Detecting Protein Complexes in PPI Networks: The Roles of Interactions

Xiaoke Ma, Lin Gao*
School of Computer Science and Technology, Xidian University, 710071, PR China
Email: maxiaoke8218@163.com; lgao@mail.xidian.edu.cn

***Abstract*—Studying protein complexes is very important in biological processes since it helps reveal the structure-functionality relationships in protein complexes. Most of the available algorithms are based on the assumption that dense subgraphs correspond to complexes, fail to take into account the inherence organization within protein complex and the roles of edges. To investigate the roles of edges in PPI networks, we show that the edges connecting less similar vertices in topology are more significant in maintaining the global connectivity, indicating the weak ties phenomenon in PPI networks. By using the concept of bridgeness, a reliable virtual network is constructed, in which each maximal clique corresponds to a core. By this notion, the detection of the protein complexes is transformed into a classic all-clique problem. A novel core-attachment based method is developed, which detects the cores and attachments, respectively. Finally, a comprehensive comparison between the existing algorithms and our algorithm has been made by comparing the predicted complexes against benchmark complexes. The experimental results on the yeast PPI network show that the proposed method outperforms the state-of-the-art algorithms and analysis of detected modules by the present algorithm suggests that most of these modules have well biological significance in context of complexes, implying that the role of interactions is a critical and promising factor in extracting protein complexes.**

keywords: protein-protein interaction network; protein complexes; weak tie effect; clique

I. Introduction

Interpretation of the completed biological genome sequences initiated a decade of landmark studies addressing the critical aspects of cell biology on a system-wide level, including gene expression analysis [1], [2], gene disruptions detection [3], [4], identification of protein subcellular location [5], [6] and so on. Among them one important and challenge task in proteomics is the detection of protein complexes from the available protein-protein interaction (PPI) networks generated by various experimental technologies such as yeast-two-hybrid [7], affinity purification [8], mass spectrometry [9], etc.

Protein complexes, consisting of molecular aggregations of proteins assembled by multiple protein interactions, are of the fundamental units of macro-molecular organizations and play crucial roles in integrating individual gene products to perform useful cellular functions. It is confirmed by the fact that the complex 'RNA polymerase II' transcribes genetic information into messages for ribosomes to produce proteins. Unfortunately, the mechanism for most of biological activities is still unknown and hence accurately predicting protein complexes from the available PPI data has a considerable merit of practice because it allows us to infer the principles of biological processes.

The general methods for protein complexes prediction are based on experimental and computational notions. Experimentally, the Tandem Affnity Purification (TAP) with mass spectrometry [9] turns out to be popular. However, it is far away from a satisfying answer because there are several limits on the TAP [11], [8]. That's why the computational approaches are becoming promising alternatives to complement the experimental ones.

Generally, protein interaction data can be effectively modeled as a graph (also called a network) by regarding each protein as a vertex and each interaction as an edge. It is still non-trivial to design an efficient algorithm to mine protein complexes from PPI networks largely due to the fact that there has not been an exact definition for a protein complex. To overcome it, Tong *et al* [13] assumed that a complex corresponds to a dense subgraph since proteins in the same complex interact frequently among themselves, and similar discussion was also made in [14].

Markov Cluster Algorithm (MCL) [18], [17] is a popular method by simulating random walks within graphs. Molecular Complex Detection (MCODE) [19] relied on the topological structure of a network to infer the protein complexes. CFinder [20] defined a dense subgraph by using the k-cliques. Other non-topological properties such as the functional information [23] and data of protein binding interface [24] are also incorporated into algorithms with an immediate purpose to improve the accuracy of prediction. In addition, there are some other for detection of protein complexes relying solely on TAP date [25], [26], [27]. Recently, Gavin *et al* [8] have proved that the protein complexes consists of two components: core component and attachment component. It sheds light on the protein complex detection. Leung *et al* [15] proposed the CORE algorithm, a statistical framework to identify protein-complex cores. Wu *et al*. [16] presented the excellent algorithm, named by Coach. Ma *et al*. [30] also proposed a core-attachment algorithm by exploring the graph communicability. They all outperforms the existing state-of-the-art methods dramatically, indicating the critical role of the core-attachment structure in discovering protein complexes. Further information concerning the computational approaches

for predicting protein complexes can be referred to [28], [29].

At present a major problem confounding the existing computational algorithm is that, available PPI networks are too sparse, for instance, the average numbers of interactions per protein are 5.29, 6.98, and 10.62 in DIP [39], Krogan [27], and Gavin [8], respectively. In these PPI networks, many protein complexes are difficult to extract since the sparse networks are full of noises [41]. Unfortunately, previous algorithms did not pay enough attention to the problem since they only filter the noises by deleting nodes with degree 1 based on the fact that the interactions between proteins have lower reliability to the topological reliability measures [42].

Aside from issues of noise, all the existing computational approaches only make use of the topological structure information from vertices without the roles of edges. It, however, is unreasonable since an edge may play an important role in enhancing the locality or be significant in maintaining the global connectivity. For example, the famous weak ties theory [31] indicates the job opportunities and new ideas are usually from persons with weak connections. Furthermore, the weak ties can be used to characterized the topological properties of networks such as the stability of biological functions [32], the accuracy of network structure prediction [33], the structure in mobile communication networks [34]. Motivated by these observations, we pose the following question:

Q: *whether the roles of edges can be used in protein complexes detection*?

