Integrating Multiple Biomedical Resources for Protein Complex Prediction

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Bo Xu
College of Computer Science and Technology, Dalian University of Technology, Dalian, China 116023
zhyj@dlut.edu.cn

***Abstract*—Prediction of protein complexes from protein-protein interaction (PPI) networks is crucial to unraveling the principles of cellular organization. Most existing approaches only exploit high-throughput experimental PPI data to predict protein complexes. In this paper, we integrate the multiple biomedical resources for protein complex prediction by constructing attributed PPI networks, which include high-throughput data, co-expression data, genomic data, text mining data and gene ontology data. Multiple biomedical resources are complementary in attributed PPI networks. We propose a novel approach called IMBP based on attributed PPI networks. IMBP can effectively learn the degree of contributions of different biomedical resource for complex prediction. The experimental results show that IMBP can make good use of multiple biomedical data and achieve state-of-the-art performance.**

***Keywords*—protein complex prediction; multiple biomedical resources; attributed networks; gene ontolgy**

I. Introduction

In the post-genome era, a key task of systems biology is to cluster proteins, and their interactions into protein complexes which can help us understand certain biological processes and predict the proteins functions [1]. Most proteins seem to be functional only after they are assembled into a protein complex and interact with other proteins in this complex.

High-throughput experimental technologies such as yeast-two-hybrid and mass spectrometry, along with computational predictions, have detected huge amount of protein-protein interaction (PPI) data for numerous organisms [2]. These data can be represented as undirected graphs, in which nodes represent proteins and edges represent interactions between pairs of proteins. There have emerged a series of computational methods to detect protein complexes in PPI networks, including MCODE [3], CMC [4], COACH[5] and ClusterONE [6].

To enhance the quality of predicted complex, additional biomedical resources, such as gene expression data and gene ontology (GO), have been applied to protein complex prediction. For example, Feng et al. [7] used microarray data to weight PPI networks, which could better represent the actual interaction networks than the initial binary PPI networks. Besides, sequenced genomes data and a large amount of biomedical literature also contain valuable information for protein complex prediction. However, few studies have exploited such biomedical resources in detecting protein complex.

In this paper, we seek to integrate multiple biomedical resources to improve protein complex prediction. These resources include high-throughput experimental PPI data, GO, gene expression data, sequenced genomes data and biomedical literature. Firstly, we used these biomedical resources data to construct attributed PPI networks, in which node attributes represented the GO annotations of the protein node and edge attributes represented the type of associations between two protein nodes. The attributed PPI networks integrated from multiple biomedical resources offer two unique advantages: (i) various types of evidence are mapped onto a single, stable set of proteins, thereby facilitating comparative analysis; (ii) different types of associations can partially complement each other, leading to increased coverage. We then proposed IMBP algorithm for protein complex prediction based on attributed PPI networks. IMBP can predict protein complexes based on both the dense topological structure and GO annotation similarity of attributed PPI networks. Particularly, IMBP can distinguish the degree of contributions between the different types of biomedical resources for complex prediction task. Finally, we showed that IMBP was competitive or superior in performance, compared with the state-of-the-art methods.

The rest of paper is organized as follows: In Section II, we describe the IMBP approach in details. We investigate the performance of the IMBP approach on different yeast PPI datasets in Section III. Finally, we conclude and present our future plan in Section IV.

II. Methods

A. Biomedical Resources

Most of studies only make use of PPI data from high-throughput experiments to detect protein complexes. However, these PPI data often have a high false positive rate and an even higher false negative rate. Further improvements for complex prediction can be obtained by integrating other biological evidences with high-throughput PPI data.

GO is currently one of the most comprehensive and well-curated ontology databases in the bioinformatics community. GO provides GO terms and GO slims to describe gene product characteristics. Due to the similar biological properties of protein complex, GO are valuable biological evidences to high-throughput PPI data for protein complex prediction.

Since proteins which interact with each other are likely to exhibit similar gene-expression profiles, gene expression data have been widely exploited to annotate protein functions and predict novel PPI. Similarly, sequenced genomes data can be used to predict novel interactions between proteins based on systematic genome comparisons. In addition, published biomedical literature are another important source of PPI data. In this study, we concentrate on protein complex prediction task. Therefore, we import the PPI data of co-expression type,

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

genomic type and text mining type from STRING which is one of the largest databases of known and predicted protein interactions [8].

