objective function F(Nest(:, :, i))
(25) **end for**
(26) Find the largest objective function $F\text{max} = \max(F(\text{Nest}(:, :, i)))$, $i = 1, 2, \ldots, np$;
(27) Find the best solution *Nestbest*, F(Nestbest) = Fmax;
(28) **end while**
(29) Complex_t = Nestbest;
(30) **end for**
(31) Complex = $(\text{Complex}_1, \text{Complex}_2, \ldots, \text{Complex}_{12})$
(32) Refinement procedure;
End

Algorithm 1: ICSC algorithm.

and high coexpression. For $e(v_i, v_j) \in E_t$, if the weight w_{ij} is larger than mean(w)/wth, the node pair (v_i, v_j) is denoted as one initial nest, where mean(w) is the average weight of G_t. Protein complex cores often correspond to the small, dense, and reliable subgraphs in PPI networks, but the node pairs may have overlaps with each other. So the node clustering coefficient (NCC) was used to filter out the overlapping nests, which is defined as follows:

$$\text{NCC}(v) = \frac{2n_v}{k_v(k_v - 1)}, \tag{8}$$

where k_v is the degree of node v, n_v is the number of links connecting the k_v neighbors of node v to each other. Because the PPI network has a large number of nodes and edges, many nodes may have the same value of node clustering coefficient. In this study, the weighted node clustering coefficient (WNCC) was defined to distinguish the importance of nodes in the dynamic PPI network. For two initial nests (v_i, v_j) and (v_i, v_k), if $\text{WNCC}(v_i) \geq \text{WNCC}(v_j)$ and $\text{WNCC}(v_i) \geq \text{WNCC}(v_k)$, they are merged into (v_i, v_j, v_k). The WNCC of node v is defined as

$$\text{WNCC}(v) = \frac{\sum \text{We}}{k_v(k_v - 1)}, \quad e \in n_v, \tag{9}$$

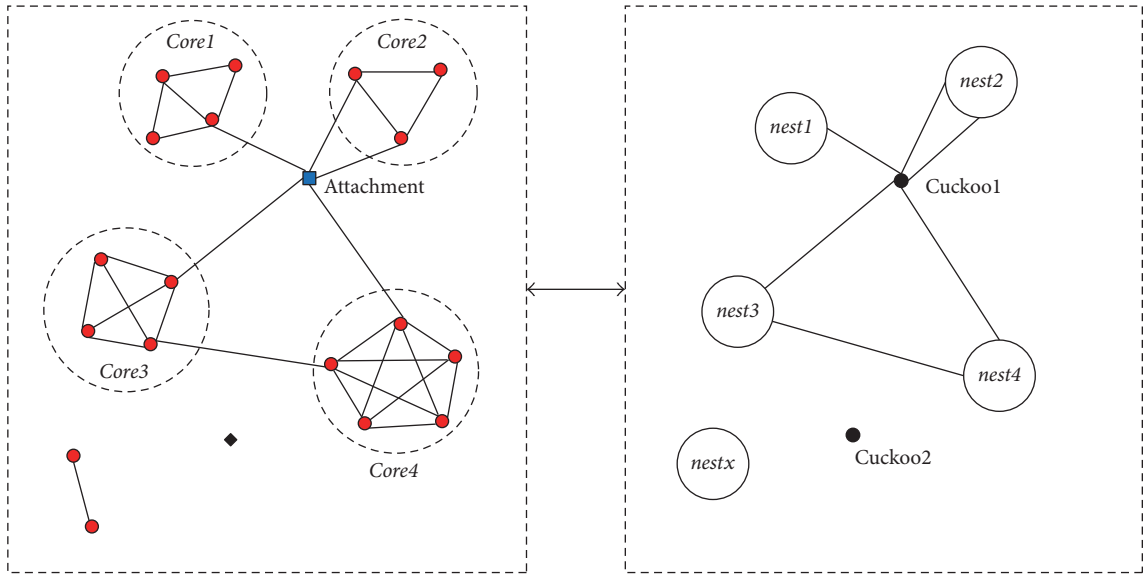
where We is the weight of edge $e \in n_v$; k_v and n_v have the same meanings as in NCC.