In this paper, we investigate the possibility to extract protein complexes by exploring the roles of edges and provide an affirmative answer to the above question. First, we prove that in PPI networks the edges connecting less similar nodes are more significant in maintaining the global connectivity. By using the weak ties, a reliable virtual network is constructed from the original PPI network, in which each maximal clique corresponds to a protein complex. A novel core-attachment based method is developed, which detects the cores and attachments, respectively. A comprehensive comparison between the existing algorithms and our algorithm has been made by comparing the predicted complexes against benchmark complexes. The experimental results on the yeast PPI network show that the proposed method outperforms DPClus [36], DECAFF [37], MCL [18], MCODE [19] and Coach [16]. Further, the analysis of detected modules by the present algorithm suggests that most of these modules have well biological significance in context of complexes, suggesting that the roles of edges are critical and promising in discovering protein complexes.

This paper is organized as follows. Section II shows the weak ties phenomenon in PPI networks. Section III contains the algorithm. The experiments and conclusion are proposed in Section IV and V, respectively.

II. Weak ties phenomenon in PPI networks

Edges in a network usually have two roles to play: some contribute to the global connectivity like the ones connecting two clusters while others enhance the locality like the ones inside a cluster. In social networks, the two roles are reflected as two important phenomena, being respectively the homophily and weak ties effects [38]. Homophily demonstrates that connections are more likely to be formed among individuals with close background, common characteristics. On the other hand, the weak ties phenomenon shows that the less similar individuals are prone to be connected with weaker strength. It has been proved that the weak ties phenomenon exists in the mobile communication [34] and document networks [35]. But, the weak ties effects are less studied for the PPI network.

To investigate the weak ties effects in PPI networks, we quantify how the topological structure changes according to an edge percolation process. In detail, if the weak ties effect exists in terms of topological similarity, the network disintegrates faster when we delete edges successively in an ascending order of the similarity than in descending order. Similar to [35], two measures are employed to quantify how topological structure changes when the edges are removed. The first one is the fraction of vertices contained in the giant component, represented by R_{GC}. The second one is the normalized susceptibility, defined as

$$\tilde{S} = \sum_{s<s_{max}} s^2/N, \tag{1}$$

where s is the size of a connected subgraph, N is the size of the whole network and the sum includes all connected components. An obvious gap occurs when the network disintegrates [43].

Prior to study the percolation, the definition of bridgeness of an edge should be discussed. In [35] it is defined as

$$B = \sqrt{C_u C_v}/C_{(u,v)}, \tag{2}$$

where (u, v) is an edge with u, v being the endpoints, C_u is the size of the maximal clique containing vertex u and $C_{(u,v)}$ is the size of the maximal clique containing (u, v). It fails to take into account the difference between u and v. Actually, if (u, v) is a bridge, the roles of u, v should differ greatly. Therefore, a new bridgeness is defined as

$$B_{(u,v)} = (1 - J(u, v))\frac{\sqrt{C_{u\setminus v} C_{v\setminus u}}}{C_{(u,v)}}, \tag{3}$$

where $J(u, v)$ is the Jaccard similarity, i.e., $J(u, v) = \frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|}$ with $N(u)$ being the neighbors, and $C_{u\setminus v}$ is the size of the maximal clique containing u without v. The $1 - J(u, v)$ measures the dissimilarity between the pair of endpoints while the latter component quantifies the relation between the neigbors of two endpoints. The topological similarity for protein pair is defined as

$$Sim = A + \beta A^2 + \beta^2 A^3, \tag{4}$$

where β is parameter controlling the relevant importance of each component and A is the adjacency matrix of the network involved. Here, we set β=0.618.

Fig.1 shows the edge percolation results on the DIP data, which shows R_{GC} decreases much faster when the less similar edges are removed firstly. As shown in Fig.1(b), a sharp peak occurs when the edges removed from the weakest to the

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

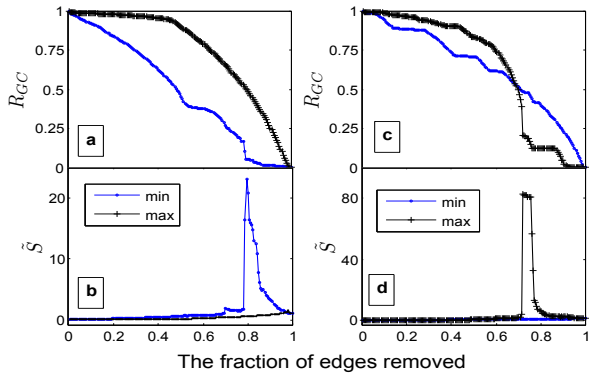

Fig. 1. Plots(a) and (b) are for the topological similarity, while (c) and (d) are for bridgeness. In (a) and (b), the min- (max-) lines represent the processes where the edges are removed from the least (most) similar to the most (least) similar ones. In (c) and (d), the min- (max-) lines denote the processes where the edges with smaller (larger) bridgeness are removed firstly.

strongest one, demonstrating the disintegration of the networks involved. Careful comparison of Fig.1(a)(b) further shows that no percolation phase transition appears since there is no clear peak. These strongly supports the weak ties phenomenon in the PPI networks. How good the bridgeness characterizes the weak ties phenomenon has been investigated in Fig.1(c)(d). Fig.1 (c) indicates that R_{GC} decreases much faster when the stronger bridges are removed firstly. As shown in Fig.1(d), a sharp peak occurs when the edges removed from the strongest to the weakest one, demonstrating the disintegration of the networks involved. It is enough to assert that the bridgeness is an excellent alternative to describe the tie strength.