B. Construction of Attributed PPI Networks

We define an attributed PPI network as a 6-tuple $G=(V,E,A_v,A_e,F_v,F_e)$ where V is the set of protein vertices, E is the set of PPIs. $A_v=\{GS_1,GS_2,...GS_n\}$ is the set of GO slim attributes for protein vertices, and F_v is a function that returns the set of GO slim attributes of a protein vertex. Each protein vertex p_i in V has a set of GO slim attributes $F_v(p_i)=\{GS_{i1},GS_{i2},...,GS_{im}\}$, where $m=|F_v(p_i)|$ and $F_v(p_i)\subseteq A_v$. Likewise, $A_e=\{T_1,T_2,...T_s\}$ is the set of type attributes for PPIs, and F_e is a function that returns the set of type attributes of a PPI. Each PPI e_i in E has set of type attributes $F_e(e_i)=\{T_{i1},T_{i2},...,T_{ir}\}$, where $r=|F_e(e_i)|$, $F_e(e_i)\neq\emptyset$ and $F_e(e_i)\subseteq A_e$. In this study, the type attributes of PPIs include high-throughput type, co-expression type, genomic type and text mining type, namely, $A_e=\{T_1,T_2,T_3,T_4\}$.

(a)

Protein	GO attributes
p1	GS1
p2	GS1,GS2
p3	GS1,GS2
p4	GS1,GS2
p5	GS1,GS2,GS3
p6	GS1,GS3
p7	GS1
p8	GS1,GS3
p9	GS2,GS3
p10	GS2,GS4

PPI	Type
e1	T1,T4
e2	T2
e3	T1,T2,T4
e4	T1,T3,T4
e5	T1,T2
e6	T1,T2,T3,T4
e7	T2,T3,T4
e8	T1,T3,T4
e9	T1,T4
e10	T3
e11	T1,T2
e12	T1,T3
e13	T3,T4

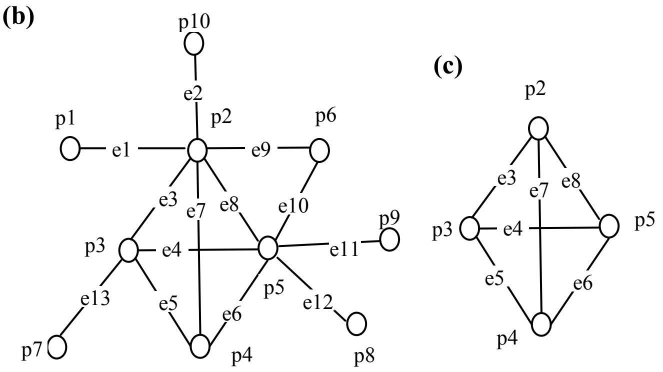

Fig.1. Illustration example of attributed PPI networks: (a) GO slim attributes of protein vertices and type attributes of PPIs. T1: high-throughput type; T2: co-expression type; T3: genomic type; T4: text mining type. (b) An attributed PPI networks. (c) The subgraphs induced by {GS1,GS2}.

Figure 1 shows an example of an attributed PPI network where the GO slim attributes of protein vertices and type attributes of PPIs are given in Fig. 1a. It can be seen that each protein vertex has a GO slim attribute set and each edge has a type attribute set. For instance, p_1 has one GO slim attribute “GS1”, and e_3 has three type attributes including “T1” (high-throughput type),”T2”(co-expression type) and “T4”(text mining type).

Given the set of ontology slim attributes A_v, we define an attribute set S as a subset of A_v $(S\subseteq A_v)$. Moreover, we denote by $V(S)\subseteq V$ the vertex set induced by S (i.e., $V(S)=\{p_i\in V|S\subseteq F_v(p_i)\}$) and by $E(S)\subseteq E$ the edge set induced by S (i.e., $E(S)=\{(p_i,p_j)\in E|p_i,p_j\in V(S)\}$). The subgraph $G(S)$, induced by S, is the pair $(V(S),E(S))$. Fig. 1c is the subgraph induced by the attribute set $\{GS_1,GS_2\}$.

