A Novel Protein Complex Identification Algorithm Based on the Integration of Local Network Topology and Gene Ontology

Jia-wei LUO[1*]
School of Information Science and Engineering,
Hunan University
Changsha, China
luojiawei@hnu.edu.cn

Xiao-ping WANG[2]
School of Information Science and Engineering,
Hunan University
Changsha, China
wxp8129680@163.com

***Abstract*—The identification of protein complexes is an essential step to understand the principles of cellular organization and biochemical phenomena. A large dataset of experimentally detected protein-protein interactions (PPI) has been determined using high-throughput experimental techniques. However, these datasets usually contain spurious interactions, which complicate the accurate identification of protein complexes by using computational methods. In this study, a novel method is developed to predict protein complexes based on PPI network topology and gene ontology. A protein functional similarity is performed to estimate the reliability of this interaction. A minimum cut-based method is used to detect protein complexes in the weight network. Experimental results show that our method performs better than several efficient, existing clustering algorithms.**

Keywords: Complexes; Topolog; Gene ontology; Weight network; Minimum cut.

I. INTRODUCTION

In the post genomic era, one of the most challenging tasks is the prediction of protein complexes from PPI networks. In cells, protein complexes can perform various functions, such as dynamic signaling, and act as cellular machines, rigid structures, and post-translational modification systems [1]. In general, PPI networks can be represented as undirected graphs G (V, E), in which the node V denotes proteins and the edge E corresponds to interactions among these nodes. Protein complexes generally correspond to dense subgraphs in a PPI network because proteins in the same complex are highly interactive with one another. Many algorithms, such as Cfinder [2], CMC [3], IPCA [4], MCODE [5], and PCP [6], based on the hypothesis have been proposed to discover protein complexes.

Gavin et al. [7] proved that a protein complex comprises two components, core and attachment. Many researchers applied this concept to design detecting protein complex, such as COACH [8], Hunter [9], and Core [10].

Protein interaction data produced by high-throughput experiments are often associated with high false–positive results, which may have a negative effect on the discovery of complex algorithms. For this reason, many computational approaches have been proposed to assess the reliability of high-throughput protein interaction data. Liu et al. [3] proposed the AdjustCD weighting method, which applies an iterative procedure that relies solely on network topology to calculate the reliability of a binary protein interaction. Their experimental results show that the iterative scoring method can effectively reduce the effect of random noise on the performance of a complex prediction method. Lubovac et al. [11,12] introduced two PPI graph weighing schemes to calculate the weight of each interaction in the graph as similarity indicated in the gene ontology (GO) [13] terms of the corresponding proteins and obtained benign outcome. Luo et al. [14] used protein semantic similarity based on GO terms to filter interactions with low GO semantic similarity and added other interactions with high GO semantic similarity for a new reliable network. This new reliable network can effectively reduce the effect of random noise.

In this study, a novel method is developed to identify protein complexes based on topological characteristics and GO in PPI networks. First, a weighted method for each pair of interacting proteins is proposed on the basis of AdjustCD distance [3] and GO semantic similarity. Protein complexes are then detected on the basis of minimum cut in the weighted network. We compare our method with CMC [3], Core [10], PCP [6], Cfinder [2], Coach [8], and IPCA [4] in terms of specificity, sensitivity, and f-score by using two datasets of yeast protein interactions, Krogan [15] and DIP [16]. The experimental results show that our approach is more efficient than these algorithms.

II. METHODS

A. GO semantic similarity

GO [13] is considered as a *de facto* standard for the annotation of gene products. GO is applied to organize data regarding gene function in a directed acyclic graph (DAG) of terms and their relationships. DAG consists of three sub-ontologies, such as molecular function, biological process, and cellular component. One important aspect of GO application is to measure semantic similarity between gene products. The similarity of two gene products based on GO annotations can be considered as the similarity of two sets of GO terms. Considering the definition of term-wise similarity, we can determine the similarity of two proteins annotated by two sets of GO terms. In this study, the average of three semantic similarity values is used as the semantic similarity of the interaction. With an edge (u, v) $\in$ E in the graph G, we assume that a protein u is associated with the following GO terms $\{t_{u1}, ..., t_{ui}\}$, and that a protein

v is associated with the following GO terms $\{t_{v1}, ...,t_{vj}\}$. The similarity between proteins u and v based on gene ontology can be defined as follows:

