Identifying Protein Complexes Based on Local Fitness Method

Jun Ren
School of Information Science and Engineering
Central South University
Changsha, China
College of Information Science and Technology
Hunan Agricultural University
Changsha, China
renjun19@163.com

Jianxin Wang
School of Information Science and Engineering
Central South University
Changsha, China
jxwang@mail.csu.edu.cn

Min Li
School of Information Science and Engineering
Central South University
Changsha, China
Limin@mail.csu.edu.cn

***Abstract*—Identifying protein complexes from a PPI network is crucial to understand principles of cellular organization and functional mechanisms. However, it is still a difficult task because protein complexes have various topologies in PPI networks. In the paper, a novel protein complex identifying method, named LF-PIN, is proposed based on local fitness method. Firstly, LF-PIN calculates each PPI's weight based on its clustering value in the PPI network and selects seed edges by the edge weight. Then, protein complexes are extended from seed edges based on the evaluation of their neighbors' fitness values until their fitness reach the local maximum value. We apply the proposed algorithm LF-PIN and other nine previous algorithms, including HC-PIN, NFC, MCODE, DPClus, IPCA, CPM, MCL, CMC and Core-Attachment, to the PPI network of *S.cerevisiae* and compare their performances. Experimental results show that LF-PIN outperforms other competing algorithms in terms of matching with known complexes and functional enrichment.**

Keywords-edge clustering value; local fitness method; protein complex; PPI network

I. INTRODUCTION

Identifying protein complex is important in understanding the cellular organizations and functional mechanisms. However, the experimental methods to discover protein complexes are costly and time-consuming. Fortunately, with the development of high-throughput techniques such as yeast-two-hybrid [1], mass spectrometry [2], and protein chip technologies [3], protein-protein interactions (PPIs) are increasing fast and available conveniently. Furthermore, many evidences have indicated that cliques or dense subgraphs in PPI networks have high correlation with protein complexes [4]. Thus, many algorithms, such as CPM [5], IPCA [6], CMC [7], Core-Attachment [8], and DPClus [9], identify protein complexes in PPI networks based on detecting cliques or dense subgraphs.

These algorithms can only find protein complexes corresponding to dense subgraphs in PPI networks. However, many protein complexes are not dense subgraphs. To solve this problem, many researchers investigated topologies of protein complexes in PPI networks and found that many protein complexes are densely connected within themselves but sparsely connected with the rest of the PPI network [4,10]. Thus, *Wang et al*. [11] defined λ-module as protein complex model and proposed an agglomerative algorithm HC-PIN. As both dense subgraph and sparse subgraph have possibility to be λ-modules, HC-PIN can identify protein complexes with different densities. However, it is difficult to decide the λ threshold value because the λ values of known protein complexes in the PPI network are very different.

To solve this problem, *Lancichinetti et al.* defined a fitness function *f* of a subgraph and proposed NFC algorithm to identify protein complexes as subgraphs with the maximum value of fitness in a PPI network [12]. NFC can identify complexes with different densities and λ values. Complexes identified by NFC algorithm generally have large size. However, known protein complexes usually have small size. Thus, compared with the algorithms based on detecting dense subgraph, NFC algorithm usually has poor performance in identifying protein complexes.

To overcome above issues, we define a subgraph's fitness as the product of the subgraph's density and its λ value and propose a new algorithm, named LF_PIN, to identify protein complexes by extending each seed edge to a subgraph until its fitness reaches the local maximum value. LF_PIN chooses seed edges according to the edge clustering value because we find that the higher clustering value a PPI has, the more likely it is to be in a protein complex. The experimental results of *S.cerevisiae* show that LF_PIN outperforms the other competing algorithms in terms of matching with known complexes and functional enrichment.

II. METHOD

A. Quantitative Definition of Complex

As mentioned in introduction, most existing algorithms, such as CPM, IPCA, CMC, Core-Attachment, DPClus, HC-PIN and NFC, have their own drawbacks. To solve these

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

issues, we define a new fitness function f of a subgraph as the product of the subgraph's density and λ value. Then, we develop a new protein complex model as the subgraph with the maximum value of fitness in the PPI network. Like NFC algorithm, using this protein complex model can identify protein complex with different λ values. Moreover, compared with the complex identified by NFC algorithm, the size of the identified complex based on the new protein complex model is closer to that of known protein complex because our fitness function of a subgraph takes into account both the subgraph's density and its λ value. To describe the new fitness function f and the protein complex model clearly, we give some related definitions firstly.