In addition, the relations between the topological similarity and bridgeness are studied in Fig.2, which demonstrates that there is a negative correlation between bridgeness and topological similarity, i.e., the weaker the similarity between a pair of proteins is, the stronger its bridgeness is. In next section, we will show how the bridgeness can be used to discover complexes.

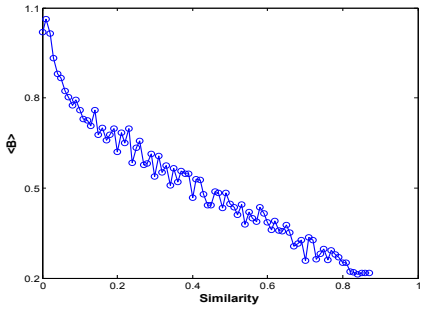

Fig. 2. The relation between bridgeness and topological similarity in PPI networks: <B> is the average bridgeness values of edges with same topological similarity.

III. Method

The key idea behind our algorithm consists of two main steps: (1) construct a reliable network by exploring the roles of edges; (2) identify the protein complexes.

A. Constructing a reliable network

Gavin *et al* [8] have pointed out that the core of a complex has relatively more interactions while the attachment proteins bind to the core proteins to form a protein complex, implying that the connectivity of a core is better than the whole complex.

To assess the topological proximity of a core, the measure of proximity of a pair of vertices should be handled beforehand. The most commonly used measures for topological proximity would be the graph distance, that is, the length of the shortest path connecting the pair of vertices. However, this quantity is not appropriate for the biological networks largely because of the two drawbacks: first, it does not take into account the global structure of the networks; second, it is very susceptible to the noises, e.g., a single missing edge may reduce proximity significantly. Thus, vertices connected via different paths are likely to be functionality closer than vertices connected via a single path. In detail, give an edge, say (u, v), it is reasonable to consider that the information transferred from u to v through the right channels. The more the channels, the better connectivity is. Actually, in biological network, the genetic information is transferred by the pathways. From the aspect of the graph theory, it is natural to consider the channels as various walks connecting u, v. Likewise, we also take into consideration the strength of paths: the strength of the effect via longer paths with more intermediate vertices is very likely to be lower than via shorter ones with fewer intermediaries. Given a walk of length k, say $v_1 \to v_2 \to \cdots \to v_{k+1}$, its strength is defined as the product of the weights on each edge in the walk, i.e., $\prod_{i=1}^{k} w_{i,i+1}$ where $w_{i,j}$ is the weight on the edge (v_i, v_{i+1}).

Given a un-weighted PPI network, how to assign weights to edges is one of the key steps in our algorithm. As shown in Fig.2, there is a negative correlation between bridgeness and topological similarity. Thus, a novel similarity for interactions based on the bridgeness in Eq.(3) is proposed as

$$D(u, v) = exp(-B_{(u,v)}). \tag{5}$$

The larger the bridgeness of an interaction is, the less similar it is.

Now it is sufficient to deal with the similarity between a pair of proteins via various lengths of walks. $(D^k)_{uv}$ denotes the sum of strengths of all walks of length k connecting u and v [44]. Since the connectivity in cores is high, any pair of proteins in the same core should be tightly connected by short walks. Therefore, the similarity for a pair of proteins is the sum of strengths of walks connecting them, which can be a generalization of Eq.(4) as

$$S = W + \beta W^2 + \beta^2 W^3, \tag{6}$$

where W is a matrix with element $(W)_{ij} = D(i, j)$.

For any protein pairs, if the similarity between them is large enough, we have enough reason to believe they should be connected, otherwise, un-connected. To construct a virtual and reliable network, similar to [30], a definition is proposed as

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Definition 1 *The reliable network* $\Phi(G,\tau) = (V_\tau, E_\tau)$ *of the PPI* $G = (V, E)$ *is the graph with* $V_\tau = V$ *and* $E_\tau = \{(u,v)|u,v \in V, \psi(S_{u,v},\tau) = 1\}$*, where* $\psi(x,\tau)$ *is a function defined as*

$$\psi(x,\tau) = \begin{cases} 1 & \text{if } x \geq \tau, \\ 0 & \text{otherwise.} \end{cases}$$

There are two good physic interpretation for S: first of all, if the similarity of a pair of proteins is considered as the reliable score on the corresponding edge, $\Phi(G)$ can also be considered as a reliable network, second, it can be understood as a perturbation of the original network by adding edges between vertices if there are enough short walks connecting them and deleting edges between vertex pairs if there are fewer short walks connecting them.

In this way, the core of a protein complex corresponds to a maximal clique in the virtual network. In the follows, we design algorithm to discover complexes by extracting cores and attachments, respectively.