$$ss(u,v) = 1/3 \times \sum_{D} \max_{i,j} Sim(t_{ui}, t_{vj}), D = \{BP, CC, MF\} \quad (1)$$

where Sim (t_{ui}, t_{vj}) is the similarity between the GO terms t_{ui} and t_{vj} based on the ISM method, an improved measure proposed by Yang et al. [17]. ss (u, v) of each edge e (u, v) is normalized as follows:

$$SS(u,v) = \frac{ss(u,v) - \min_{e(u_i,u_j)\in E}\{ss(u_i,u_j)\}}{\max_{e(u_i,u_j)\in E}\{ss(u_i,u_j)\} - \min_{e(u_i,u_j)\in E}\{ss(u_i,u_j)\}} \quad (2)$$

B. A novel functional similarity based on topology and GO annotation

Several approaches to PPI graph weighting use the graph-theoretic methods, such as CD-distance [3], FSWeight [19], and AdjustCD [3]. These methods were proposed based on the principle that the higher the number of common interactors shared by two proteins, the more likely they are functionally related. The complex detected approach [3], which is employed to predict protein complexes that take advantage of the AdjustCD to obtain the expected outcome. The AdjustCD distance of edge (u,v) is defined as follows:

$$AdjustCD(u,v) = \left\lceil \frac{2 \mid Nu \cap Nv \mid}{(Nu + \lambda u + Nv + \lambda v)} \right\rceil \quad (3)$$

where λu and λv are used to penalize proteins with very few neighbors as in FSWeight [19]. However, these approaches only use the topology of the graph to induce weighting. The GO annotation regarding the molecular function of the proteins and their involvement in biological processes or cellular components likely increases the reliability of PPI, thereby reducing the number of false positives.

Based on this information, a new method is proposed to assess the reliability between two interacting proteins u and v by combining topology and GO semantic. This combination is expressed as follows:

$$FS(u,v) = (AdjustCD(u,v) + SS(u,v))/2 \quad (4)$$

C. Interaction Closeness Between Core and Peripheral

The IPCA algorithm [4] defines the interaction closeness (IC) (v, K) of a vertex v to a subgraph K, which is used to determine the mechanism by which a vertex is closely connected to a subgraph. In this study, IC is generalized to the weight network and defined as follows:

$$IC(v,K) = \frac{\sum FS(u,v)}{\mid V_K \mid}, u \in (N_v \cap V_K) \quad (5)$$

where Nv corresponds to all the direct neighbors of a given vertex v. $|V_K|$ represents the number of vertices in the core K.

D. The Definition of λ-modules

The definition of λ- module has been proposed in a previous study [20] and defined as follows:

Given an undirected graph G (V, E) and a threshold λ, a subgraph $H \subseteq G$ is a λ-module if the following condition is observed:

$$\sum_{v\in H} In\deg ree(H,v) > \lambda * \sum_{v\in H} Out\deg ree(H,v) \quad (6)$$

where λ is a parameter determined by user.

III. OUR PROPOSED ALGORITHM

Our method operates in two main stages. In the first stage, a new weight network is constructed on the basis of AdjustCD and the GO semantic similarity of the PPI network according to Eq. (4). In the second stage, the protein complexes in the network G' = (V, E, w) are detected. First, the protein cores in the network G' = (V, E, w) are identified using a CreateCore algorithm by detecting q-connected graphs with a minimum cut-based method. Then, a noise-filtering method is applied to process the PPI data, in which the iterations with low GO similarity less than a given threshold α are filtered. Some proteins are then selected as attachments to form complexes with the cores by an expanding core algorithm. A merge-complex algorithm is applied to process the complex set, which merges the complexes with a high overlap. The details of Algorithm 1 are presented as follows:

Algorithm1.Detecting the complexes in the weight network;
Input: the PPI network G' = (V, E, w), q, α, T, λ and the IC threshold T;
Output: CP: the set of protein complexes;

1. CG = CreateCore (G',q);//create complex cores from G' //(V, E, w) and reference [9]
2. for each e (vi,vj)∈E
3. If SS (vi; vj) < α, then delete e (vi, vj) from E;
4. end for
5. for each cg∈CG
6. CP = Expanding_core (cg,T,λ); //add the attachment to //form a complex
7. end for
8. CP = merge_complex (CP); //merge some complexes //from CP, references [9]
9. Output the set of protein complexes