C. Ontology Correlated Cliques

Definition 1 (Ontology correlated clique) Given a protein vertex set C and an edge set E_c in the induced subgraph $G(S)(C\subseteq V(S),E_c=\{(p_i,p_j)\in E(S)|p_i,p_j\in C\})$, an ontology correlated clique is a pair $((C,E_c),S)$, such that for each protein vertex p_i in C, the degree of p_i is $|C|-1$. S is the common ontology attribute set of C.

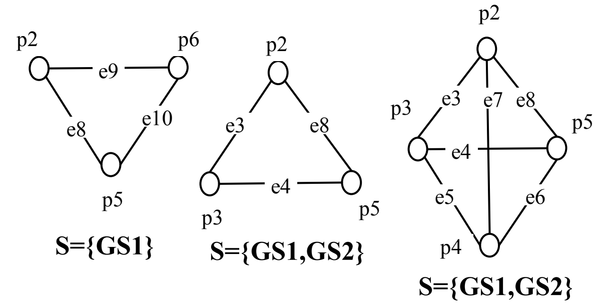

Fig. 2. Illustration examples of ontology correlated cliques.

Given an attributed PPI network, we can mine many Ontology correlated cliques with different common ontology attribute sets. Figure 2 shows three ontology correlated cliques of the attributed PPI networks in Fig. 1.

Definition 2 (Structural correlated function η) Given an ontology slim attribute set S, the structural correlation of S, $\eta(S)$, is given as:

$$\eta(S)=\frac{|K_S|}{|V(S)|} \qquad (1)$$

where K_S is the set of vertices in ontology correlated cliques in $G(S)$. We only considered the cliques whose size is no less than 3.

Structural correlation function was proposed in the recent study [9]. We can use this function to measure the dependence between ontology attribute set S and the density of the associated vertices. In other words, it indicates how likely S is to be part of cliques. Therefore, the larger structural correlation function $\eta(S)$ gets, the more valuable ontology attribute set S is. Given an attributed graph $G=(V,E,A_v,A_e,F_v,F_e)$, the density of G , $density(G)$, is given as:

$$density(G) = \frac{2 \bullet \sum_{e_i \in E} \sum_{T_j \in F_e(e_i)} w_j}{|V| \bullet |V-1|} \qquad (2)$$

where w_j is the weight of the T_j type attribute and $\sum_{T_j \in F_e(e_i)} w_j$ is the weight of the edge e_i.

In general, different type attribute evidences have different importance for protein complex prediction task. To model the degree of contributions of type attribute evidences, we assign a weight to each type attribute. In the equation (2), w_j denotes the weight of contribution of T_j type attribute evidences. Moreover, the edges of attributed networks may have different weights due to their different type attributes.

Then we define the clique score for ontology correlated clique $((C, E_c), S)$ as follows:

$$C\_Score((C,E_c),S) = \eta(S) \bullet |C| \bullet |S| \bullet density(C,E_c) \qquad (3)$$

where S is the common ontology attribute set of C.

D. Weight Mechanism

We design an automatically method to learn the different weight for each type attribute of PPIs.

Firstly, we construct attributed PPI networks with GO slims annotations and each type PPI data in turn. The attributed PPI networks only have one type PPIs, and we set $w_1 = w_2 = w_3 = w_4$ =0.25. Secondly, we use the cliques mining algorithm [10] to enumerate all maximal cliques with size no less than 3 from initial attributed PPI networks constructed with one type PPI data such as high-throughput type, and calculate the ontology attribute set for each maximal clique. All maximal ontology correlated cliques make up the candidate clique set $Candidate$. The maximal ontology correlated cliques generally overlap with each other. Thirdly, we use the same method as CMC [4] to generate the seed clique set $Seed$ where the seed cliques are non-overlapping.

