Journal of Theoretical Biology 412 (2017) 107–112

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Combining sequence and Gene Ontology for protein module detection in the Weighted Network

Yang Yu[a,c,*], Jie Liu[a], Nuan Feng[b,**], Bo Song[a], Zeyu Zheng[c]

[a] Software College, Shenyang Normal University, Shenyang 110034, PR China
[b] College of Information Technology, Shenyang Institute of Technology, Fushun 113122, PR China
[c] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, PR China

ARTICLE INFO

Keywords:
Protein complex
Protein interaction
Gene Ontology
The Weighted Network

ABSTRACT

Studies of protein modules in a Protein-Protein Interaction (PPI) network contribute greatly to the understanding of biological mechanisms. With the development of computing science, computational approaches have played an important role in locating protein modules. In this paper, a new approach combining Gene Ontology and amino acid background frequency is introduced to detect the protein modules in the weighted PPI networks. The proposed approach mainly consists of three parts: the feature extraction, the weighted graph construction and the protein complex detection. Firstly, the topology-sequence information is utilized to present the feature of protein complex. Secondly, six types of the weighed graph are constructed by combining PPI network and Gene Ontology information. Lastly, protein complex algorithm is applied to the weighted graph, which locates the clusters based on three conditions, including density, network diameter and the included angle cosine. Experiments have been conducted on two protein complex benchmark sets for yeast and the results show that the approach is more effective compared to five typical algorithms with the performance of f-measure and precision. The combination of protein interaction network with sequence and gene ontology data is helpful to improve the performance and provide a optional method for protein module detection.

1. Introduction

Modularity is a ubiquitous phenomenon in various network systems (Lorenz et al., 2011). A biological network manifests a modular organization and consists of different functional modules. Much of a cell's activity is organized as a network composed of lots of the interacting modules (Chen et al., 2014; Segal et al., 2003), and altering the connections between the different modules may affect changes in cellular properties and functions. A protein module, composed of interdependencies of proteins, is a group of proteins giving rise to the target-specific function whose function is separable from those of other modules (Sharma et al., 2015). Since biologists have found that cellular functions and biochemical events are coordinately carried out by each other in protein modules, and the modular structure of a complex network is critical to functions, identifying such functional modules (or complexes) in PPI networks is of utmost importance as it assists in understanding the structural and functional properties of a biological network and also aids in describing the evolutionary orthology signal. Moreover, it is proposed that a disease is a result of the breakdown of a particular functional module (Barabási et al., 2011), and it has been demonstrated that the modular structure is of great significance in aiding the diagnosis, prevention, and therapy of deadly diseases, especially in cancer research (Segal et al., 2004; Thiagalingam, 2006). Recently a novel concept of modular pharmacology (MP) has emerged (Wang et al., 2012) in pharmacological research. Therefore, based on the above reasons, it is extremely important and necessary to identify functional modules in networks.

In the recent past, a variety of classic clustering approaches, such as density-based clustering (Adamcsek et al., 2006; Altaf-Ul-Amin et al., 2006; Bader and Hogue, 2003), hierarchical clustering (Arnau et al., 2005; Holme et al., 2003; 2010), partition-based clustering, (King et al., 2004; Pfeiffer and Pfeiffer, 2007) and flow simulation-based clustering (Cho et al., 2007; Enright et al., 2002; Pereira-Leal et al., 2002), have been introduced to identify protein complexes from protein interaction data. In recent years, a number of new approaches (Hwang et al., 2008; Inoue et al., 2010; Lecca and Re, 2015; Nepusz et al., 2012; Wu et al., 2009; Yu et al., 2015), utilizing some novel computational models to identify protein modules in a PPI network, has been emerging. Especially, the sources of other biological information have been recently employed to the detection of protein modules

* Corresponding author at: Software College, Shenyang Normal University, Shenyang 110034, PR China.
** Corresponding author.
E-mail addresses: yuyangsd1204@126.com (Y. Yu), 279901222@qq.com (N. Feng).

http://dx.doi.org/10.1016/j.jtbi.2016.10.010
Received 6 January 2016; Received in revised form 25 October 2016; Accepted 26 October 2016
Available online 29 October 2016

in PPI networks (Andreopoulos et al., 2009; Feng et al., 2010; Kouhsar et al., 2016; Lakizadeh et al., 2015; Li et al., 2015; Maraziotis et al., 2007). Though using computational approaches to detect protein functional modules in PPI networks has received considerable attention and researchers have proposed many detection ideas and schemes over the past few years, how to efficiently identify protein modules by means of multiple sources of biological information is still a vital and challenging scientific problem in computational biology.