After nest detection in the previous steps, the nests are fixed. It is time to find cuckoos around the nests. In $G_t(V_t, E_t)$, if protein $v_i \in V_t$ is not in any nests, it is denoted as a cuckoo.

As a "cuckoo" in $G_t(V_t, E_t)$, there are many "nests" around "cuckoo"; the similarities between "cuckoo" and "nest" is measured based on the closeness between cuckoo_i and nest_j, defined as follows:

$$\text{closeness}\left(\text{cuckoo}_i, \text{nest}_j\right) = \frac{\left|N_{\text{cuckoo}_i} \cap \text{nest}_j\right|}{\left|\text{nest}_j\right|}, \tag{10}$$

where N_{cuckoo_i} is the set of all cuckoo_i's neighbors, $|N_{\text{cuckoo}_i} \cap \text{nest}_j|$ is the number of vertices in nest_j connected with



(a) Core-attachment structure (b) ICSC algorithm

Figure 2: The corresponding relationships between ICSC algorithm and the clustering procedure of a PPI network.

cuckoo_i, and $|\text{nest}_j|$ is the number of vertices in nest_j. In order to keep the diversity of population, the roulette wheel selection was used. For a cuckoo_i, if $\text{closeness}(\text{cuckoo}_i, \text{nest}_j) > 0$, the nest_j is selected to construct the roulette wheel.

The objective function F is defined as follows:

$$F\left(C^1, C^2, \ldots, C^k\right) = \sum_{i=1}^{k} \frac{C_{\text{in}}^i}{C_{\text{in}}^i + C_{\text{out}}^i},$$
$$C_{\text{in}}^i = \frac{2 \times |E|}{|V| \times (|V| - 1)}, \tag{11}$$
$$C_{\text{out}}^i = \frac{W_{ki}}{|V|},$$

where $(C^1, C^2, \ldots, C^k)$ is a clustering result determined by a nest; C^i represents a cluster. $|E|$ is the number of edges in the cluster C^i; $|V|$ is the number of nodes in the cluster C^i. W_{ki} is the number of edges with one node in C^i and another node outside C^i. Finally, the same or highly overlapping protein complexes are filtered out.

2.4. Time Complexity Analysis of ICSC Algorithm. The time complexity is used to estimate the efficiency of the ICSC algorithm. The maximal iterations *maxiter* is for the external loop; each iteration produces np solutions. In order to generate solutions, there are three main operations, generating the cuckoo, calculating the closeness, and calculating the objective function. Let nv be the number of proteins in G_t and ne be the number of interactions in G_t. The time complexity of generating the cuckoos is $O(\text{nv})$. The time complexity of calculating closeness is $O(\text{nc} * \text{nn})$, where nc is the number of cuckoos; nn is the number of nests. The time complexity of calculating the objective function is $O((\text{nv} - \text{nc})^2)$. In summary, the time complexity of ICSC algorithm is $O(\text{maxiter} * \text{np} * (\text{nv} + \text{nc} * \text{nn} + (\text{nv} - \text{nc})^2))$, which is equivalent to $O(\text{maxiter} * \text{np} * \text{nv}^2)$.

3. Experiments and Results

The proposed ICSC algorithm was implemented in Matlab R2015b and executed on a quad-core processor 3.30 GHz PC with 8 G RAM.

3.1. Experimental Dataset. In this study, four PPI datasets DIP [34] (version of 20160114), Krogan et al. [35], MIPS [36], and Gavin et al. [11] were employed to evaluate our algorithm. All the data used were *Saccharomyces cerevisiae* which have false positive and false negative interactions in the datasets. In this study, self-interactions and repetitive interactions are removed for data preprocessing. After preprocessing, the DIP dataset consists of 5028 proteins and 22302 interactions, the Krogan dataset consists of 2674 proteins and 7075 interactions, the MIPS dataset consists of 4546 proteins and 12319 interactions, and the Gavin dataset consists of 1430 proteins and 6531 interactions.

