Not All Protein Complexes Exhibit Dense Structures in *S. cerevisiae* PPI Network

Bolin Chen, Jinhong Shi
Division of Biomedical Engineering
University of Saskatchewan
Saskatoon, Canada
Email: {boc135, jis043}@mail.usask.ca

Fang-Xiang Wu*
Division of Biomedical Engineering
Department of Mechanical Engineering
University of Saskatchewan
Saskatoon, Canada
Corresponding email: faw341@mail.usask.ca

***Abstract*—Various algorithms have been proposed to identify protein complexes from PPI networks, based on the assumption that protein complexes are densely connected subgraphs. In this study, we conclude that most known protein complexes do not exhibit dense structures in *S. cerevisiae* PPI network, but maintain starlike structures in the network. Moreover, vertices of protein complexes are not sparsely connected with the rest components of the network. Many vertices tend to have more outgoing interactions than they have within protein complexes. Based on starlike properties of known protein complexes, we propose a random-star algorithm to identify protein complexes in PPI networks. Predictions are evaluated in terms of the average *f-score*. After excluding similar clusters, we finally obtaine 744 predictions with the average *f-score* at 0.51.**

***Keywords*-protein complex, protein-protein interaction, dense subgraph, starlike structure**

I. Introduction

Protein complexes are essential molecular entities that carry out major cellular processes [1]. They consist of groups of proteins that physically bind together in living cells [2]. Understanding them is an essential step for our attempt towards unraveling the intricate biological systems [3].

Various computational algorithms have been proposed to identify protein complexes according to their topological structures in protein-protein interaction (PPI) networks. The most commonly used assumption is that protein complexes exhibit dense structures in PPI networks. Algorithms, such as the maximal clique algorithm [4], MCODE [5], RNSC [6], DEM [7], LCMA [8], MCL [9], and the graph entropy based algorithm [10], are proposed based on this assumption. The other assumption for protein complexes is the core-attachment structures [2]. Many core-attachment approaches [3][11] are developed to identify the cores and attachments of protein complexes, separately. Although most of those algorithms are efficient and helpful, their accuracy is still limited, which is due to not only the intricate connections of PPI networks, but also the unclear characters of protein complexes.

In this paper, we first investigate properties of protein complexes in a *S. cerevisiae* PPI network of DIP. We find that most protein complexes do not exhibit dense structures in the PPI network, but are sparsely internally connected in terms of both the density and the average degree. We introduce a cyclic-level model to describe the relationship between protein complexes and their surrounding neighbours. Statistic results show that protein complexes have distinct statistic characters, which indicates that they are identifiable in PPI networks. Moreover, we conclude that most protein complexes exhibit starlike structures in the PPI network. Proteins are more likely to have interactions with only one or more hub-proteins within complexes, and most of them tend to have frequently connections with proteins out of complexes. Based on this character, we finally propose a random-star algorithm to identify protein complexes in PPI networks. Numerical experiments are conducted on the PPI network of DIP. Predicted results show that the algorithm can output protein complexes with high accuracy, which is very promising in predicting protein complexes.

II. Materials and Methods

A PPI network can be represented as an undirected graph $G = (V(G), E(G))$, where $V(G)$ is the set of vertices (individual proteins), and $E(G)$ is the set of edges (protein interactions). Let $H = (V(H), E(H))$ be a subgraph of G, the neighbours of H can be defined as

$$N(H) = \{v|(u, v) \in E(G), u \in V(H), v \in V(G) \backslash V(H)\}.$$

Without loss of generality, in this paper we do not distinguish concepts of PPI networks, protein complexes and proteins from graphs, subgraphs and vertices, respectively.

A. Protein complexes and their relative neighbours

We introduce a cyclic-level model to represent the relationship between a protein complex and the rest components of a PPI network. From inside to outside, they are (1) the core level, (2) the inner boundary (IB) level and (3) the outer boundary (OB) level, respectively. To be more precise, let P be a protein complex, then the core level consists of vertices that interact with proteins only in the complex,

$$Core(P) = \{v|v \in V(P), N(v) \subset V(P)\},$$

while the IB level consists of vertices of the complex, but have interactions with proteins out of the complex,

$$IB(P) = \{u|(u, v) \in E(G), u \in V(H), v \in V(G) \backslash V(H)\}.$$

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

The OB level is made up of all proteins that have interactions with proteins in the complex, but are not components of the complex, which is $OB(P) = N(P)$.

The cyclic-level model provides a meticulous way to describe a protein complex in a PPI network. Specifically, edges of a vertex can be divided into three categories, which incident with vertices (1) in the inside level, (2) in the same level and (3) in the outside level, respectively. Then the degree $d(v)$ of a vertex v is decomposed into three kinds of degrees: $d_i(v)$, $d_l(v)$ and $d_o(v)$, which represent the number of each kind of edges, respectively.