B. A core-attachment algorithm

The first task is to extract all the maximal cliques in the virtual network, known as the classic all cliques problem, an NP Hard problem. Therefore, the exact algorithms are prohibit largely due to the complexity. Thus, we turn to some heuristic algorithms to avoid the time issue. The Coach algorithm detects the dense subgraphs very quickly and accurately from each vertex's neighborhood graphs [16]. We adopt the Protein-complex core mining algorithm in the Coach to identify approximately all cliques in the communicability graph $\Phi(G)$. Of course, others can be used to identify the cliques, for example, the greedy algorithm, the tabu search and so on.

What we would like to point out is that although we adopt the same strategy to detect the cores our algorithm differ greatly from Coach algorithm for two reasons: first, our algorithm detects core in a virtual network based on the weak ties phenomenon and L-percolation, while the Coach on the original network; second, the strategies for the attachment vary greatly.

Given a core denoted by a induced subgraph $G(U)$ with U is the protein set of the core and the virtual network $\Phi(G) = (V, E_\tau)$, one crucial step when revealing the attachments is to construct the candidate protein set $CS(U)$. For simplicity, we limit ourselves to only these proteins connected to at least one in U, i.e. $CS(U) = \{v|v \in V \setminus U, \exists u \in U \Rightarrow (u,v) \in E_\tau\}$.

What remains to be done is to determine the correct membership of each protein in $CS(U)$. To quantify the closeness of $v \in U$ to the core, a function is need. Here we present one based on the concept of bridgeness. If v is an attachment of G_U, there should be no protein $u \in U$ such that interaction (u,v) is bridge. In other words, there must be many short walks connecting v and vertices in U. Thus, we can define a new similarity function as

$$cl(v,U) = \frac{\sum_{u \in U} S_{vu}}{|U| + 1}, \tag{7}$$

which quantifies the average closeness of v to U from the aspect of connectivity. The larger $cl(v,U)$ is, the more walks connecting v and the core. Thus, a vertex $v \in S$ is selected as an attachment when the $cl(v,U) \geq acl(U \cup N(U)) = \frac{\sum_{v \in S} cl(v,U)}{|N(U)|+|U|}$, indicating that the selected attachment has more connection ways with U than the average connectivity in $N(U)$.

The procedure can be described in Algorithm 1.

Algorithm 1 Our algorithm

Input:

G: the PPI network;

τ: the parameter controls the strengthes of edges;

Output:

PC: the set of protein complexes;

1: Compute the bridgeness for each interaction in G according to Eq.(3);
2: Compute similarity matrix S as shown in Eq.(6);
3: Construct the virtual network $\Phi(G)$;
4: Extract the cores using Protein-complex core mining algorithm [16];
5: Detect the attachments for each core;

IV. Experiments

In this section we test the performance of our algorithm by verifying the prediction accuracy and biological meanings of the predicted protein complexes. it is coded using the MATLAB version 7.11.

A. Dataset

The Database of Interaction Protein (DIP) [39] is adopted, which consists of 4928 proteins and 17,201 interactions. To evaluate the predicted protein complexes, a benchmark set is constructed from the MIPS [40], Aloy *et al*. [21] and SGD database [22] based on the Gene Ontology (GO) notations. This benchmark set consists of 428 protein complexes.

In order to make a comprehensive comparison, the four competing algorithms, Coach algorithm [16], DPClus [36], DECAFF [37] and MCL [18], are selected deliberately.

B. F-measure and Coverage rate

The basic information of predictions by various compared algorithms is summarized in Table 1. From it, the MCL identifies 1116 complexes, of which 193 mach 242 real protein complexes; DPClus extracts 1143 complexes, of which 193 match 274 real complexes, DECAFF detects 2190 protein complexes, of which 605 match 243 ones and Coach reveals 746 complexes, of which 289 match 249 real ones. Our algorithm predicts 604 protein complexes, out of which 230 match 220 real ones in the benchmark.

Fig.3 shows the overall comparison in terms of F-measure [25] and coverage rate [12] on the DIP data. Although it is 2.9% lower than Coach algorithm, the F-measure of our algorithm is 43.2%, which is 16.7%, 16.5% and 6.0% higher than MCL, DPClus and DECAFF, respectively. It demonstrates that

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

TABLE I
THE RESULTS OF VARIOUS ALGORITHMS USING DIP DATA

	MCL	DPClus	DECAFF	Coach	Our's
Predicted complexes	1116	1143	2190	746	620
covered proteins	4930	2987	1832	1832	1702
N_{cp}	193	193	605	285	230
N_{cb}	242	274	243	249	220

TABLE II
STATISTICAL SIGNIFICANCE OF PROTEIN COMPLEXES OBTAINED BY VARIOUS ALGORITHMS ON DIP DATA

	MCL	DPClus	DECAFF	Coach	Our's
Predicted complexes	1116	1143	2190	746	620
Significant complexes	312	352	1653	622	519
Proportion (%)	34.2	30.8	75.5	83.4	83.7

our algorithm can predict protein complexes very accurately. From Fig.3, it is very easy to see that our method obtains the highest coverage rate of 42.8%, which is 7.9%, 9.6%, 11.4% and 16.2% higher than Coach,MCL, DPClus and DECAFF, respectively. It shows that the predicted complexes by our algorithm can cover the most proteins involved in the real complexes.