Gene expression data was retrieved from GEO (Gene Expression Omnibus, GSE3431) [37]. After preprocessing, the dataset contains 7074 genes in 3 cell life cycles, each cycle having 12 time points. The GSE3431 dataset contains 4876 proteins in the DIP dataset (coverage rate: 4876/5028 = 96.98%), 2644 proteins in the Krogan dataset (the coverage rate: 2644/2674 = 98.88%), 4446 proteins in the MIPS dataset (the coverage rate: 4446/4546 = 97.80%), and 1418 proteins in the Gavin dataset (the coverage rate: 1418/1430 = 99.16%).

The GO database is currently one of most comprehensive ontology databases in bioinformatics. GO-slims data are

Table 1: The number of proteins and interactions in SPIN and DPIN on four datasets.

Datasets		SPIN	DPIN timestamp t											
			1	2	3	4	5	6	7	8	9	10	11	12
DIP	Protein	5028	860	1029	863	671	645	598	530	1000	1194	638	690	489
	Interactions	22302	1103	1608	1337	839	835	752	627	1861	2447	950	1026	569
Krogan	Protein	2674	336	379	320	256	206	189	202	580	626	304	330	250
	Interactions	7075	334	464	331	234	210	184	213	1025	1081	314	373	258
MIPS	Protein	4546	737	897	781	583	570	531	470	839	1014	523	616	402
	Interactions	12319	1097	1443	1183	754	684	642	504	1238	1637	878	1207	700
Gavin	Protein	1430	177	228	215	135	112	102	96	379	419	174	190	146
	Interactions	6531	242	334	317	150	135	118	135	1019	1043	230	264	184

cut-down version of the GO ontologies [17], which is available at http://www.yeastgenome.org/download-data/curation. GO-slim data provide GO terms to explain gene product feature in biological process (BP), molecular function (MF), and cellular component (CC). we used GO-slims to annotate PPI data.

The standard protein complex CYC2008 [38] is used to evaluate our clustering results, which includes 408 protein complexes and covers 1492 proteins.

In this study, three-sigma principle is used to construct the dynamic PPI networks based on four static PPI networks (SPIN) DIP, Krogan, MIPS, and Gavin in combination with GSE3431 gene expression dataset. There are 12 timestamps per cycle in GSE3431, so each dynamic PPI network contains 12 subnetworks, as shown in Table 1. These 12 subnetworks have different sizes.

3.2. Evaluation Metrics. Three commonly used metrics *sensitivity (SN)*, *specificity (SP), and F-measure* [8, 25, 39] are used to measure the efficiency of the proposed ICSC algorithm and evaluate the performance of the clustering results:

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ F\text{-measure} &= \frac{2\times \mathrm{SN}\times \mathrm{SP}}{\mathrm{SN}+\mathrm{SP}}, \end{aligned} \tag{12}$$

where TP is the number of predicted protein complexes which are matched with 408 standard protein complexes, FP is the number of predicted protein complexes which are not matched with anyone of 408 standard protein complexes, and FN is the number of standard protein complexes which are not matched with predicted protein complexes [8, 25]. The overlapping score OS is used to evaluate the matching degree between predicted protein complexes and standard protein complexes:

$$\mathrm{OS}\left(\mathrm{pc}, \mathrm{sc}\right) = \frac{\left|V_{\mathrm{pc}} \cap V_{\mathrm{sc}}\right|^2}{\left|V_{\mathrm{pc}}\right| \times \left|V_{\mathrm{sc}}\right|}, \tag{13}$$

where V_{pc} and V_{sc} denote the node sets of predicted protein complex pc and standard protein complex sc, respectively. The threshold of OS is set for 0.2 [8, 40]; that is, if OS(pc, sc) is greater than 0.2, the predicted protein complex pc is considered to match standard protein complex sc. OS(pc, sc) = 1 shows that the predicted protein complex pc is perfectly matched with the standard protein complex sc. The p value [41], which illustrates the probability that a protein complex is enriched by a given functional group, was used to evaluate the biological significance of the predicted protein complexes in this study:

$$p\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i}\binom{N-F}{C-i}}{\binom{N}{C}}, \tag{14}$$

where N, C, and F are the sizes of the whole PPI network, a protein complex, and a functional group in the network, respectively, and k is the number of proteins in the functional group in the protein complex [41]. For a protein complex, the smaller the p value is, the higher the biological significance is. The protein complex is considered to be insignificant if p value is greater than 0.01.