B. The number of edges, density, relative density and radius

Given two adjacent levels L_1 and L_2, the set of edges that incident with vertices only in L_1 is denoted by $E(L_1)$, and the set of edges that incident with vertices between L_1 and L_2 is denoted by $E(L_1, L_2)$. Therefore, the number of edges in L_1 and between L_1 and L_2 are

$$m(L_1) = |E(L_1)| \text{ and } m(L_1, L_2) = |E(L_1, L_2)|,$$

respectively.

The density of L_1 can be measured by the commonly used definition

$$Q(L_1) = \frac{2 \cdot m(L_1)}{n(L_1) \cdot (n(L_1) - 1)},$$

where $n(L_1)$ is the number of vertices in L_1. However, when it comes to the density of two adjacent levels, edges in both levels should not be counted. The density is given as

$$Q(L_1, L_2) = \frac{m(L_1, L_2)}{n(L_1) \cdot n(L_2)}.$$

The relative density of two levels or the relative density between a level and two adjacent levels are defined as

$$RQ(L_1|L_2) = \frac{Q(L_1)}{Q(L_2)} \text{ and } RQ(L_1|L_1, L_2) = \frac{Q(L_1)}{Q(L_1, L_2)}.$$

The concept of radius in the cyclic-level model is more important. Suppose each vertex of a subgraph H is assigned an unit area $S = \pi r^2$, where $r = 1$, then the overall area of the subgraph should be $S(H) = \pi \sqrt{n(H) \cdot r}^2$. It gives a quantitative definition about how large a subgraph should cover if the density of a network is equally distributed. For a single level L_1, the radius is defined as $r(L_1) = \sqrt{n(L_1)}$, while for two adjacent levels L_1 and L_2, the radius is defined as $r(L_1, L_2) = \frac{\sqrt{n(L_1)} + \sqrt{n(L_2)}}{2}$.

III. Experiments and Results

A. Data Source

Protein complex data are collected from the database of MIPS [12], CYC2008 [13], YHTP2008 [13], and from the paper of Spirin and Mirny [4]. After removing redundant protein complexes, we finally obtain 2,165 protein complexes as the *gold standard*. There are 870 protein complexes consisting of more than five proteins, and 462 of them consisting of more than ten proteins. However, 563 protein complexes are made up of only two proteins, and another 369 complexes contain only three proteins.

The PPI data are downloaded from the database of interacting proteins (DIP) [14]. The file, named as *Scere20120228.txt*, containes 5004 proteins and 22010 interactions after removing all redundant data, including interactions between *S. cerevisiae* and other species, interactions between the same proteins (loops) and the same interactions between two proteins (multiple edges).

B. Statistic Results

We test known protein complexes data on the PPI network of DIP, and find that almost all vertices of protein complexes have interactions with proteins both in and out of the complexes. Only three protein complexes have core level. Therefore, in most cases the IB level consists of all vertices in a protein complex.

1) The Number of Edges: The number of edges for each level of protein complexes shows distinct properties. Fig 1(a) illustrates $m(IB)$, $m(IB, OB)$ and $m(OB)$ against to their relative radius. We can clearly see from Fig 1(a) that the value of $m(IB)$ and $m(IB, OB)$ rise gradually with the increase of the radius, while the value of $m(OB)$ does not have such significant property. We also test the number of edges for the further outer boundary level of current OB ones. They do not show significant differences from characters of randomly selected subgraphs in the PPI network, which indicates that structures of protein complexes are different from those of randomly selected ones.

2) Density and Relative Density: It is hard to be convinced that protein complexes have dense structures. Fig 1(b) illustrates the density distribution for protein complexes that consist of more than five proteins. The densities for most protein complexes are less than 0.4, and the value of the average density is only 0.31.

However, protein complexes are still relatively dense than their neighbours. Fig 1(c) gives the scatter diagram of relative density $RQ(IB|IB, OB)$ and $RQ(IB|OB)$. We can see from Fig 1(c) that values of most relative densities are large than 1. In fact, the average value of $RQ(IB|IB, OB)$ is 2.45, and the average value of $RQ(IB|OB)$ is 10.54.

Moreover, if values of $Q(IB)$, $Q(IB, OB)$ and $Q(OB)$ are compared vertically, which are plotted against to the value of $r(IB)$, they show clear boundaries for those densities. The scatter diagram is illustrated in Fig 1(d). From top to bottom, the red region is the values of $Q(IB)$, the blue region is the values of $Q(IB, OB)$, and the green region is the values of $Q(OB)$. This character can be used to evaluate whether a predicted result is a protein complex.