Based on author's knowledge, there are few methods based on the primary sequence information in the feature selection in the weighted PPI network constructed from the gene ontology information. Thus, in this paper inspired by this observation, we present a novel algorithm called CSeq-GO (Combining Sequence and Gene Ontology for Protein Module Detection) to discover protein complexes from the weighted PPI network. The proposed algorithm consists of mainly three parts: weighted graph construction, feature selection and protein module detection. Moreover, the included angle cosine as the similarity measure is introduced to locate protein complex based on the sequence biological information. The topological properties are based on the fact that proteins are relatively connected densely in the complex (Bader and Hogue, 2003) and protein amino acid background frequency is virtually the axiomatic fact that "sequence specifies structure," which gives rise to an assumption that knowledge of the amino acid sequence might be sufficient to estimate the interacting property between two proteins for a specific biological function. Therefore, the topological and biological features are both of considerable importance for a complex. This algorithm is helpful to capture more biological clusters and experiments conducted on the two public datasets show that the proposed algorithm outperforms five state-of-the-art clustering algorithms in terms of f-measure and precision.

2. Material and methods

In this part, the protein complex detection is described in detail. PPI network can be represented as an undirected weighted graph G =(V, E), where V is the nodes set, corresponding with proteins, E is the set of weighted edges, representing interactions between pairs of proteins. In CSeq-GO, the input is the weighted PPI graph and complex is considered as a subgraph in the whole PPI network, which represents a subset of nodes with a specific set of edges connecting among them.

2.1. The weighted graph construction

It is argued that the detection of protein complexes can greatly be improved by taking into account network weights globally (Nepusz et al., 2012). In this paper, gene ontology is employed to construct the weighted graph. Gene Ontology (GO) is a comprehensive resource across species describing gene and gene product biological properties related to biological process, molecular function, and cellular component. It provides us with promising ways to characterize the functional relationship between pairs of proteins and to infer the interaction between them at functional level (Consortium, 2004; Zhang and Tang, 2016).

The reliability of protein interactions is computed by the definition that qualifies the functional correlation of two proteins using Gene Ontology(GO) annotations based on semantic similarity, which has been used in information science to evaluate the similarity between two concepts in a taxonomy (Resnik, 1995). We use semantic similarity to construct the weighted graph based on the gene ontology and protein interaction information.

In this section, the PPI network is transformed into a weighted graph based on gene ontology information, where the weights are computed by the BMA (best-match) strategy of Lin's method (Lin, 1998) by utilizing the tool of FastSemSim. The attribution to each interaction reflects the degree of confidence and represents the confidence level and the related equations are in (1–3).

$$sim_{MAX}(A, B) = MAX_{t1\in GO(A), t2\in GO(B)}(sim(t_1, t_2)) \quad (1)$$

$$sim_{AVG}(A, B) = AVG_{t1\in GO(A), t2\in GO(B)}(sim(t_1, t_2)) \quad (2)$$

$$sim_{BMA}(A, B) = \frac{AVG_{t_1}(MAX_{t_2} sim(t_1, t_2)) + AVG_{t_2}(MAX_{t_1} sim(t_1, t_2))}{2} \quad (3)$$

2.2. Feature selection

The topological features of this paper mainly include the density and the diameter of the subgraph. The density is used based on the theory that proteins of complex in the internal parts links more closely than the external part. The subgraph diameter is selected based on small world characteristics of the network (Chakrabarti, 2005). Based on my previous study (Yu et al., 2013), the biological characteristics of the background frequency of the amino acids is introduced as the biological characteristic.

(1) Density: Node degree is the sum of the edge weight for a node v. Cluster density is defined in (5).

$$dg_w(w) = \sum_{e=(u,v)\in E} w(e) \quad (4)$$

$$den_w(G) = \frac{2^* \sum_{e\in E} w(e)}{(|V|^*(|V| - 1))} \quad (5)$$

$|V|$ is the number of vertexes and $w(e)$ is the weight of the edge e in a cluster.