From Fig.3, we can make a conclusion that our algorithm is obviously outperform the MCL, DPClus and DECAFF, and it makes a better balance between the F-measure and Coverage rate than the Coach. Such results further demonstrate that the critical phenomenon in the PPI can be used for enhancing the prediction accuracy.

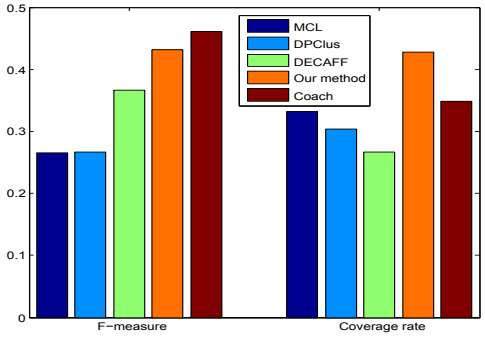

Fig. 3. The performance comparison for various algorithms on DIP data: the F-measure and Coverage rate.

C. P-value

To further investigate the biological significance of the predicted complexes, the P-value is adopted here. The functional homogeneity P-value is the probability that a given set of proteins is enriched by a given functional group merely by chance, following the hypergeometric distribution. It is the probability of co-occurrence of proteins with common functions. Accordingly, a low P-value of a predicted complex indicates that the collective occurrence of these proteins in the complex does not merely combine by chance and thus achieves high statistical significance. The values are calculated by the GO::TermFinder [10].

We discarded all clusters with P-value above a cutoff threshold. In the experiments, we chose a cutoff of 10^{-2} for each protein complex because it offers a compromise between complex-cluster matching rate and a clustering passing rate.

Table II shows the comparison results in terms of the proportion of significant protein complexes over all predicted ones. In the Table, our algorithm achieves the best performance (83.7%), implying the majority of predicted complexes are significant. Furthermore, the Coach has a comparative performance with our algorithm but the MCL and DPClus can only predict a small proportion of significant complexes. To further demonstrate the predicted protein complexes, 5 protein complexes with very low P-values, predicted by our method. The second columun is Table III refers to the ratio of the annotated proteins to ones in the identified complex.

The P-values of predicted complexes by our algorithm support that the role of interactions in PPI is a promising on enhancing the accuracy of prediction.

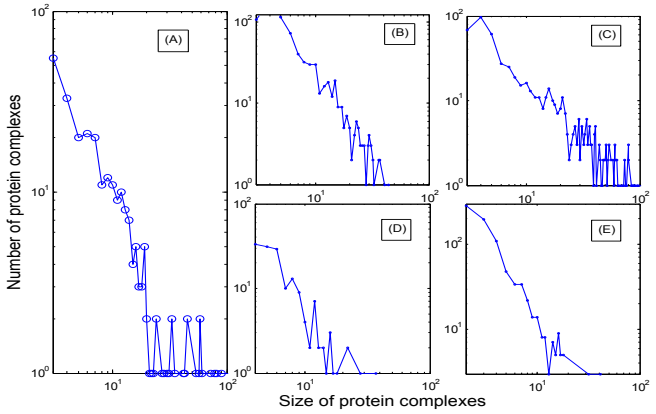

Fig. 4. Protein complex size distribution of various method and the benchmark set: (A) the benchmark set; (B) the Coach; (C) our algorithm; (D) the DPClus; (E) the MCL.

The module size distribution of predicted protein complexes for each compared methods on the DIP network has been shown in Fig.4. From it we can conclude that the major trend generated by our algorithm is very similar to that of the complexes in the benchmark set, which suggest that the definition of protein complex based on the weak tie effect is reasonable. However, the Coach can identify much less modules than these in the benchmark set, and its trend is different from that of the benchmark set. What we would like to point out is that the size distributions of the DPClus and MCL algorithms are very different from the previous ones.

Notice that our algorithm is quite different from those based on discovering the dense subgraphs because it makes use of the weak tie effect and various length paths. To verify the difference on the densities of the predicted complexes, we compared the Coach algorithm with our method in terms of the graph densities of the predicted complexes, shown in the Fig.5. It is easy to figure out that more than 50% complexes

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

TABLE III
SELECTED COMPLEXES PREDICTED BY OUR METHOD ON DIP DATA

ID	Match	P-value	Predicted complexes	Function
1	90.5%	5.44E-44	YBL002W YBR009C YBR154C YDL140C YDL150W YGL070C YJR063W YKL144C YKR025W YNL113W YNR003C YOR116C YOR151C YOR207C YOR210W YOR224C YOR341W YPR010C YPR110C YPR187W YPR190C	DNA-directed RNA polymerase activity
2	94.4%	8.77E-40	YDL150W YKL144C YKR025W YNL151C YNR003C YOR116C YOR207C YPR110C YBL002W YBR154C YDR045C YJR063W YNL113W YOR224C YOR341W YPR010C YPR187W YPR190C	RNA polymerase activity
3	100%	7.57E-26	YPL138C YDR469W YBR175W YHR119W YBR258C YAR003W YKL018W YLR015W	histone methyltransferase activity (H3-K4 specific)
4	88.2%	1.49E-20	YBL093C YBR253W YDR443C YNL025C YNL236W YOR140W YBR193C YCR081W YDL005C YER022W YGL151W YGR104C YHR041C YOL051W YOL135C YPL042C YPL248C	transcription regulator activity
5	100%	2.64E-21	Q0085 YBL099W YDR298C YDR377W YJR121W YKL016C YML081C-A YPL078C YPR020W	proton-transporting ATPase activity, rotational mechanism

predicted by the Coach algorithm whose densities are more than 0.9, while only 40% complexes predicted by our method whose densities are larger than 0.9. Furthermore, our algorithm can discover more protein complexes whose densities in range [0.6 0.9], which suggest that the density is not the only manner to characterize the protein complex and others are necessary and reasonable.

