Protein Complex Discovery from Protein Interaction Network with High False-Positive Rate*

Yunku Yeu[1], Jaegyoon Ahn[1], Youngmi Yoon[2], and Sanghyun Park[1]

[1] Dept. of Computer Science, Yonsei University,
3rd Enginnering Bldg. 533-1, Shinchon-dong, Seodaemun-gu Seoul, Korea
[2] Division of Information Technology, Gachon university of Medicine & Science,
1108 Gachon-Kwan, Yonsu-dong, Yonsu-gu, Incheon, Korea
{yyk,ajk}@cs.yonsei.ac.kr, ymyoon@gachon.ac.kr,
sanghyun@cs.yonsei.ac.kr

Abstract. Finding protein complexes and their functions is essential work for understanding biological process. However, one of the difficulties in inferring protein complexes from protein-protein interaction(PPI) network originates from the fact that protein interactions suffer from high false positive rate. We propose a complex finding algorithm which is not strongly dependent on topological traits of the protein interaction network. Our method exploits a new measure, GECSS(Gene Expression Condition Set Similarity) which considers mRNA expression data for a set of PPI. The complexes we found exhibit a higher match with reference complexes than the existing methods. Also we found several novel protein complexes, which are significantly enriched on Gene Ontology database.

Keywords: data mining; machine learning; protein interaction; protein complex.

1 Introduction

Most proteins are known to function within complicated cellular pathways, interacting with other proteins either in pairs or as components of larger complexes [1]. Therefore finding protein complexes is fundamental process for understanding biological functions and cellular organization. Protein complex can be modeled as an undirected graph whose node is a protein, and edge is a physical interaction between two protein nodes. Large scale of PPI information can be abstracted into a PPI network, so finding protein complexes is same as finding subgraph in this global map of protein interactions. Because protein complexes are groups of proteins that interact with others, they are shown as dense subgraphs in PPI network. Several algorithms based on clustering dense region or cliques were proposed to discover protein complexes from PPI networks.

MCODE [2] gives weights to nodes that are densely connected, predicts complexes, and then does postprocess for optimal result. Markov clustering algorithm (MCL) [3] partitions the graph by discriminating strong and weak flows in the graph. DPClus

[*] This work was supported by National Research Foundation of Korea funded by the Korean Government under Grant (No. 2010-0003965).

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2011, LNCS 6623, pp. 177–182, 2011.

[4] finds protein complexes whose densities are more or equal to a threshold value, by expanding clusters starting from seeded vertices.

The algorithms presented above assume that protein complex can be represented as a densely connected subgraph or with several topological structures such as clique or spoke model. That is, these algorithms are strongly dependent on PPI network. Unfortunately, protein interactions are known to suffer from high false positive and negative rates [5], and so protein complexes discovered from that PPI data also suffer from high false rates.

In this article, we propose a complex finding algorithm which is not strongly dependent on topological traits of the PPI network. Our algorithm uses Clustering Coefficient measure to find dense region of the network, and also exploits mRNA expression data to find out the edges of which proteins are similarly transcribed under the same conditions. The mRNA data helps efficient exclusion of false positive protein interactions. We applied our algorithm to the PPI network of *Saccharomyces cerevisiae*, and showed better prediction accuracy compared with existing complex finding algorithms. The number of complexes we found is from about 20 times to 40 times larger than compared algorithms. We also enriched the predicted protein complexes using Gene Ontology database.

2 Method

2.1 Data Description

We downloaded PPI sets of *Saccharomyces cerevisiae* from DIP [6] and BioGRID [7] database. PPI set from DIP (20071007) is manual-curated, and accordingly includes relatively small number of PPIs which are relatively accurate (4,823 proteins and 16,914 interactions). PPI set from BioGRID (3.1.69) includes about 10 times more PPIs than DIP, but likely to contain some false positives (5,920 proteins and 162,378 interactions).

Table 1. Description of reference complexes data. Three numbers of each column represents original data, preprocessed data with DIP, and preprocessed data with BioGRID respectively.

Database	#. complex	#. protein	Avg. size of a complex
MIPS complex	81 / 62 / 77	885 / 492 / 783	12.35 / 9.53 / 11.62
CYC2008	105 / 79 / 105	967 / 609 / 967	10.84 / 9.55 / 10.84

Reference protein complexes of *Saccharomyces cerevisiae* for validating the predicted protein complexes were downloaded from MIPS database [8] and CYC2008 [9]. Each reference complex is a list of proteins without edge information. Because a sparse PPI network (i.e. DIP network) may not contains some edges among member nodes of reference complexes, some of those complexes are divided into 2 or more partitions on PPI network. In this case, we used each sub-component as a reference complex. Then, we excluded reference complexes which contain less than 5 proteins in them. Table 1 shows the information of these reference complexes. We use four cases for reference complex data as follow: MIPS with DIP, MIPS with BioGRID, CYC2008 with DIP, and CYC2008 with BioGRID. Finally we got mRNA expression data from Gasch yeast dataset [10]. It contains 2,994 genes with 173 conditions.