3) The average degrees of protein complexes: The vertices of protein complexes tend to have more outgoing interactions than they have within the protein complexes.

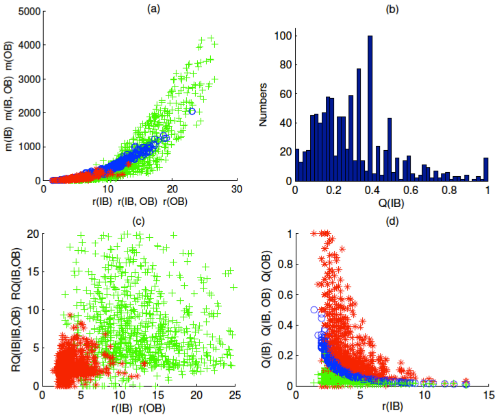

Figure 1. Statistic results for protein complexes with no less than five proteins. The scatter diagrams are plotted according to the value of different statistics and their relative radius. (a) the number of edges; (b) the distribution of density; (c) the scatter diagram of relative density; (d) the scatter diagram of density.

Table 1 summarizes the average degrees of both IB and OB vertices for all protein complexes with more than three vertices, in terms of $\bar{d}_i(v)$, $\bar{d}_l(v)$ and $\bar{d}_o(v)$.

Table 1. The average degree for IB and OB vertices

IB vertices	OB vertices
-	$\bar{d}_i(v) = 1.21$
$\bar{d}_l(v) = 2.30$	$\bar{d}_l(v) = 4.48$
$\bar{d}_o(v) = 18.02$	$\bar{d}_0(v) = 30.28$

It is hard to conclude that protein complexes are dense structures in PPI networks. The average degree of vertices within protein complexes is only 2.30. Considering that the average degree of a minimal connected subgraphs almost equals to 2, the small value of the average degree of vertices within protein complexes indicates that the number of connections within them may only suffice for them to be connected subgraphs.

C. The Structures of Protein Complexes

We conclude that protein complexes exhibit starlike structures in *S. cerevisiae* PPI network. For all known protein complexes of yeast, we draw pictures of them and their relative neighbors. Although some of them are densely connected within themselves (such as the top red regions in Fig 2(a) and Fig 2(b)), there are approximately 70% of them tending to exhibit starlike structures (such as structures in Fig 2(c) and Fig 2(d)). Most protein complexes have one or more hub-proteins, where all other proteins only interact with them within complexes. It is noteworthy that proteins tend to have more outgoing interactions than they have within complexes, no matter whether they are hub-proteins or not.

IV. Algorithm and Results

Based on above statistic characters and starlike structures of known protein complexes, we propose a random-star algorithm to identify protein complexes from PPI networks. Since the overall degree of vertices in protein complexes are usually very large, we consider only those highly connected vertices in the PPI network. All vertices of the network are divided into three categories: (1) core vertices $(d(v) \geq 50)$, (2) important vertices $(3 < d(v) < 50)$ and (3) trivial vertices $(d(v) \leq 3)$. The upper threshold is assigned according to the average maximum degree, which is 49.6 for all known protein complexes. The lower threshold is selected based on the fact that they do not significantly affect the structure of protein complexes. However, they can be changed according to properties of a PPI network.

A. The Random-Star Algorithm

The algorithm is described as follows:

Input: A PPI network G.
Output: A group of starlike clusters.

```
Initialize the random-times $T$ and a threshold $p$.
for $i = 1 : T$ do
    Initialize core vertices list $L_{core}$, important vertices list $L_{impt}$, and trivial vertices list $L_{tvil}$.
    while $L_{core}$ is not empty do
        Randomly select a core vertex $v \in L_{core}$ and let $C = N(v) \backslash L_{tvil}$, $L_{core} = L_{core} \backslash \{v\}$.
        Randomly select a vertex $u \in C$.
        while $u$ is not empty do
            Let $C_1 = N_C(u)$ and $C_2 = C \backslash (C_1 \cup \{u\})$.
            Get a random number $r \sim U(0, 1)$.
            if $r \geq p$ then
                Randomly select a vertex $u \in C_1$, and let $C = C_1$.
            else
                Randomly select a vertex $u \in C_2$, and let $C = C_2$.
            end if
        end while
        Output a cluster.
    end while
end for
```

B. Accuracy Evaluation

We use *f-score* as the measure to evaluate the accuracy of predictions for protein complexes. It is defined as the harmonic mean of $precision = k/n(H)$ and $recall = k/n(P)$, where $n(H)$ and $n(P)$ are the number of proteins in a predicted cluster H and a known protein complex P respectively, and k is the number of proteins they have in common. The *f-score* is defined as

$$f\text{-}score = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot k}{n(H) + n(P)}.$$