For a weighted graph G, the density of a subgraph H ($H \subseteq G$) is donated as q_H, which is defined as follows:

$$q_H = 2 * m_H / n_H * (n_H - 1) \tag{1}$$

where m_H and n_H are the number of edges and vertices in H.

For a weighted graph G, the weighted degree of a vertex v is donated as $d_w(v)$, which is the sum of weights of the edges connecting v.

$$d_w(v) = \sum_{u \in V, <u,v> \in E(G)} w(u,v) \tag{2}$$

where m_H is the number of edges in H and n_H is the number of vertices in H.

For a vertex v in a subgraph $H \subseteq G$, its weighted in-degree, denoted as $d_w^{in}(H,v)$, is the sum of weights of edges connecting vertex v to other vertices in H, and its weighted out-degree, denoted as $d_w^{out}(H,v)$, is the sum of weights of edges connecting vertex v to other vertices in G-H.

$$d_w^{in}(H,v) = \sum_{u,v \in H, <u,v> \in E} w(u,v) \tag{3}$$

$$d_w^{out}(H,v) = \sum_{v \in H, u \notin H, <u,v> \in E} w(u,v) \tag{4}$$

Based on the weighted in-degree and weighted out-degree of a vertex in a subgraph $H \subseteq G$, the λ value of the subgraph H, denoted as λ_H, is defined as the sum of weighted in-degrees of vertices in H, divided by the sum of weighted out-degrees of vertices in H.

$$\lambda_H = \sum_{v \in H} d_w^{in}(H,v) \Big/ \sum_{v \in H} d_w^{out}(H,v) \tag{5}$$

Based on the density and λ value of a subgraph $H \subseteq G$, the fitness of the subgraph H, denoted as f_H, is defined as the product of the density and λ value of H.

$$f_H = q_H * \lambda_H \tag{6}$$

Based on the above definition of fitness, the fitness of a vertex v ($v \in G$ and $v \notin H$) with respect to a subgraph H, denoted as f_H^v, is defined as the difference of the fitness of the subgraph H with and without vertex v.

$$f_H^v = f_{H+\{v\}} - f_{H-\{v\}} \tag{7}$$

where $f_{H+\{v\}}$ is the fitness of the subgraph H with vertex v added to it and $f_{H-\{v\}}$ is the fitness of the subgraph H with v removed from it.

When subgraph is a singleton edge, it has the maximum density of 1. When the subgraph is the whole graph, it has the maximum λ value of ∞. Generally, with the expanding of a subgraph, its λ value is increasing and its density is decreasing. Thus, by expanding from an edge, we can obtain a subgraph with the maximum value of fitness and output it as a complex. The process of a complex extending from an edge is adding neighbor vertices into the subgraph or removing vertices from the subgraph when the inclusion of a new neighbor vertex or the elimination of one vertex from the subgraph will increase the subgraph fitness.

B. Selection of Seed Edges

How to select the seed edges is very important for identifying protein complexes. Many evidences have indicated that a PPI with high edge clustering value in a PPI network has high possibility to be in a protein complex [11, 13]. Thus, in the section, we define the edge weight by quantitatively analyzing the correlation of the PPI's edge clustering value in a PPI network and its possibility in a protein complex and choose seed edges by edges' weights.