We define the contribution degree of T_i type attribute, $C\_Degree(T_i)$, as follows:

$$C\_Degree(T_i) = \frac{\sum_{((C_i,E_{ci}),S_i) \in Seed_{Ti}} C\_Score((C_i,E_{ci}),S_i) \bullet |E_{ci}|}{|E_{Ti}|} \qquad (4)$$

where $Seed_{Ti}$ denotes the seed clique set generated from the attributed PPI networks constructed with T_i type PPIs and E_{Ti} is edge set of T_i type PPIs. The weight of T_i type attribute, w_i, in the whole attributed PPI networks is

$$w_i = \frac{C\_Degree(T_i)}{\sum_{i=1}^{s} C\_Degree(T_i)} \qquad (5)$$

E. The IMBP Algorithm

The IMBP algorithm broadly consists of two phases. In the first phase, IMBP constructs attributed PPI networks with GO slims annotations and each type PPI data alone, and learns the weight of contribution for each type PPIs in turn. In the second phase, IMBP uses the whole attributed PPI networks and the weights of the PPI attributes to predict the protein complexes. Firstly, IMBP generates the seed clique set $Seed$ from the whole attributed PPI networks using the same method as in the first phase. Then IMBP augments the seed cliques by including the close neighbor protein vertex one by one. IMBP uses the closeness score to measure how closely a protein vertex P_i with ontology attribute set S_i is connected to a seed clique $((C_j, E_{cj}), S_j)$, where $P_i \notin C_j$. The closeness score of P_i with respect to $((C_j, E_{cj}), S_j)$ is defined as follows:

$$P\_Score((P_i,S_i),((C_j,E_{cj}),S_j)) = \frac{|S_i \cap S_j|}{|S_j|+1} \bullet \frac{\sum_{e_l \in E_p} \sum_{T_k \in F_e(e_l)} w_k}{|C_j|} \qquad (6)$$

where E_p is the set of the edges between P_i and $((C_j, E_{cj}), S_j)$. If the $P\_Score((P_i,S_i),(C_j,S_j)) \geq extend\_thres$, then P_i is added to the seed clique $((C_j, E_{cj}), S_j)$. Thus the final predicted complexes will be generated by adding the close neighbor proteins to the seed cliques. After preliminary experiments, we set $extend\_thres = 0.2$.

III. Experimental results

A. Datasets and Evaluation Metrics

The high-throughput PPI dataset used in our experiment is Gavin dataset [11]. The PPI data of co-expression type, genomic type and text mining type are imported from STRING database [8]. Besides, the GO slim data are downloaded from http://www.yeastgenome.org. The benchmark complex dataset is CYC2008 [12].

To keep our evaluation metrics as the same as the most studies, we chose the precision, recall and F-score as the major evaluate measures. In addition, we also reported accuracy (Acc) in our experiments.

In our experiments, we constructed the attributed PPI networks with Gavin high-throughput dataset, STRING data and GO slim data.

B. The Contributions of Different Biomedical Data

In this experiment, we evaluated the contributions of different biomedical data including high-throughput type data, co-expression type data, genomic type data and text mining type data. IMBP can automatically learn the weights for each type attribute of PPI data in order to distinguish the different contributions of each biomedical data for protein complex prediction task. The statistics of the contributions of different biomedical data was listed in Table 1. The text mining type attribute contributed the largest weight 0.43 and the co-expression type attribute contributed the smallest weight 0.09.

Furthermore, we performed IMBP on the attributed PPI networks constructed with each type biomedical data and GO slim data. The experimental results were listed in Table 2. It could be seen that even if the high-throughput data and text mining data were used independently, they achieved high F-score on the attributed PPI networks. We also noticed that the weight mechanism significantly improved the performance of IMBP on the attributed PPI networks. Moreover, the experimental results indicated that IMBP made good use of multiple biomedical resources, and achieve better performance than only used separate biomedical resource.

TABLE 1

The contributions of different biomedical data for protein complex prediction. Extend_thres is set 0.2.

		High-throughput	**Co-expression**	**Genom**	**Text Mining**
Attributed PPI Networks	**PPIs**	6351	12490	103	2086
	Weight	0.22	0.09	0.26	0.43

TABLE 2

The performance comparison on different biomedical data. "Weight" denotes that the IMBP performs on the whole attributed PPI networks using weight mechanism.