3.3. Parameter Analysis. The proposed algorithm ICSC has three parameters, the maximum iterations *maxiter*, the cuckoo populations' size np, and the weight threshold wth. The maximum number of iterations *maxiter* measures the convergence performance of the algorithm, and the populations' size np can guarantee the diversity of the population. The convergence curve of ICSC algorithm on the first subnetwork of the dynamic PPI network was shown in Figure 3. The horizontal axis is the number of iterations, and the vertical axis is the objective function value. Figure 3 illustrates that the ICSC algorithm converges with 30 iterations. The populations' size np is from 5 to 30; the objective function reaches its maximum value at np = 15. In this study, we set maxiter = 100, np = 15.

In ICSC method, cuckoo_i chooses the most suitable nest_j to form a protein complex; the quality of nest_j directly determines the accuracy of protein complexes, and the value of weight threshold wth directly affects the quality of the nest. If the value of wth is too small, a small amount of protein pairs is selected in a nest; the clustering results are not accurate. On the contrary, if the value of wth is too large, lots of meaningless protein complexes are predicted. Therefore, it is critical to select the appropriate value of wth. Matching Rate

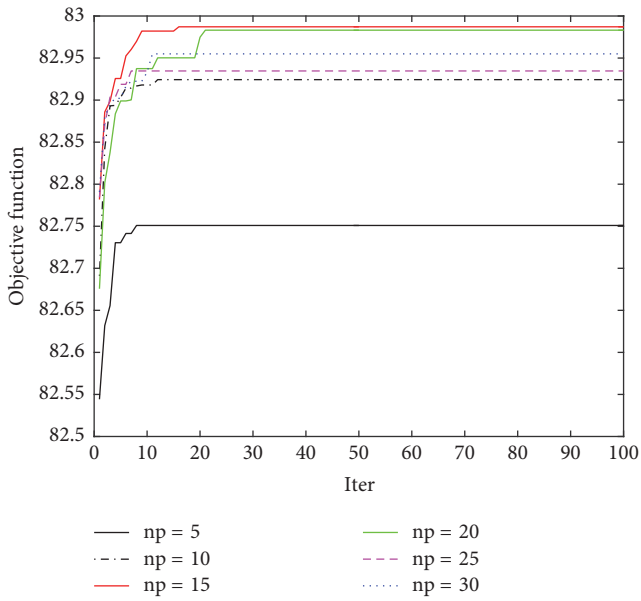


Figure 3: Convergence curve of ICSC algorithm on number 1 subnetwork of DIP dataset.

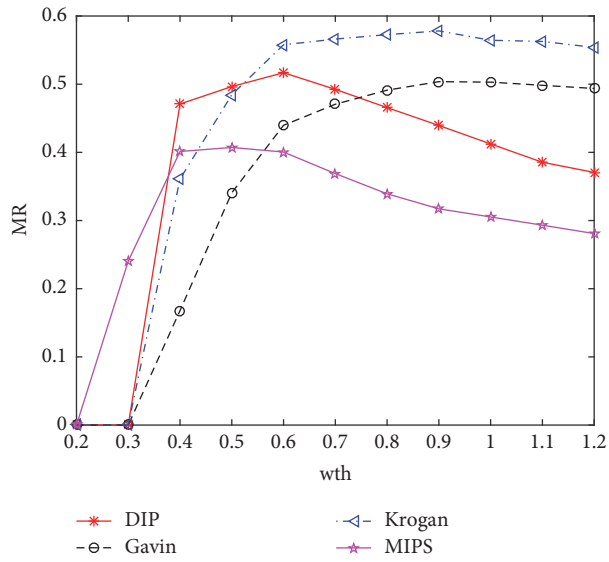


Figure 4: Influence of parameters wth on four DPIN.