The edge clustering value is defined based on the edge clustering coefficient which is defined as the number of triangles to which a given edge belonged, divided by the number of triangles that might potentially include it [13]. The basic idea behind this definition is that a "small-world" network is composed of modules and many triangles exist within these modules and edges connecting vertices in different modules are included in few or no triangles [13]. Thus, the higher clustering coefficient an edge has, the more likely it is to be in a module. As PPI networks are "small-world" networks and their modules generally correspond to protein complexes [14], it seems reasonable to evaluate the possibility of a PPI to be in a complex by the PPI's clustering coefficient. However, this evaluation may be problematic because PPI networks are disassortative networks and edge clustering coefficient is not suitable for disassortative networks because of the small number of short cycles [13]. To break through such limitations, *Wang J et al.* calculated the common neighbors instead of the triangles or high-order cycles and defined the clustering value of edge $<u,v>$ as follows [11].

$$ECV(u,v) = |N_u \cap N_v|^2 / |N_u| * |N_v| \tag{8}$$

where N_u is the set of neighbors of vertex u and N_v is the set of neighbors of vertex v, respectively.

Wang J et al. pointed out that the higher clustering value an edge has, the more likely it is to be in a protein complex. [11]. However, they did not quantitatively analyze the correlation of the PPI's edge clustering value and its possibility in a protein complex. We give a quantitative analysis in Fig.1. Fig.1 shows the percentage of PPIs in protein complexes with respect to different range of edge clustering value (*ECV*). Here, the PPI network is a yeast PPI network which is downloaded from DIP database and the protein complex set is provided by *Pu S et al.* in [15].

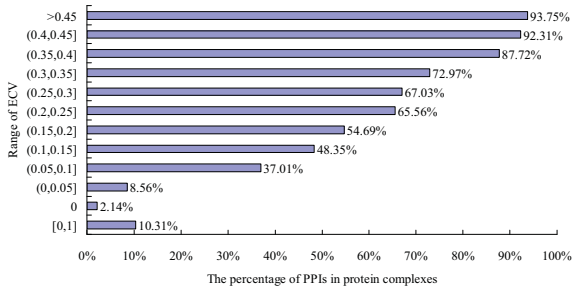

Figure 1. The percentage of PPIs in protein complexes with respect to different range of ECV

As shown in Fig.1, when ECV in the range of =0, (0,0.05], (0.05, 0.1], (0.1,0.15], (0.15, 0.2], (0.2,0.25], (0.25, 0.3], (0.3,0.35], (0.35, 0.4], (0.4,0.45], and >0.45, the percentage of PPIs in the protein complex set are 2.1%, 8.6%, 37.0%, 48.3%, 54.7%, 65.6%, 67.0%, 73.0%, 87.7%, 92.3%, and 93.8%, respectively. It is obviously that with the increase of a PPI's ECV, the possibility of the PPI to be in a protein complex is also increasing. Thus, the possibility of a PPI in a complex is an increasing function of the PPI's clustering value. Considered that a PPI with ECV=0 also has little possibility to be in a protein complex, the weight of this PPI is set as a small constant. So the weight of an edge <u,v> in an un-weighted PPI network G is calculated as:

$$w(u,v) = \alpha + \beta * ECV(u,v) \qquad (9)$$

where α is a given small constant that reflect the possibility of the PPI with *ECV*=0 in a protein complex. Considered that the average weight of the whole PPI network is equal to 1, the value of β is equal to $(1-\alpha)/\gamma$, where γ is the average *ECV* of the whole PPI network.

C. Algorithm LF-PIN

Based on quantitative description of protein complexes, we propose a novel overlapping clustering algorithm LF-PIN (based on **L**ocal **F**itness). The detailed description of algorithm LF-PIN is shown in Fig. 2. The input of algorithm LF-PIN is a parameter α and an un-weighted PPI network which is described as a simple undirected graph *G(V, E)*.

Algorithm LF-PIN has four stages: weight calculating, seed selecting, seed expanding, and outputting. Firstly, algorithm LF-PIN calculates the weights of all edges in *G* according to the formula (9) and generates the weighted PPI network $G^W(V, E, W)$. Then, seed edges are selected as edges