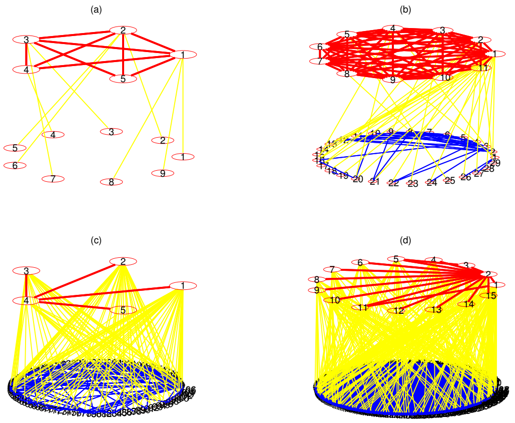

Figure 2. The structure of protein complexes. The top red region of each subgraph represents a protein complex (IB vertices), while the bottom blue region illustrates relative OB vertices. The yellow lines indicate interactions between them. (a) a protein complex with dense inner connections and sparse outer connections; (b) a protein complex with dense inner and sparse outer connections; (c) a protein complex with starlike inner connections and dense outer connections; (d) a larger starlike protein complex.

It is a measure that balances both the true positive predictive rate and the true positive rate.

C. Predicted Results

We vary values of the probability threshold from 0.9 to 0.1, the overall accuracy of prediction increase gradually from only 0.28 to 0.43. The smaller the threshold, the more starlike clusters are generated.

Output clusters can be first excluded according to values of $Q(IB)$, $Q(IB, OB)$ and $Q(OB)$. After fitting boundaries of densities in Fig 1(d), we obtain the upper boundary line and the lower boundary line as

$$f_u = \frac{1.8}{r(IB)} \text{ and } f_l = \frac{0.9}{r(IB)},$$

respectively.

Since we randomly run 100 times for each core proteins (about 200 core proteins in the PPI network), the number of predictions far exceed the number of known protein complexes. After excluding similar predictions according to known protein complexes, we finally obtain 744 predictions of protein complexes, with the average *f-score* at 0.51. It indicates that our proposed random-star algorithm can be a promising method in terms of predicting protein complexes.

V. Conclusion

In this paper, we first analyze statistic properties of protein complexes in *S. cerevisiae* based on the PPI network of DIP. We have concluded that most protein complexes exhibit starlike structures in the PPI network, rather than the dense structures. Moreover, most proteins in those complexes have interactions with proteins out of complexes, and many of them even have more outgoing interactions than they have within complexes.

Based on statistic properties of protein complexes, we propose a random-star algorithm to generate starlike subgraphs in PPI networks. Although it still need to be further improved, the best group of predictions still report protein complexes with high accuracy.

Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., et al., CORUM: the comprehensive resource of mammalian protein complex-2009. *Nucleic Acids Research*, 2010, *38*, D497-D501.

[2] Gavin, A. C., Bösche, M., Krause, R., Grandi, P., et al., Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 2002, *415*, 141-147.

[3] Yu, L., Gao, L., Kong, C., Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics*, 2011, *11*, 3826-3834.

[4] Spirin, V., Mirny, L. A., Protein complexes and functional modules in molecular networks. *PNAS*, 2003, *100*, 12123-12128.

[5] Bader, G. D., Hogue, C. W., An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003, *4:2*.

[6] King, A. D., Przulj, N., Jurisica, I., Protein complex prediction via cost-based clustering. *Bioinformatics*, 2004, *20*, 3013-3020.

[7] Georgii, E., Dietmann, S., Uno, T., Pagel, P., et al., Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 2009, *25*, 933-940.

[8] Li, X. L., Tan, S. H., Foo, C. S., Ng, S. K., Interaction Graph mining for Protein Complexes Using Local Clique Merging. *Genome Informatics*, 2005, *16*, 260-269.

[9] Dongen S. V., Graph Clustering by Flow Simulation. Ph. D. Thesis, University of Utrecht, 2000.

[10] Kenley, E. C., Cho, Y. R., Detecting protein complexes and functional modules from protein interaction networks: A graph entropy approach. *Proteomics*, 2011, *11*, 3825-3844.

[11] Pang C. N., Krycer J. R.. Lek A. et al., Are protein complexes made of cores, modules and attachments?. *Proteomics*, 2008, *8*, 425-434.

[12] Mewes, H. W. , Dietmann, S., Frishman, D., Gregory, R., et al., MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Research*, 2008, *36*, D196-D201.

[13] Pu, S., Wong, J., Turner, B., Cho, E., et al., Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 2009, *37*, 825-831.

[14] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., et al., The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 2004, *32*, D449-D451.