(MR) is defined to verify the influence of different values of wth. Nest is the set of initial nests of the dynamic PPI network; SC is the set of standard protein complexes CYC2008, and MR(Nest, SC) is defined as follows:

$$\mathrm{MR}\,(\mathrm{Nest}, \mathrm{SC}) = \frac{(\mathrm{NI}/\,|\mathrm{Nest}| + \mathrm{SI}/\,|\mathrm{SC}|)}{2}, \tag{15}$$

where NI is the number of nests which are included in the standard protein complexes, |Nest| denotes the number of nests in Nest, SI is the number of standard protein complexes which are included in Nest, and |SC| denotes the number of protein complexes in SC. The experiments on four dynamic PPI networks with wth from 0.2 to 1.2 were carried out to verify the influence of parameters wth. The results were showed in Figure 4. From Figure 4, in Krogan and Gavin datasets, the MR tends to be stable while wth is greater than or equal to 0.8. In DIP datasets the MR reaches its maximum value at wth = 0.6 and then gradually declines, and the downward trend is from 0.6 to 0.8. The MR curve in MIPS dataset is similar to DIP. Therefore, the value of wth is set as 0.8 in this study.

3.4. Clustering Results. The performance of ICSC is compared with six other previously proposed methods: MCODE, MCL, CORE, CSO, ClusterONE, and COACH. All the six methods were run on the dynamic PPI networks constructed by three-sigma principle based on DIP, Krogan, MIPS, and Gavin datasets. The clustering results are shown in Table 2, where PC is the total number of predicted protein complexes, MPC is the count of predicted protein complexes which were matched, and MSC is the number of matched standard protein complexes. Perfect is the count of predicted protein complexes and standard complexes are perfectly matched; that is, OS(pc, sc) = 1. AS represents the average size of the predicted protein complexes. The comparison results are also showed in Table 2, from which it is clear that ICSC performs better than other six methods in terms of *sensitivity (SN)* and MPC. The *F*-measure of ICSC is the highest on DIP, Krogan, and MIPS while on the Gavin the *F*-measure of ICSC it was a bit less than that of ClusterONE. The Perfect values of ICSC on DIP and MIPS are 64 and 50, respectively, and are far superior to other algorithms.

In Table 2, the *perfect* value of ICSC on DIP is 64. The degree distribution of perfectly matched protein complexes is calculated in Table 3. The *degree* refers to the number of protein nodes contained in the protein complex. There are 408 protein complexes in the standard protein complexes CYC2008; 172 complexes contain 2 protein nodes accounting for 42.16%. However, the MCODE, CSO, and COACH cannot predict this part of protein complexes. The degree of 149 protein complexes greater than or equal to 4 accounted for 36.52% of all standard protein complexes, only a small part of which can be predicted by MCL, CORE, and ClusterONE. It is clear that ICSC algorithm achieved the best performance in these two aspects.

In order to clearly show the clustering results, we visualize the 265th standard protein complex of CYC2008 "nuclear exosome complex" in Figure 5. As shown in Figure 5(a), there are 12 proteins in this standard protein complex. The clustering results of other five methods MCODE (b), MCL (c), CORE (d), ClusterONE (e), and ICSC (f) are all from Krogan dataset. The blue nodes are proteins that are correctly predicted, the red nodes are proteins that are not identified, and the green nodes are the proteins that are wrongly identified. MCODE method only successfully predicted six proteins. Although MCL successfully predicted all 12 proteins in the protein complex, MCL also produced 3 incorrect proteins. The accuracy of CORE is the lowest; only 2 proteins are successfully predicted. Our method ICSC accurately

Table 2: The performance comparison of several typical algorithms on four datasets.

Dataset	Algorithms	SN	SP	*F*-measure	PC	MPC	MSC	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
	ICSC	*0.8385*	*0.4186*	*0.5585*	*1997*	*836*	*247*	*64*	*3.5613*
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
	ICSC	*0.6314*	*0.5966*	*0.6135*	*761*	*454*	*143*	*23*	*3.3338*
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
	ICSC	*0.7181*	*0.3028*	*0.4260*	*1691*	*512*	*207*	*50*	*3.7534*
Gavin	MCODE	0.2612	0.7548	0.3881	155	117	77	6	5.3484
	MCL	0.4411	0.6417	0.5228	321	206	147	25	5.0312
	CORE	0.4336	0.5735	0.4938	347	199	148	26	2.8184
	CSO	0.3109	0.773	0.4434	185	143	91	6	5.9405
	ClusterONE	0.4797	0.6413	0.5488	368	236	152	19	5.2826
	COACH	0.3477	0.6966	0.4585	234	163	94	5	6.312
	ICSC	*0.5033*	*0.5704*	*0.5347*	*540*	*308*	*104*	*10*	*3.6093*

Table 3: The degree distribution of predicted protein complexes perfectly matched (OS = 1) on DIP datasets.