Algorithm: LF-PIN
Input: Un-weighted PPI networks $G = (V, E)$, parameter α
Output: Identified Clusters
Process:
1. Weight Calculating:
   $W = \phi$;
   for each edge $(v_i, v_j) \in E$ **do** { $w(v_i, v_j) = ECV(v_i, v_j)$; $W \leftarrow w(v_i, v_j)$; }
   γ = the average value of W;
   for each weight $w(v_i, v_j) \in W$ **do** $w(v_i, v_j) = \alpha + ((1-\alpha)/\gamma) * w(v_i, v_j)$;
2. Seed Selecting:
   β = the average value of W;
   $Es = \phi$;
   for each edge $(v_i, v_j) \in E$ **do** { **if** $w(v_i, v_j) \geq \beta$ **then** $Es \leftarrow (v_1, v_2)$;}
   sort all edges in Es to queue Sq in non-increasing order of edge weights;
3. Seed Expanding:
   $C = \phi$; //queue C is used to store the identified clusters.
   while $Sq \neq \phi$ **do**
   { $(v_1, v_2) \leftarrow Sq$; // the first edge (v_1, v_2) in Sq is selected.
   $H = \{v_1, v_2\}$; // cluster H is initialized as two vertices v_1 and v_2.
   flag1 = 1;
   while flag1 = 1 **do**
   { flag1 = 0;
   for each neighbor vertex v_i of H in $G^W(V, E, W)$ **do**
   $f_H^{v_i} = f_{H+\{v_i\}} - f_{H-\{v_i\}}$;
   sort all neighbor vertex of H to queue Vq in non-increasing order by their f value;
   while $Vq \neq \phi$ **do**
   { $v_3 \leftarrow Vq$; flag3 = 1;
   if $f_H^{v_3} < 0$ **then**
   { $C = C \cup \{H\}$;
   remove edges which include vertices of H from Sq;
   break; }
   else
   { $H = H + \{v_3\}$; //the neighbor vertex v_3 is added to H.
   $f_H^{v_1} = f_{H+\{v_1\}} - f_{H-\{v_1\}}$;
   $f_H^{v_2} = f_{H+\{v_2\}} - f_{H-\{v_2\}}$;
   if $f_H^{v_1} < 0$ **or** $f_H^{v_2} < 0$ **then**
   { $H = H - \{v_3\}$; flag3 = 0; }
   else
   { flag2 = 1;
   while flag2 = 1 **do**
   { flag2 = 0;
   for each vertex v_j of H in $G^W(V, E, W)$ **do**
   { $f_H^{v_j} = f_{H+\{v_j\}} - f_{H-\{v_j\}}$;
   if $f_H^{v_j} < 0$ **then**
   { $H = H - \{v_j\}$; flag2 = 1; **break**;}
   }}}
   if flag3 = 1 **then** { flag1 = 1; **break**;}
   }}
   if $Vq = \phi$ **and** flag3 = 0 **then**
   { $C = C \cup \{H\}$;
   remove edges which include vertices of H from Sq;}
   }}
4. output C.

Figure 2. Decription of algorithm LF-PIN

whose weights no less than average weight and sorted into seed queue *Sq* in non-increasing order by the edge weight. Thirdly, when the seed queue *Sq* is not null, LF-PIN will always select the first edge in *Sq* as the seed edge and gradually add neighbor vertex or remove vertex decided by

the measure of vertex fitness. Each loop will stop when the seed edge is not removed and all neighbor vertices have negative fitness, and an identified cluster is produced. At the same time, all the edges which include vertices in the identified cluster are removed from *Sq*. The seed expanding processes will stop when the seed queue *Sq* is null. At last, LF-PIN outputs all identified clusters.

III. Result and Discussion

To evaluate the performance of algorithm LF-PIN, we compare it with nine previous competing algorithms: CPM, IPCA, CMC, Core-Attachment, DPClus, MCODE, HC-PIN, NFC, and MCL. MCODE, DPClus and IPCA are density-based local search algorithms. CPM, CMC and Core-Attachment identify overlapping protein complexes based on detecting cliques in PPI networks. NFC is also a fitness-based local search algorithm. MCL is a fast and highly scalable cluster algorithm for networks based on stochastic flow. HC-PIN is a fast hierarchical clustering algorithm. NFC, MCL and HC-PIN can all identify protein complexes with different density. The values of the parameters in each algorithm are selected from those recommended by the authors.