2.2 GECSS(Gene Expression Condition Set Similarity)

We can expect proteins which are transcribed from genes that are co-expressed under same set of conditions are likely to be involved in the same cellular function. And the connected set of PPIs of which condition sets are similar is likely to be also functionally related. In order to measure the similarity between condition sets of PPIs, we firstly find CS (Condition Set) of each PPI, which is defined in Definition 1. After that, GECSS which is defined in Definition 2 can be calculated from CSs.

Definition 1. CS (Condition Set). Let there be two genes g_1 and g_2 which transcribe two proteins of a given PPI *e*. CS_e is defined as condition set where expression levels of both g_1 and g_2 are high or low.

In order to decide whether each expression value is high or low, we apply k-Means clustering to the expression levels on each condition (We set k=3). Once we get CSs of all PPIs, we can calculate similarity between two condition set CS_1 and CS_2 as follows:

$$CS_{sim}(CS_1, CS_2) = \frac{|CS_1 \cap CS_2|}{|CS_1 \cup CS_2|}. \tag{1}$$

The CS_{sim} is 1 if two condition sets are identical, 0 if the two sets are exclusive. Using the definition 1 and the CS_{sim}, we can calculate similarity of a protein set using definition 2.

Definition 2. GECSS (Gene Expression Condition Set Similarity). Given protein set *P*, PPI set $E = \{(x, y) \mid x \in P, y \in P\}$, and CS set $CS_E = \{CS_i \mid CS_i$ corresponds to each $E_i \in E\}$,

$$GECSS = \frac{1}{n} \sum_{CS_1, CS_2 \in CS_E} CS_{sim}(CS_1, CS_2), (CS_1 \neq CS_2). \tag{2}$$

, where *n* is the size of *E*. GECSS has range from 0 to 1, and becomes large if all proteins in *P* have similar mRNA expression patterns. GECSS has the merit that several expression patterns of the protein set can be represented by one single value.

2.3 Algorithm

Finding dense subgraph is similar to the maximal clique enumeration problem which is NP-hard, so we apply some heuristics. The proposed algorithm consists of two main steps: 1) seeds searching, and 2) expanding.

At first, we sort all nodes in the network according to degree in descending order. Next, for each node, all possible node pairs are found which can make a clique with the node. Then, three features of each node pairs are calculated. Those features are a number of common neighbor, averaged Clustering Coefficient (CC), and GECSS. Because these features have different range of possible values, each feature values are standardized by their ranking in each feature. After that, three scores of each pair are summed, and the best scored pair and current node construct the seed for complex finding.

After finding the seed, expanding procedure is called. In this procedure, the seed is expanded with its neighbors considering density and GECSS. At first, a neighbor list is created and sorted according to number of connections to the seed. For all nodes in the list, we calculate the changed value of averaged CC (ΔCC) and GECSS (ΔGECSS) if current node is inserted. Time complexity of calculating averaged CC and GECSS are $O(n^3)$ and $O(e^2 \cdot m)$ respectively (n : the number of nodes, e : the number of edges, m : the number of microarray conditions). But time complexity of adding single node and edge can be reduced to $O(n)$ and $O(em)$, which ensures the practical runtime of the algorithm. If the product of ΔCC and ΔGECSS is greater than 1.0, current node is inserted into candidate. It means if a node makes GECSS decrease greatly because of different expression pattern, it cannot be joined into the candidate even though it could form a dense subgraph.

Our algorithm has only 1 parameter, which is threshold of CC (*CC_threshold*) to determine ΔCC. Range of *CC_threshold* is between 0.0 and 1.0. Higher value of the parameter results more fine-grained clusters and produces large number of clusters.

3 Experimental Results

In order to evaluate the performance of our algorithm, we compared our result with MCODE and MCL. To decide whether a predicted complex is matched with reference complexes, we used affinity score as follows:

$$\text{Affinity}(A,B) = \frac{|A \cap B|^2}{|A| \times |B|}. \tag{3}$$

, where A and B are set of proteins (i.e. one of the predicted complexes or reference complexes). A predicted complex is said to be matched, if there exists reference complex of which affinity score is greater or equal to 0.2.