Algorithm	Prefect	Degree (=2)	Degree (≥4)
MCODE	6	0	3
MCL	14	10	1
CORE	39	32	1
CSO	11	0	5
ClusterONE	15	11	2
COACH	13	0	6
ICSC	64	**47**	**6**
CYC2008	408	172 (42.16%)	149 (36.52%)

predicted 9 proteins and achieved the best performance in identifying protein complexes.

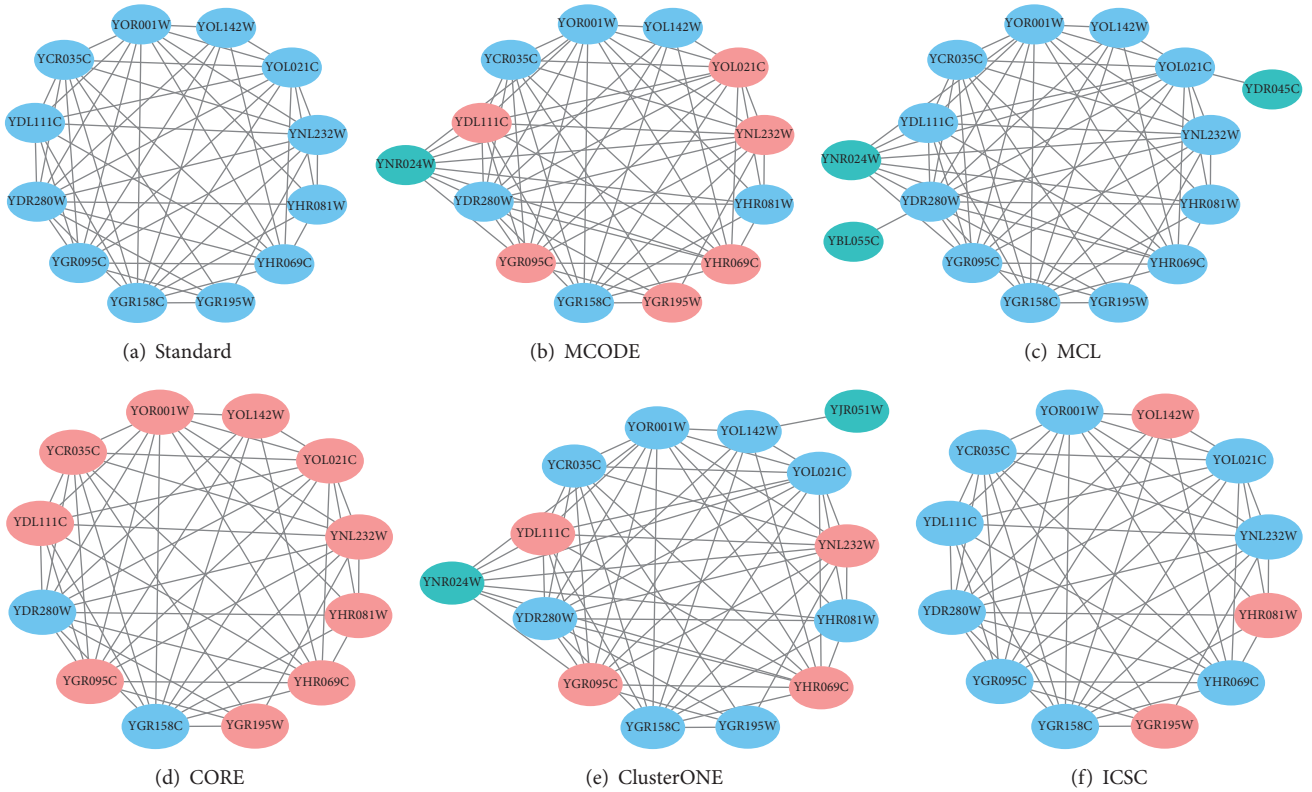
To evaluate the biological significance and functional enrichment of protein complexes identified by ICSI, we randomly selected five predicted protein complexes and calculated the p value of on biological process ontologies based on Krogan datasets by using GO: termFinder (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl). The results are showed in Table 4. The proteins in bold have well matched standard protein complexes. From Table 4, it is obvious that four protein complexes have larger OS values and lower p values, which illustrates that the ICSC algorithm is effective, and these protein complexes are reliable and biologically meaningful.