The original un-weighted protein interaction network of *S.cerevisiae* is downloaded from DIP database [16], which is released on June 14, 2010. We remove all self-connecting interactions and repeated interactions. The final network, named *YPPI*, includes 4,938 proteins and 21,759 interactions.

In the section, the effect of parameter α on clustering results is discussed firstly. Then, all the identified complexes of LF-PIN and those of nine other algorithms are compared with the known protein complexes in [15]. Finally, all the identified complexes of LF-PIN and those of nine other algorithms are compared in terms of functional enrichment.

A. Effect of parameter α on clustering results

Firstly, we apply the proposed algorithm LF-PIN to the un-weighted PPI network *YPPI* and evaluate the effect of parameter α on clustering results in Table 1. Parameter α is the weight of a PPI with *ECV*=0. As the average weight of the whole PPI network is equal to 1 and the PPI with *ECV*=0 has much less possibility to be in a protein complex than a randomly selected PPI, the value of α is usually much small than 1. So, we change the values of parameter α from 0 to 0.5 with 0.1 increments and achieve six different output sets of clusters from the *YPPI*. Table 1 lists their characteristics. The overlapping rate of a complex set is used to evaluate the overlap of all complexes in the set and defined as [17]:

$$Or_{Cset} = \sum_{C_i \in Cset} |C_i| \Big/ |\cup C_i| \qquad (10)$$

where *Cset* is a cluster set, $|C_i|$ is the number of vertices in cluster C_i, $|\cup C_i|$ is the total number of vertices in *Cset*.

As shown in Table 1, with the increase of α, the number of identified complexes and the overlapping rate of the identified complex set are increasing slowly, the average size and maximum size of identified complexes are decreasing slowly, and the minimum size of identified complexes is fixed. Changes of all characteristics are small, which means that LF-PIN has good robustness. So, in the paper, the value of parameter α is randomly selected as a medium value of 0.2.

TABLE I. The effect of the varying of α on clustering

α	Number	Average size	Maximum size	Minimum size	Overlapping rate
0	342	4.78	96	2	1.11
0.1	358	4.53	88	2	1.13
0.2	365	4.55	88	2	1.17
0.3	381	4.51	84	2	1.21
0.4	392	4.37	62	2	1.22
0.5	395	4.40	62	2	1.24

B. Comparison with known complexes

To directly validate the effectiveness of algorithm LF-PIN for identifying protein complexes, we compare the protein complexes predicted by LF-PIN and other nine algorithms with the known protein complexes obtained from [15]. Here, the results of LF-PIN and other nine algorithms are assessed using an evaluation metric employed by earlier studies [9, 10, 11], which is to determine how effectively a predicted complex (*Pc*) matches a known complex (*Kc*). The overlapping score *OS(Pc, Kc)* between a predicted complex *Pc* and a known complex *Kc* is calculated as:

$$OS(Pc, Kc) = |V_{Pc} \cap V_{Kc}|^2 \Big/ |V_{Pc}| * |V_{Kc}| \qquad (11)$$

where $|V_{Pc}|$ is the number of proteins in the predicted complex and $|V_{Kc}|$ is the number of proteins in the known complex. A predicted complex *Pc* and a known complex *Kc* are considered as a match if their overlapping score *OS(Pc, Kc)* is no less than a specific threshold. If *OS(Pc, Kc)* is equal to 1, we say that they are perfectly matched. The percentage of matched predicted complexes of LF-PIN and other algorithms is shown in Fig. 3.

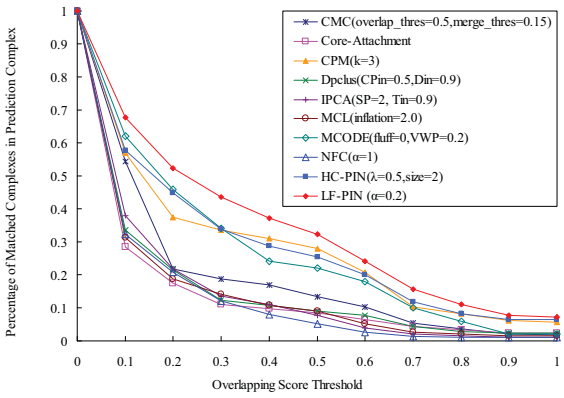

Figure 3. Comparison of the percentage of matched predicted complexes of LF-PIN and other algorithms with respect to different overlapping scores threshold.