The neighborhood N_{cg} of a core cg is denoted as $N_{cg} = \{u \mid u = \bigcup_{c\in cg} N_v\}$, where N_v corresponds to all the direct neighbors of vertex v. For a weighted graph G, the weighted

degree of a vertex v is denoted as $d_w(v)$, which is the sum of the weights of the edges connecting v. For each core cg, $d_w(v)$ of each neighbor v is computed. The neighbor v with a max $d_w(v)$ is selected to merge into the core cg if the IC (v, cg) > T. Once a neighbor v of core cg is merged into the core cg, the core cg is updated. The neighbors of the core cg are then reconstructed. Once none of the neighbors of the core cg can be merged into the core cg, the extension of the core cg is terminated, that is, a complex is formed. On the basis of the definition of λ-module, we discard the complex cg∉λ-module. The expanding core procedure is described in detail in algorithm 2.

Algorithm 2. Expanding core (CG,T,λ)

Input: CG: set of complex cores, T, λ;
Output: CP: set of final complex

```
1.  CP = ϕ; Set Sq = ϕ; // initialization
2.  for each cg in CG
3.    for each v ∈ N (cg)
4.      compute d_w(v) and put v into the Set Sq;
5.    end for
6.    while (! Sq .empty()) :
7.       choose vertex i owning the biggest d_w(v) to in Sq
         and delete it from the Set Sq;
8.      calculate IC(i,cg) according to Eq.(5);
9.      If IC (i,cg) > T, then  add i into cg;
10.       for each m∈N(i):
11.           if m∉ core cg and m∉Sq, then calculate d_w(m)
              and insert m into Set Sq;
12.       end for
13.   end while
14.  If cg∈λ-module, then put cg into complex set CP;
15. end for
16. return CP
```

IV. Results and Discussion

A. Datasets

We apply our method on the two datasets of yeast protein interactions, namely, Krogan [15] and DIP [16]. The Krogan dataset consists of 2,675 proteins and 7,080 interactions among the proteins, and the DIP dataset has 4,930 proteins and 17,201 interactions. To evaluate the predicted complexes, we use two gold-standard sets of protein complexes in our experiments. One set comprises 236 hand-curated complexes, including three or more proteins from CYC2008 [21]. The other gold-standard set of 428 complexes is the union of Aloy [22], MIPS [23], and SGD databases [24] known as "Combined." Both of these datasets are used as benchmark complexes in this study because they have been employed in many computational approaches [25,26].

Based on known protein complexes, an overlapping score [20] OS (x,y) between a predicted protein complex x and a known protein complex y is used to determine the extent by which these two variables match each other. They are considered as matching if OS (x,y) ≥ 0.2 in this study.

B. Comparison with other approaches

We compared our method with six previous competing algorithms, such as CMC [3], Core [10], PCP [6], IPCA [4], Coach [8], and CFinder [2] with recommended parameters. Furthermore, we evaluate the experimental result with four evaluation metrics, namely, specificity (Sp), sensitivity (Sn), f-score, and p-value, which are described in a previous study [20].

Table 1 shows the comparison results between these algorithms and our method in terms of DIP data with two benchmark datasets (Here, we set q = 0.6, α = 0.5, T = 0.12, λ = 0.6 for our method). The Sp and f-score of our method are higher than that of other algorithms with the two benchmark data. The Sn of our method is lower than that of Core, IPCA, and Coach because the number of identified complexes by our method is less than these algorithms. Nevertheless, the overall performance of our method is significantly more efficient than that of other algorithms.

Similarly, our method can achieve the highest f-score on Krogan dataset (Here, we set q=0.75, α=0.5,T=0.08, λ=0.9) as shown in Table 2 by providing the highest Sp and comparable Sn, indicating that our method can predict protein complexes very accurately.

Considering the interactions with higher weight in which AdjustCD and GO terms are combined in the networks, we describe that our method achieves higher Sp and f-score than the other approaches.

Table 1. Results of various algorithms using DIP datasets with two benchmark datasets.

Algorighm	#complex	Average Size	Combined(428)			CYC2008(236)		
			Sp	Sn	f-score	Sp	Sn	f-score
Cfinder (k = 3)	245	10.21	0.34	0.26	0.30	0.29	0.32	0.31
CMC	166	6.75	0.53	0.37	0.44	0.52	0.42	0.46
Core	581	5.10	0.23	0.48	0.31	0.21	0.56	0.30
PCP	313	6.65	0.21	0.28	0.24	0.21	0.35	0.27
Coach	879	7.63	0.34	0.59	0.43	0.30	0.67	0.42
IPCA	1241	7.49	0.38	0.58	0.46	0.35	0.67	0.46
Our method	427	12.80	0.70	0.49	0.58	0.63	0.54	0.58

Table 2. Results of various algorithms using Krogan datasets with two benchmark datasets.