4. Conclusion

Protein complexes are involved in multiple biological processes, and thus detection of protein complexes is essential to understanding cellular mechanisms. There are many methods to identify protein complexes but cannot reflect dynamics of protein complexes. In this study, we have presented a novel protein complex identification method ISCS according to the core-attachment structure of protein complexes. First, a weighted dynamic PPI network is constructed, which integrates the gene expression data and GO terms information. Then, we find functional cores and cluster protein

Table 4: Function enrichment analysis of five predicted protein complexes detected form Krogan dataset.

Number	Predicted protein complex	p value	Gene Ontology	OS
(1)	**YOR001W YOL021C YNL232W YHR069C YGR158C YGR095C YDR280W YDL111C YCR035C**	4.25e − 24	GO:0071042 Cluster frequency: 100%	0.75
(2)	**YNL136W** YML041C **YJL081C YHR090C YGR002C** YDR334W	5.98e − 07	GO:0006325 Cluster frequency: 100%	0.23
(3)	**YNL317W YLR277C YKL059C YKL018W YJR093C YGR156W YDR301W YDR195W** YBL010C	2.45e − 16	GO:0031124 Cluster frequency: 88.9%	0.41
(4)	**YPR034W YMR091C YMR033W YLR357W YLR321C YLR033W YKR008W YFR037C YDR303C YCR052W YCR020W-B**	1.35e − 28	GO:0031498 Cluster frequency: 100.0%	0.64
(5)	**YPR110C YOR340C YOR210W YNL248C YJR063W YJL148W YDR156W YBR154C**	4.81e − 16	GO:0098781 Cluster frequency: 100.0%	0.57



(a) Standard (b) MCODE (c) MCL

(d) CORE (e) ClusterONE (f) ICSC

Figure 5: Visualization of the 265th standard protein complex "nuclear exosome complex."

attachments based on the CS algorithm. Compared with the other competing clustering methods, ICSC can effectively identify the protein complexes and has higher precision and accuracy.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

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

References

[1] E. A. Winzeler, D. D. Shoemaker, A. Astromoff et al., "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, 1999.

[2] A. Lakizadeh, S. Jalili, and S.-A. Marashi, "PCD-GED: Protein complex detection considering PPI dynamics based on time series gene expression data," *Journal of Theoretical Biology*, vol. 378, pp. 31–38, 2015.

[3] A. J. Link, J. Eng, D. M. Schieltz et al., "Direct analysis of protein complexes using mass spectrometry," *Nature Biotechnology*, vol. 17, no. 7, pp. 676–682, 1999.

[4] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] A. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[6] X. Lei, Y. Ding, H. Fujita, and A. Zhang, "Identification of dynamic protein complexes based on fruit fly optimization algorithm," *Knowledge-Based Systems*, vol. 105, pp. 270–277, 2016.

[7] M. Bertolaso, A. Giuliani, and L. De Gara, "Systems biology reveals biology of systems," *Complexity*, vol. 16, no. 6, pp. 10–16, 2011.

[8] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[9] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.

[10] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471-472, 2012.

[11] A.-C. Gavin, P. Aloy, P. Grandi et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[12] H. C. Leung, Q. Xiang, S. M. Yiu, and F. Y. Chin, "Predicting protein complexes from PPI data: a core-attachment approach," *Journal of Computational Biology*, vol. 16, no. 2, pp. 133–144, 2009.

[13] M. Wu, X. Li, C.-K. Kwoh, and S.-K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics*, vol. 10, article 169, 2009.