Generally, a predicted complex and a known complex are considered as a match if their overlapping score is no

less than 0.2 [9, 10, 11]. We can see from Fig. 3 that when overlapping score's threshold is equal to 0.2, more than half of complexes detected by LF-PIN are matched by the known complexes. This ratio is much higher than those identified by other nine competing algorithms at the same threshold. Furthermore, Fig. 3 shows that for each overlapping score's threshold, the percentage of matched complexes in the complex set identified by LF-PIN is much higher than those in the complex sets identified by other nine competing algorithms. All these indicate that LF-PIN outperforms other nine competing algorithms in terms of matching with the known complexes.

To compare the performance of LF-PIN with those of other nine competing algorithms, we calculate *Sensitivity*, *Specificity*, *F-score*, the number and the percentage of perfect matches of the ten algorithms and list them in Table 2. Here, the overlapping score's threshold is selected as 0.2. *Sensitivity*, *Specificity* and *F-score* are three popular evaluation criteria that are used to quantify the quality of protein complexes detection methods. *Specificity* (*Sp*) is the fraction of the predicted complexes that are matched by the known complexes among all the predicted complexes [10]. *Sensitivity* (*Sn*) is the fraction of the known complexes that are matched by the predicted complexes among all the known complexes [10]. *F-score* combines the *Sensitivity* and *Specificity* and is defined as [11]:

$$F-score = 2 * Sn * Sp/(Sn + Sp) \qquad (12)$$

TABLE II. COMPARISON OF *SENSITIVITY*, *SPECIFICITY*, *F-SCORE*, THE NUMBER AND THE PERCENTAGE OF PERFECT MATCHES OF LF-PIN AND OTHER NINE COMPETING ALGORITHMS

	Num-ber	Perfect match	Sn	Sp	F-score
LF-PIN (α=0.2)	365	26 (7.12%)	0.472	0.523	0.496
HC-PIN (λ=0.5, size=2)	265	17 (6.42%)	0.318	0.449	0.373
NFC (α=1)	518	5 (0.97%)	0.277	0.209	0.238
MCODE (*fluff*=0, *VWP*=0.2)	50	1 (2.00%)	0.057	0.460	0.101
DPClus (CP_{in}=0.5, D_{in}=0.9)	1200	27 (2.25%)	0.651	0.216	0.324
IPCA (*SP*=2, T_{in}=0.9)	3839	40 (1.04%)	0.933	0.219	0.355
CPM (*k*=3)	197	11 (5.58%)	0.185	0.376	0.248
MCL (*inflation*=2.0)	929	15 (1.61%)	0.450	0.187	0.264
CMC (AdjstCD=1, *overlap_thres*=0.5, *merge_thres*=0.15)	1130	21 (1.86%)	0.576	0.219	0.317
Core-Attachment	1358	31 (2.28%)	0.589	0.174	0.268

As shown in Table 2, the number of complexes identified by LF-PIN is 365, which is much less than those of DPClus, IPCA, MCL, CMC, and Core-Attachment. Obviously, the more complexes dose an algorithm identify, the more perfect matches and matched complexes dose the algorithm identify. Thus, LF-PIN identifies less perfect matches than DPClus, IPCA, and Core-Attachment, and its *Sensitivity* value is less than those of DPClus, IPCA, CMC, and Core-Attachment. However, we can see from Table 2 that the percentage of perfect matches in the identified complexes and the *Specificity* value of LF-PIN are both higher than those of the nine other algorithms, which means that the percentages of perfect matches and matched complexes in the complexes identified by LF-PIN are both higher than those of the nine other algorithms. Moreover, LF-PIN has the highest *F-score* value in the ten algorithms. Even compared with the highest *F-score* value of the nine other algorithms, 33.18% improvement can be obtained by using LF-PIN algorithms.