Finally, we used recall, precision, and F1 score to measure the performance of the algorithm. Each of those measures can be defined as follow:

$$Hit(A,B) = \{A_i \in A | \text{Affinity}(A_i, B_j) \geq 0.2, \exists B_j \in B\}. \tag{4}$$

$$\text{Recall} = \frac{|Hit(R,P)|}{|R|}. \tag{5}$$

$$\text{Precision} = \frac{|Hit(P,R)|}{|P|}. \tag{6}$$

$$\text{F1 score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{7}$$

, where A and B are set of proteins, P is predicted complex set and R is reference complex set. Table 2 shows the results of each algorithm when applying the optimal parameter. The optimal parameters were obtained through iterative experiments on main parameter of each algorithm (*CC_Threshold* in GECSS, *Main inflation value* in MCL, *Node score cutoff* in MCODE). When DIP data was used, our algorithm showed better F1 score than other algorithms. When BioGRID data was used, our algorithm showed similar F1 score with MCL, but the number of complexes is about 40 times more than MCL and 20 times more than MCODE. Low precision and high recall of our algorithm is likely due to the relatively small number of reference complexes. To justify this, we enriched the non-hit complexes ($C - C_{hit}$), with Gene Ontology (GO) database. The enriched complexes can be regarded to be involved with some biological functions.

For four combinations of PPI sets and reference sets, we could get 113, 124, 1780, and 1723 non-hit complexes which do not match with any reference complexes. Those non-hit complexes are enriched by GO using FuncAssociate [11], and 84.7%, 83.2%, 74.7%, and 74.8% of the non-hit complexes were enriched with significance level $\alpha < 0.05$. This means that the predicted complexes have much possibility to be involved with some biological functions, even though they do not match with reference complexes.

Table 2. Experimental results

Dataset (Reference, PPI)	Algorithm (optimal param.)	#. Predicted complex	Recall	Precision	F1 score
MIPS, DIP	GECSS (0.6)	157	0.4677	0.2101	**0.2900**
	MCL (2.0)	253	**0.5161**	0.1304	0.2082
	MCODE (0.1)	39	0.1774	**0.2820**	0.2178
CYC2008, DIP	GECSS (0.6)	157	**0.5063**	0.2802	**0.3608**
	MCL (2.4)	200	**0.5063**	0.2050	0.2918
	MCODE (0.1)	39	0.2278	**0.4871**	0.3104
MIPS, BioGRID	GECSS (0.8)	1,865	**0.7142**	0.0455	0.0856
	MCL (3.6)	53	0.0909	**0.1320**	**0.1076**
	MCODE (0.1)	90	0.0649	0.0555	0.0598
CYC2008, BioGRID	GECSS (0.8)	1,865	**0.7904**	0.0761	**0.1388**
	MCL (3.0)	55	0.0952	**0.2000**	0.1290
	MCODE (0.1)	90	0.0476	0.0555	0.0512

4 Conclusion

The key idea of the presented algorithm is that the members of a protein complex show similar mRNA expression patterns and form dense subgraph in PPI network. These two properties are used to complement each other. By exploiting mRNA expression data, false positive protein interaction can be efficiently excluded. By exploiting network topology, we can exclude false nodes which have similar expression pattern but no relation with the complex. We compared our algorithm with existing methods using known reference complexes and GO terms. The experimental result shows that our method provides better or equivalent accuracy than existing methods, and provides abundance of candidates which are significantly enriched by GO terms.

For future works, we plan to select large-sized, and GO-enriched complexes among abundance of our inferred complexes, and perform case studies in order to validate their biological significance.

References

1. Kumar, A., Snyder, M.: Protein complexes take the bait. Nature 415, 123–124 (2002)
2. Barder, G.D., Hogue, C.W.V.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4(2) (2003)
3. Enright, A.J., Van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7), 1575–1584 (2002)
4. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S.: Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics 7, 207 (2006)
5. Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D.: Protein interactions: Two methods for assessment of the reliability of high throughput observations. Mol. Cell Proteomics 1, 349–356 (2002)
6. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The database of interacting proteins:2004 update. Nucleic Acids Research 32(Database issue), D449–D451 (2004)
7. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a. general repository for interaction datasets. Nucleic Acids Research 34(Database issue), D535–D539 (2006)
8. Güldener, U., et al.: CYGD: the comprehensive yeast genome database. Nucleic Acids Research 33(Database issue), D364–D368 (2005)
9. Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S.: Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research 37(3), 825–831 (2009)
10. Gasch, A.P., et al.: Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell 11(12), 4241–4257 (2000)
11. Berriz, G.F., King, O.D., Bryant, B., Sander, C., Roth, F.P.: Characterizing gene sets with FuncAssociate. Bioinformatics 19(18), 2502–2504 (2003)