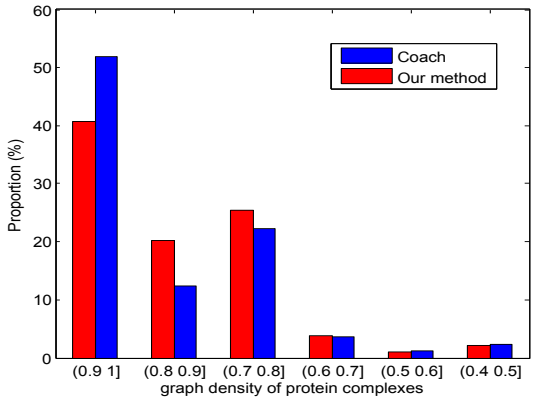

Fig. 5. The comparison on the density of predicted protein complexes from various algorithm.

D. Effects of parameters

The effects of parameters involved in the algorithm are studied in this subsection.

The value of τ controls the size of a core, the total number of cores in the virtual graph, and the connectivity 'strength' of the network involved. Therefore, we investigate its effect on the size of the virtual network. Fig.Fig.6 shows how the number of edges in the virtual network changes for various values of τ. From it, we cans see that the size of the virtual graph decreases dramatically when the value of τ increases from 0 to 0.4. Specifically, the size is approximately 3×10^4 if $\tau = 0.02$. The reason is that when the value of τ increases, only the edges whose connectivity is strong enough are maintained.

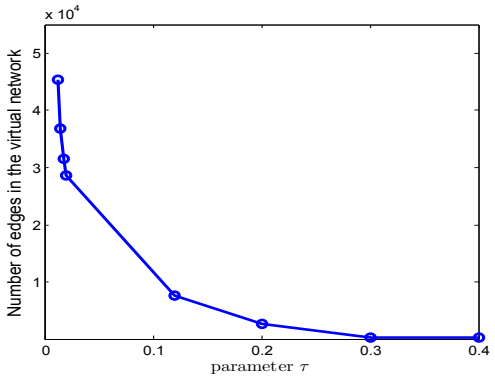

Fig. 6. The comparison on the density of predicted protein complexes from various algorithm.

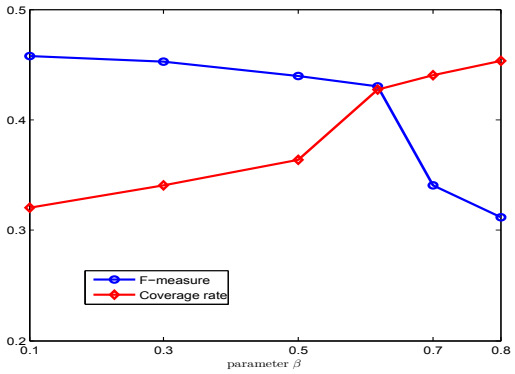

Fig. 7. The plot of the F-measure and Coverage rate for different values τ.

The parameter β controls the weights on the edges. Thus, we study its effect on the accuracy of prediction. Fig.7 demonstrates that the F-measure decreases, while the coverage rate increases when β increases. A possible reason is that the size of a maximal clique in the virtual network decreases when β increases, resulting in many small cores by dividing the large cores in the virtual graphs with small β. As β increases, more and more proteins in the PPI data are covered because

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

the number of predicted protein complexes increases. For this reason, the coverage rate keeps increasing. The value of β is 0.618.

E. Robustness analysis

The robustness analysis for the proposed algorithm is discussed in this subsection. The benchmark networks adopted here originated from Ref.[12]. In detail, from the protein complexes annotated in the MIPS database [40], an interaction network named a *test graph* is constructed by regarding each protein as a node and connecting each pair of nodes in the same complexes. The test graph has a poor value for assessing the robustness of the algorithms because each protein complex corresponds to a clique in the test graph. To solve this problem, the *altered graphs* are constructed from the test graph by adding or deleting the edges in various proportions. For the sake of convenience, the altered graph is denoted by $AG_{add,del}$ where add and del show the percentage of added and deleted edges, respectively.

In this experiment, only the MCL and Coach algorithms are selected for a comparison. The reason is that it is reported that the MCL is the most robust algorithms [12], and the Coach algorithm is the best core-attachment based method.