To give a more complete comparison, we compare the *F-score* of LF-PIN and other nine algorithms with respect to different overlapping score's thresholds in Fig. 4. As shown in Fig.4, LF-PIN algorithm has the highest value of *F-score* in the ten algorithms when overlapping score's threshold is no less than 0.2, which means it has the best performance for identifying protein complexes in the ten algorithms.

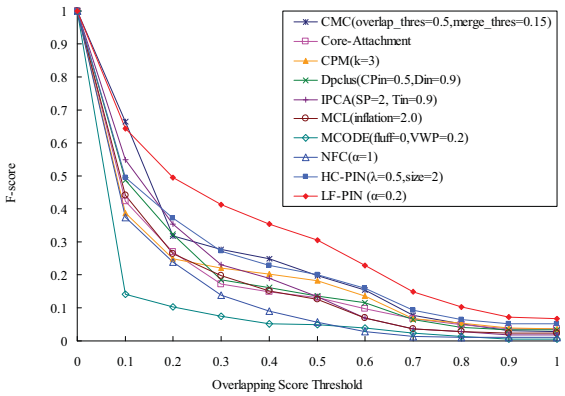

Figure 4. Comparison of *F-score* of LF-PIN and other algorithms with respect to different overlapping score's thresholds.

C. *Comparison with other algorithms in terms of functional enrichment*

To evaluate the effectiveness of LF-PIN, we also compare it with the nine algorithms based on the enrichment of biological function. For this validation, the P-value of a cluster with a given GO term is used to estimate whether the proteins in the cluster are enriched for the GO term with a statistically significant probability compared to what one would expect by chance. The smaller P-value indicates the identified cluster is not accumulated at random and is more biologically significant than the one with a larger P-value [18]. As a protein complex can have various P-values for various GO terms, its P-value defaults to its minimum P-value.

We compare the functional enrichment of complexes identified by each algorithm in Table 3. Table 3 lists the percentages of the identified complexes whose P-value falls within <E-15, [E-15, E-10], [E-10, E-5], [E-5, 0.01] and ≥0.01. Generally, a complex with P-value≥0.01 is considered insignificant and that with P-value<0.01 is considered significant. As shown in Table 3, 73.97% complexes identified by LF-PIN are significant. The percentage is much higher than those of other algorithms (except MCODE). On the contrary, the percentage of

insignificant complexes identified by LF-PIN is only 26.03%, which is much less than that detected by other algorithms (except MCODE). Furthermore, the percentage of insignificant complexes identified by LF-PIN is less than half of those identified by NFC, DPClus, IPCA, MCL, CMC, and Core-Attachment. The statistical results show that LF-PIN is more effective for identifying significant proteins complexes than other algorithms (except MCODE).

TABLE III. COMPARING THE FUNCTIONAL ENRICHMENT OF PROTEIN COMPLEXES IDENTIFIED BY LF-PIN, AND THE NINE OTHER ALGORITHMS

Algorithms	<E-15	[E-15,E-10]	[E-10,E-5]	[E-5, 0.01]	≥0.01
LF-PIN	28(7.7%)	41(11.2%)	83(22.7%)	118(32.3%)	95(26.0%)
HC-PIN	26(9.8%)	14(5.3%)	42(15.9%)	84(31.7%)	99(37.4%)
NFC	25(4.8%)	22(4.3%)	81(15.6%)	124(23.9%)	266(51.3%)
MCODE	8(16.0%)	0(0.0%)	26(52.0%)	14(28.0%)	2(4.0%)
DPClus	10(0.8%)	32(2.7%)	155(12.9%)	329(27.4%)	674(56.2%)
IPCA	161(4.2%)	244(6.4%)	500(13.0%)	810(21.1%)	2124(55.3%)
CPM	10(5.2%)	15(7.6%)	49(24.9%)	42(21.3%)	81(41.1%)
MCL	22(2.4%)	32(3.4%)	114(12.3%)	239(25.7%)	522(56.2%)
CMC	33(2.9%)	40(3.6%)	191(16.9%)	292(25.9%)	574(50.8%)
Core-Attachment	39(2.7%)	37(2.9%)	122(9.0%)	287(21.1%)	873(64.3%)