[14] R. V. Solé, R. Ferrer-Cancho, J. M. Montoya, and S. Valverde, "Selection, tinkering, and emergence in complex networks. Crossing the land of tinkering," *Complexity*, vol. 8, no. 1, pp. 20–33, 2002.

[15] B. Chen, W. Fan, J. Liu, and F. X. Wu, "Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 177–179, 2014.

[16] J. Wang, X. Peng, M. Li, Y. Luo, and Y. Pan, "Active protein interaction network and its application on protein complex detection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp. 37–42, 2011.

[17] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, and B. Xu, "Protein complex prediction in large ontology attributed protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 3, pp. 729–741, 2013.

[18] S. M. Van Dongen, Graph clustering by flow simulation, 2001.

[19] H. Wu, L. Gao, J. Dong, and X. Yang, "Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks," *PLoS ONE*, vol. 9, no. 3, Article ID e91856, 2014.

[20] P. Manikandan, D. Ramyachitra, and D. Banupriya, "Detection of overlapping protein complexes in gene expression, phenotype and pathways of Saccharomyces cerevisiae using Prorank based Fuzzy algorithm," *Gene*, vol. 580, no. 2, pp. 144–158, 2016.

[21] J. Zhao, X. Lei, and F. Wu, "Identifying protein complexes in dynamic protein-protein interaction networks based on Cuckoo Search algorithm," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1288–1295, Shenzhen, China, December 2016.

[22] X. Lei, F. Wang, F.-X. Wu, A. Zhang, and W. Pedrycz, "Protein complex identification through Markov clustering with firefly algorithm on dynamic protein-protein interaction networks," *Information Sciences*, vol. 329, pp. 303–316, 2016.

[23] J. Zeng, D. Li, Y. Wu, Q. Zou, and X. Liu, "An empirical study of features fusion techniques for protein-protein interaction prediction," *Current Bioinformatics*, vol. 11, no. 1, pp. 4–12, 2016.

[24] F. Luo, J. Liu, and J. Li, "Discovering conditional co-regulated protein complexes by integrating diverse data sources," *BMC Systems Biology*, vol. 4, no. 2, article no. 4, 2010.

[25] M. Li, X. Wu, J. Wang, and Y. Pan, "Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data," *BMC Bioinformatics*, vol. 13, no. 1, article 109, 2012.

[26] C. Tu, "A dynamical method to estimate gene regulatory networks using time-series data," *Complexity*, vol. 21, no. 2, pp. 134–144, 2015.

[27] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, "A comparison of the functional modules identified from time course and static PPI network data," *BMC Bioinformatics*, vol. 12, article no. 339, 2011.

[28] J. Wang, X. Peng, M. Li, and Y. Pan, "Construction and application of dynamic protein interaction network based on time course gene expression data," *Proteomics*, vol. 13, no. 2, pp. 301–312, 2013.

[29] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, article 519, 2006.

[30] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[31] X. Shang, Y. Wang, and B. Chen, "Identifying essential proteins based on dynamic protein-protein interaction networks and RNA-Seq datasets," *Science China Information Sciences*, vol. 59, no. 7, Article ID 070106, 2016.

[32] X.-S. Yang, *Nature-inspired metaheuristic algorithms*, Luniver Press, 2010.

[33] A. M. Reynolds and C. J. Rhodes, "The Lévy flight paradigm: Random search patterns and mechanisms," *Ecology*, vol. 90, no. 4, pp. 877–887, 2009.

[34] I. Xenarios, Ł. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[35] N. J. Krogan, G. Cagney, H. Yu et al., "Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[36] U. Güldener, M. Münsterkötter, M. Oesterheld et al., "MPact: the MIPS protein interaction resource on yeast.," *Nucleic acids research.*, vol. 34, pp. D436–441, 2006.

[37] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, "Cell biology: Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

[38] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic Acids Research*, vol. 37, no. 3, pp. 825–831, 2009.

[39] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[40] X. Lei, Y. Ding, and F.-X. Wu, "Detecting protein complexes from DPINs by density based clustering with Pigeon-Inspired Optimization Algorithm," *Science China Information Sciences*, vol. 59, no. 7, Article ID 070103, 2016.

[41] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics*, vol. 7, article 207, 2006.