Algorithm	#complex	Average Size	Combined(428)			CYC2008(236)		
			Sp	Sn	f-score	Sp	Sn	f-score
Cfinder (k = 3)	115	10.89	0.53	0.24	0.33	0.56	0.31	0.40
CMC	111	7.91	0.66	0.32	0.43	0.66	0.37	0.48
Core	304	5.27	0.36	0.41	0.39	0.35	0.51	0.41
PCP	242	4.34	0.57	0.36	0.44	0.58	0.44	0.50
Coach	388	7.52	0.55	0.43	0.48	0.53	0.53	0.53
IPCA	396	8.84	0.75	0.36	0.44	0.71	0.36	0.48
Our method	149	12.50	0.80	0.38	0.52	0.77	0.44	0.56

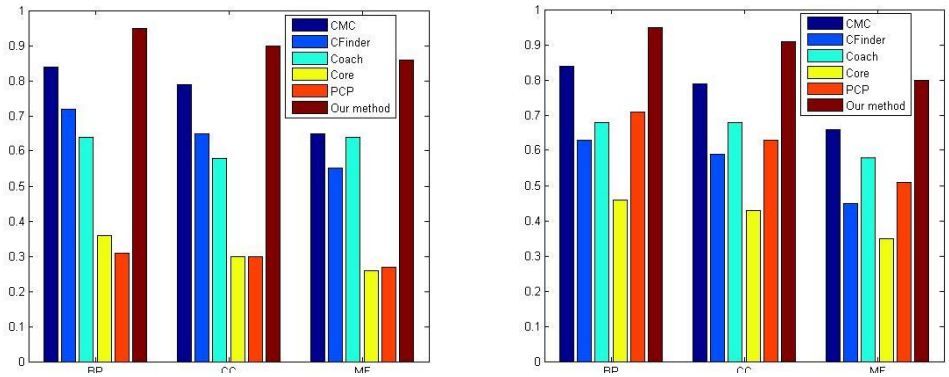

Fig. 1. Accuracy in the three aspects of BP, CC, and MF for various methods. (Left): Accuracy in the three aspects of BP, CC, and MF of various methods on the DIP data; (Right): Accuracy in the three aspects of BP, CC, and MF of various methods on the Krogan data.

To investigate the biological significance of the predicted complexes further, we use p-value based on a hypergeometric distribution and conduct GO enrichment analysis using the tool GO-TermFinder –0.86 [27].

A detected complex with corrected p-value below a cutoff threshold is considered significant (cutoff is set as 10^{-4} in our experiments). We compute the proportion (called accuracy) of significant complexes in all identified ones to evaluate the performance of various algorithms. The accuracy of the six complex detecting methods, such as CMC, Coach, Core, CFinder, PCP, and our method are evaluated separately in three GO categories (i.e., BP, CC, and MF). Fig. 1 shows the comparison results based on this measure using the Krogan and DIP data. In the case of DIP dataset (see the left panel of the Fig.1), the accuracy of our method in the aspect of BP is 95%, which is 11%, 23%, 31%, 59%, and 64% higher than CMC, Coach, CFinder, Core, and PCP, respectively. Our method still performs the best in terms of CC and MF. Furthermore, the CMC method is comparable to the performance of our method, but the PCP and the Core can only predict a small proportion of significant complexes because the PCP and the core extract contain too many small complexes. In general, predicted complexes with small sizes possibly yield large p-values [28]. In the case of Krogan dataset (see the right panel of accuracy of our method outperforms the other algorithms of the DIP dataset in terms of BP, CC, and MF. The p-value analysis shows that our method mines protein complexes with high biological significance.

V. CONCLUSIONS

In this study, a novel method is proposed, in which the topology of a PPI network and GO terms are combined to mine protein complexes. First, a novel weighted strategy based on AdjustCD and GO terms is proposed to construct a new weight network. Second, q-connected graph is mined as the core of protein complex based on minimum cut. Finally, the attachments are added to the core to form a complex based on the IC between the neighbor of the core and the core. The predicted complexes, which do not satisfy the definition of λ-module, are discarded.We applied our method on two different yeast PPI networks with two benchmark complex datasets. Experimental results show that our method could detect complexes more precisely and more meaningfully.