IV. CONCLUSION

In the post-genome era, one of the most important works is to discover the protein complexes. In this paper, we propose a new fitness-based local search algorithm LF-PIN to identify protein complexes in an un-weighted PPI network. The proposed algorithm LF-PIN and other nine competing algorithms, including MCODE, DPClus, IPCA, CPM, CMC, Core-Attachment, NFC, MCL, and HC-PIN, are applied to the protein interaction network of *S.cerevisiae* and compared their performance in terms of functional enrichment and matching with the known complexes. The experimental results show that LF-PIN identifies much more significant protein complexes with smaller P-values and generates much less insignificant protein complexes than other algorithms (except MCODE). When matching with the known protein complexes, for each overlapping score's threshold, LF-PIN has the highest percentage of matched predicted complexes and the highest *F-score* in the ten algorithms. These quantitative comparisons reveal that our algorithm LF-PIN outperforms the other previous competing algorithms in identifying protein complexes. Moreover, algorithm LF-PIN has good robustness.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grant No.61003124 and No.61073036, the Talent Foundation of Hunan Agricultural University No.06YJ10, the Ph.D. Programs Foundation of Ministry of Education of China No.20090162120073, the Freedom Explore Program of Central South University No.201012200124, the U.S. National Science Foundation under Grants CCF-0514750, CCF-0646102, and CNS-0831634, and the Natural Sciences and Engineering Research council of Canada (NSERC).

REFERENCES

[1] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," Nature, vol. 403, no. 6770, Feb. 2000, pp. 623-627.

[2] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader et al., "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," Nature, vol. 415, no. 6868, Jan. 2002, pp. 180-183.

[3] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone et al., "Global analysis of protein activities using proteome chips," Science, vol. 293, no. 5537, Sep. 2001, pp. 2101-2105.

[4] V. Spirin and L.A. Mirny, "Protein complexes and functional modules in molecular networks," Proc. Natl. Acad. Sci. USA, vol. 100, no. 21, Sep. 2003, pp. 12123-12128.

[5] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncoverring the overlapping community structure of complex networks in nature and society," Nature, vol. 435, no. 7043, Jun.2005, pp. 814-818.

[6] M. Li, J. Chen, J. Wang, B. Hu and G. Chen, "Modifying the DPClus Algorithm for Identifying Protein Complexes Based on New Topological Structures," BMC Bioinformatics, vol. 9, Sep. 2008, pp. 938.

[7] G. Liu, L. Wong, and H.N. Chua, "Complex Discovery from Weighted PPI Networks," Bioinformatics, vol. 25, no. 15, May. 2009, pp. 1891–1897.

[8] H.C.M. Leung, Q. Xiang, S.M. Yiu, and F.Y.L. Chin, "Predicting Protein Complexes from PPI Data: A Core-Attachment Approach," J. Comput Biol, vol. 16, no. 2, Feb. 2009, pp. 133-144.

[9] 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, Apr. 2006, pp. 207-219.

[10] G.D. Bader and C.W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics, vol. 4, Jan. 2003, pp. 2.

[11] J. Wang, M. Li, J. Chen, and Y. Pan, "A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks," IEEE/ACM Trans Comput Biol Bioinform, vol. 8, no. 3, May/June, 2011, pp. 607-620.

[12] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics, vol. 11, no. 3, Mar. 2009, pp. 033015.

[13] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proc. Natl. Acad. Sci. USA, vol. 101, no. 9, Mar. 2004, pp. 2658-2663.

[14] F. Luo, Y. Yang, C Chen, R. Chang, J. Zhou and R. H. Scheuermann, "Modular organization of protein interaction networks," Bioinformatics, vol. 23, no. 2, 2007, pp. 207-214.

[15] 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, Dec. 2008, pp. 825-831.

[16] I. Xenarios, L. 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 Res., vol. 30, no. 1, Jan. 2002, pp. 303-305.

[17] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, et al, "The FunCat: a functional annotation scheme for systematic classication of proteins from whole genomes," Nucleic Acids Research, vol. 32, no. 18, Oct. 2004, pp. 5539-5545.

[18] A.D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," Bioinformatics, vol. 20, no. 17, 2004, pp. 3013-3020