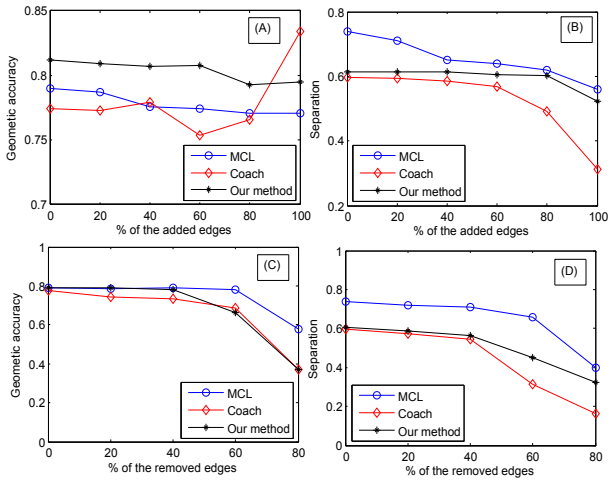

Fig. 8. The comparison on the density of predicted protein complexes from various algorithm.

The Fig.8(A) shows how the geometric accuracy fluctuates as the number of edges increases. Increasing proportions of edges were randomly added to the test graph from 0% to 100%. Both the MCL and our algorithm are barely affected by the additions of up to 100% edges, while the performance of Coach is acceptable for low values of noise, they change dramatically when the percentage of added edges increases to 40%. A good reason is that our algorithm is much more stable than the Coach algorithm is that, as the percentage of added edges increases, the added edges connecting to the vertices in different cliques yield larger complexes(through merging the small complexes). In this case, the altered graph is not suitable for correctly extracting the complexes by the Coach algorithm. However, our algorithm can remove the noise dramatically because it extracts the protein complexes in a virtual network, where some of the added edges are filtered by increasing the value of the threshold τ.

Fig.8(B) displays the impact of edge addition on the separation. We can see that both the MCL and our algorithm have good performances when the percentage of the added edges increases to 80%, while the performance of the Coach algorithm decreases when the percentage of added edges increases to 20%.

The impacts of edge removals on the geometric accuracy and separation are shown in Fig.8(C)(D), respectively. The Fig.8(C) demonstrates that both the MCL and our algorithm outperform the Coach algorithm and ours has competitive performance when the percentage of the removed edges is less than 20%. A possible reason is that, as more and more edges are deleted, it becomes more and more difficult to re-obtain the deleted edges. When the percentage of removed edges is more than 20%, the virtual network constructed by our algorithm differs greatly from the original test graph. The general trends in Fig.8(D) are similar to those displayed in Fig.8(C).

V. Conclusion

Protein complexes are key and basic molecular units in cellular functions and computational approaches to discovering accurately the unknown protein complexes hidden in the available PPI data are critical need. At present all these computational algorithms focus on the roles of proteins without taking into account the roles of interactions.

In this paper, we investigate the possibility to predict protein complexes with the roles of edges in PPI networks. Firstly, the weak tie phenomenon in the PPI network is proved by using the concept of bridge. Secondly, a reliable and virtual PPI network is constructed making use the relations of topological similarity and bridgeness. Finally, a core-attachment algorithm is designed. The experimental results demonstrate that the roles of edges in biological network is more promising than the roles of proteins, implying the significant importance of the roles of interactions.

Acknowledgment

This work was supported by the National Key NSFC (Grant No. 60933009), NSFC (Grant No. 61072103), SRFDPHE (Grant No. 200807010013) and FRFCU(Grant No. K50510030006).

References

[1] T. R. Hugher, et al, 2000 *Functional discovery via a compendium of expression profiles*, Cell, **102**:109–126.
[2] S. H. Neal, C. Amos, V. F. Nia and R. B. Jayanth, 2001 *Dynamic Modeling of Gene Expression Data*, Proc. Natl. Acad. Sci., **98**:1693–1698.
[3] P. Ross-Macdonald, et al, 1999 *Large-scale analysis of the yeast genome by transposon tagging and gene disruption*, Nature, **402**:413–418.
[4] E. A. Winzeler, et al, 1999 *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis*, Science, **285**:901–906.
[5] A. Kumar, et al, 2002 *Subcellular localizaion of the yeast proteome*, Genes. Dev., **16**:707-719.