In future work, we will still focus on how to detect complex more accurately combined the other biological knowledge and the topological characteristics.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China(Grant no.61240046) and Hunan Provincial Natural Science Foundation of China (Grant no.13JJ2017).

REFERENCES

[1] N. Zaki, D. Efimov and J. Berengueres. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. Bioinformatics, 14(1), 163, 2013.

[2] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi and T. Vicsek. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics, 22(8), 1021-1023, 2006.

[3] G. Liu, L. Wong, and H.N. Chua, Complex discovery from weighted PPI networks. Bioinformatics, 25(15), 1891-1897, 2009.

[4] M. Li, J. E. Chen, J. X. Wang, B. Hu and G. Chen. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. Bioinformatics, 9(1), 398, 2008.

[5] G. D . Bader, and C.W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. Bioinformatics, 4(1), 2, 2003.

[6] H. N. Chua, K. Ning, W. K. Sung, H.W. Leong and L. Wong. Using indirect protein–protein interactions for protein complex prediction. Journal of Bioinformatics and Computational Biology, 6(03), 435-466, 2008.

[7] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, ... and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084), 631-636, 2006.

[8] X. Ma, and L. Gao. Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability. Information Sciences, 189, 233-254, 2012.

[9] C. H. Chin, S. H. Chen, C. W. Ho, M. T. Ko and C. Y. Lin. A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. Bioinformatics,11(Suppl 1), S25, 2010.

[10] H. C. Leung, Q. Xiang, S. M. Yiu, and F. Y. Chin. Predicting protein complexes from PPI data: a core-attachment approach. Journal of Computational Biology, 16(2), 133-144, 2009.

[11] Z. Lubovac, J. Gamalielsson and B. Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. Proteins: Structure, Function, and Bioinformatics, 64(4), 948-959, 2006.

[12] Z. Lubovac, D. Corne, J. Gamalielsson, and B. Olsson. Weighted cohesiveness for identification of functional modules and their interconnectivity. In Bioinformatics Research and Development (pp. 185-198). Springer Berlin Heidelberg, 2007.

[13] J. A. Blake, J. Corradi, J. T. Eppig, D. P. Hill, J.E. Richardson, and M. Ringwald. Creating the gene ontology resource: design and implementation. Genome Res. 11(8): p. 1425-33, 2001.

[14] J. LUO, and C. LI. A Novel Method to Predict Protein Complexes Based on Gene Ontology in PPI Networks⋆. Journal of Computational Information Systems, 9(12), 5031-5039, 2013.

[15] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, ... and M. Gerstein. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440(7084), 637-643,2006.

[16] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic acids research,30(1), 303-305, 2002.

[17] H. Yang, T. Nepusz, and A. Paccanaro. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics, 28(10), 1383-1389, 2012.

[18] G. Liu, L. Wong and H. N. Chua, Complex discovery from weighted PPI networks. *Bioinformatics*, 25(15), 1891-1897, 2009.

[19] H. N. Chua, S. Wing-Kin and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 22: 1623–1630, 2006.

[20] J. Wang, M. Li, J. Chen and Y, Pan. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 8(3), 607-620, 2011.

[21] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak. Up-to-date catalogues of yeast protein complexes. Nucleic acids research, 37(3), 825-831, 2009.

[22] P. Aloy, B. Böttcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, ... and R.B. Russell. Structure-based assembly of protein complexes in yeast. Science, 303(5666), 2026-2029, 2004.

[23] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, ... and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. Nucleic acids research, 32 (suppl 1), D41-D44, 2004.

[24] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, ... and J. M. Cherry. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic acids research, 30(1), 69-72, 2002.

[25] J. Wang, D. Xie, H. Lin, Z. Yang and Y. Zhang. Identifying Protein Complexes from PPI Networks Using GO Semantic Similarity. In Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on (pp. 582-585), 2011.

[26] 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. Bioinformatics, 7(1), 207, 2006.

[27] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock. GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics, 20(18), 3710-3715, 2004.

[28] I. A. Maraziotis, K. Dimitrakopoulou and A. Bezerianos. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. Bioinformatics, 8(1), 408, 2007.