	Biomedical data	**P**	**R**	**F**	**Acc**
Attributed PPI Networks	**High-throughput**	0.615	0.262	0.368	0.463
	Co-expression	0.324	0.022	0.041	0.205
	Genom	0.571	0.017	0.033	0.11
	Text Mining	**0.702**	0.248	0.366	0.4
	Weight	0.687	**0.343**	**0.457**	**0.525**

C. Comparison of IMBP with Other Methods

In this experiment, we compared IMBP with the state-of-the-art methods: Cluster ONE [6], COACH [5], CMC [4], and MCODE [3]. The results were listed in Table 3.

As shown in Table 3, IMBP achieved an F-score of 0.457, which was significantly superior to the other methods. Cluster ONE achieved the highest Acc of 0.534. MCODE achieved the highest precision of 0.739 and the lowest recall of 0.154.

Overall, IMBP integrated multiple biomedical resources by constructing attributed PPI networks, and achieved state-of-the-art performance.

TABLE 3

Performance comparison with other approaches. The highest value in each row is in bold. Extend_thres is set 0.2 for IMBP.

PPI Networks	**Methods**	**P**	**R**	**F**	**Acc**
Attributed PPI Networks	**IMBP**	0.687	**0.343**	**0.457**	0.525
Gavin PPI Networks	**Cluster ONE**	0.568	0.331	0.418	**0.534**
	COACH	0.525	0.333	0.406	0.49
	CMC	0.608	0.218	0.321	0.474
	MCODE	**0.739**	0.154	0.255	0.384

IV. Conclusion

We constructed attributed PPI networks with multiple biomedical resources, in which multiple biomedical data were complementary. Then, we developed a clustering algorithm, IMBP, to predict protein complexes on the attributed PPI networks, which was capable of taking into account the local density of networks, as well as the similarity of GO annotations. It was encouraging to see that IMBP achieved the state-of-the-art performance.

We had to point out the state-of-the-art performance of IMBP was based on high quality GO annotation. As a future study, we will study how to exploit other biomedical resource to improve protein prediction, when no high quality GO annotation is available.

Acknowledgment

This work is supported by grant from the Natural Science Foundation of China (No. 61300088, 61272373 and 61070098).

References

[1] M. Li, J. Chen, J. Wang, B. Hu, and G. Chen, "Modifying the DPClus algorithm for identifying protein complexes based on new topological structures," BMC Bioinf., vol. 9, p. 398, 2008.

[2] A.C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, et al.,"Functional organization of the yeast proteome by systematic analysis of protein complexes", Nature, Vol 415(6868), pp.141-7, 2002.

[3] G. Bader and C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinform., vol.4, p. 2, 2003.

[4] G.M. Liu, H.N. Chua and L. Wong, "Complex discovery fromweighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.

[5] M. Wu, X.L. Li, C.K. Kwoh and S.K. Ng, "A Core-Attachment based Method to Detect Protein Complexes in PPI Networks," BMC Bioinform., vol. 10, p. 169, 2009.

[6] T. NEPUSZ, H. YU, A. PACCANARO, "DETECTING OVERLAPPING PROTEIN COMPLEXES IN PROTEIN-PROTEIN INTERACTION NETWORKS," NAT METHODS, VOL. 9, NO. 5, PP. 471-472, 2012.

[7] J. Feng, R. Jiang and T. Jiang, "A Max-Flow Based Approach to the Identification of Protein Complexes Using Protein Interaction and Microarray Data," IEEE Trans. Computational Biology and Bioinformatics, vol. 8, no. 3, pp. 51–62, 2008.

[8] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, et al., " The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Res., vol. 39, PP. 561-8， 2011.

[9] A. Silva, W.J. Meira and M.J. Zaki, "mining attribute-structure correlated patterns in large attributed graphs," Proc. Int'l Conf. on Very Large Databases (VLDB 12'), pp. 466-477, 2012.

[10] E. Tomita, A. Tanaka and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," Theor. Comput. Sci., vol. 363, no. 1, pp. 28–42, 2006.

[11] A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, "Proteome survey reveals modularity of the yeast cell Machinery," Nature, vol. 440, no. 7084, pp. 631–636, 2006.

[12] S. Pu, J. Wong, B. Turner, E. Cho and S.J. Wodak, "Up-to-date catalogues of yeast protein complexes," Nucleic Acids Res, vol. 37, no. 3, pp. 825-831, 2009.