[6] W. K. Hub, et al, 2003 *Global analysis of protein localization in budding yeast*, Nature, **425**:686–691
[7] T. Ito, T. Chila, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, 2001 *A comparehensive two-hybid analysis to explore the yeast protein interactome*, Proc. Natl. Acad. Sci., **98(8)**:4569–4574.
[8] A. C. Gavin, et al, 2002 *Functional organization of the yeast proteome by systematic analysis of protein complexes*, Nature, **415(6868)**:141–147.
[9] Y. Ho, et al, 2002 *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*, Nature, **415**:180–183.
[10] E. I. Boyle, S. Weng, J. Gollub, H. Jin, et al, 2004 *GO::TermFinder-open source software for accessing Gene Ontology information and finding sigificantly enriched Gene Ontology terms associated with a list of genes*, Bioinformatics, **20(18)**:3710–3715.
[11] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. Molina and I. Shames, 2008 *An in vivo map of the yeast protein interactome*, Science, **320(5882)**:1465–1470.
[12] S. Brohée and J. Kerbosch, 2006 *Evaluation of clustering algorithms for protein-protein interaction network*, BMC Bioinformatics, **7**: 488
[13] A. H. Tong, et al, 2002 *A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules*, Science, **295(5583)**:321-324.
[14] V. Spirin and L. A. Mirny, 2003 *Protein complexes and functional modules in molecular networks*, Proc. Natl. Acad. Sci., **100(21)**:12123-12128.
[15] H. C. Leung, Q. Xiang, S. M. Yiu and F. Y. Chin, 2009 *Predicting protein complexes from PPI data: a core-attachment approach*, Journal of Computational Biology, **16(2)**:133–144.
[16] M. Wu, X. Li, C. K. Kwoh and S. Ng, 2009 *A core-attachment based method to detect protein complexes in ppi networks*, BMC bioinformatics, **10** 169.
[17] A. J. Enright, S. V. Dongen and C. A. Ouzounis, 2002 *An efficient algorithm for large-scale detection of protein families*, Nucl. Aci. Res., **30(7)**:1575–1584.
[18] J. B. Pereira-Leal, A. J. Enright and C. A. Ouzounis, 2004 *Detection of functional modules from protein interaction networks*, Proteins, **54(1)**:49-57.
[19] G. Bader and C. Hogue, 2003 *An automated method for finding molecular complexes in large protein interaction networks*, BMC Bioinformatics, **10** 169
[20] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, 2006 *CFinder: locating cliques and overlapping modules in biological networks*, Bioinformatics, **22(8)**:1021–1023.
[21] P. Aloy, et al, 2004 *Structure-based assembly of protein complexes in yeast*, Science, **303(5666)**:2026–2029.
[22] S. S. Dwight, et al, 2002 Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go), Nucl. Acid. Res., **30(1)**:69–72.
[23] A. D. King, N. Prulj and I. Jurisica, 2004 *Protein complex prediction via cost-based clustering*, Bioinformatics, **20(17)**:3013–3020.
[24] S. H. Jung, et al, 2008 *Protein complex prediction based on mutually exclusive interactions in protein interaction network*, Genome Informatics, **21**:77–88.
[25] B. Zhang, B. Park, T. V. Karpinets and N. F. Samatova, 2008 *From pull-down data to protein interaction networks and complexes with biological relevance*, Bioinformatics, **24(7)**:979–986.
[26] A. C. Gavin, et al, 2006 *Proteome survey reveals modularity of the yeast cell machinery*, Nature, **440(7084)**:631–636.
[27] N. J. Krogan, et al, 2006 *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae*, Nature, **440(7084)**:637-643.
[28] X. Li, M. Wu, C. K. Kwoh and S. K. Ng, 2010 *Computational approaches for detecting protein complexes from protein interaction networks: a survey*, BMC Genomics, 11(suppl 1) S3.
[29] C. C. Friedel and C. Zimmer, 2009 *Identifying the topology of protein complexes from affinity purification assays*, Bioinformatic, **25(16)**:2140–2146.
[30] X. K. Ma and L. Gao, 2011 *Predicting protein complexes in protein interaction networks by using a core-attachment algorithm based on graph communicability*, Information Sciences, Accepted manuscript.
[31] M. Granovetter, 1973 *The strength of weak ties*, American Journal of Sociology, **77(6)**:1360–1380.
[32] P. Csermely, 2004 *Strong links are important, but weak links stabilize them*, Trends Biochem. Sci., **29**:331–334.
[33] L. Lü and T. Zhou, 2010 *Link prediction in weighted networks: the role of weak ties*, Europhys. Lett., **89** 18001.
[34] J. P. Onnelam, et al, 2007 *Structure and tie strengths in mobile communication networks*, Proc. Natl. Acad. Sci., **104(18)**:7332–7336.
[35] X. Cheng, F. Ren, H. Shen, Z. Zhang and T. Zhou, 2010 *Bridgeness: a local index on edge significance in maintaining global connectivity*, J. Stat. Mech., **10** P10011.
[36] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa and S. Kanaya, 2006 *Development and implementation of an algorithm for detection of protein complexes in large interaction networks*, BMC Bioinformatics, **7** 207.
[37] X. Li, C. S. Foo and S. K. Ng, 2007 *Discovering protein complexes in dense reliable neighborhoods of protein interaction networks*, CSB, 157–168.
[38] J. M. McPherson, L. Smith-Lovin, J. Cook, 2001 *Birds of a feather: Homophily in social networks*, Annual Review of Sociology, **27**:415–444.
[39] I. Xenarios, D. W. Rice, E. M. Marcotte and D. Eisenberge, 2001 *DIP: the database of interacting proteins*, Nucl. Acid. Res., **28**:289–291.
[40] H. W. Mewes, et al, 2004 *Mips: analysis and annotation of proteins from whole genomes*, Nucl. Acid. Res., **32**:41-44.
[41] C. von. Mering, et al, 2002 *Comparative assessment of large-scale data sets of protein-protein interactions*, Nature, **417(6887)**:399-403.
[42] H. N. Chuan, W. K. Sung and L. Wong, 2006 *Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions*, Bioinformatics, **22(13)**:1623–1630.
[43] D. Stauffer and A. Aharony, 1994 *Introduction to Percolation Theory*, 2nd (London: CRC Press).
[44] D. Cvetković, P. Rowlinson and S. Simić, 1997 *Eigenspaces of Graphs*, (Cambridge: Cambridge University Press).

978-1-4577-1666-9/11/$26.00 ©2011 IEEE