(2) Network diameter: Network diameter is the number of links in the shortest path between the furthest pair of nodes of a cluster.

(3) Amino acid background frequency: As for biological properties, amino acid background frequency is proposed and calculated in each subgraph and is defined in (6).

$$freq(C_i) = \frac{sum(C_i)}{\sum_{k=1}^{s} len(p_k)} \quad (6)$$

Where C_i is a kind of amino acid among twenty amino acids, $sum(C_i)$ is the count of this amino acid C_i appearing in a subgraph, $\sum_{k=1}^{s} len(p_k)$ is the sum of each protein amino acid sequence length in a subgraph, s is the size of subgraph.

(4) The included angle cosine (Yu et al., 2011): The included angle cosine value $\cos\theta$ measures the intrinsic similarity between two interaction proteins, which is introduced in our method based on the fact that proteins in the same complex have intrinsic similarity.

$$\cos\theta = \frac{\sum_{m=1}^{n} x_{im} x_{jm}}{\sqrt{\sum_{m=1}^{n} x_{im}^2 \sum_{m=1}^{n} x_{jm}^2}} \qquad \cos\theta \in [0\ 1] \quad (7)$$

where n (n=20) is the size of the vector $V = (x_1, x_2, \ldots, x_n)$ for background frequency, x_{im} and x_{jm} are the m^{th} value of the vector $V_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ from protein i and the vector $V_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ from protein j.

2.3. Algorithm description

Our detection part operates mainly in three stages: seed selection, cluster update and a key stage of update judgment. When a cluster is detected in this stage, the cluster is restricted by $\cos\theta$, density and diameter, which is defined in (8). As we know, the larger the value of cosine is, means the more similarity between proteins. If a node v satisfies the following constraint condition at the same time in (8), v will be added to this cluster (subgraph). Usually, $density \geq \lambda$ and λ is typically set as 0.7 in Refs (King et al., 2004) and diameter≤2 are adopted (Li et al., 2008). The algorithm flow and the description are shown in Fig. 1 and Fig. 2.

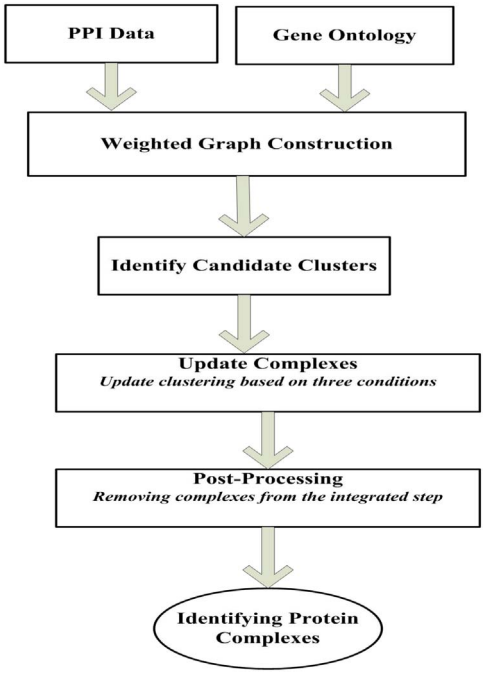

Fig. 1. Representation for Protein Complexes Detection.

Algorithm
Transform: The PPI network is transformed into
Weighted graph Based on Gene Ontology.
Input: The weighted graph G =(V,E).
Output: Identified clusters.
Process:
(1)Sorting degrees of all nodes to queue Q by
descending order.
(2) While Q not empty, do{select the first node v in Q;
Inset v into U}; Go to Update Cluster (U)
Update Cluster (U)
(1) Finding the neighbors of U and sorting
neighbors to N by non-increasing according
to their edge weight;
(2) For each node in N if a node satisfies the condition (8), add the
node to the U; Update Cluster (U);
else continue to select nodes from other
neighbors of U.
(3) If no node to be extended and the current cluster
Can not be further extended,
print the cluster U and add U to the cluster set C; Q = Q−U.
(4)post-processing in set C

Fig. 2. The algorithm for protein complex detection.

$$diameter \le \delta \ and \ density \ge \lambda \ and \ \cos\theta \ge u, \ u \in [01] \tag{8}$$

3. Results

3.1. Datasets

Two reference datasets with size no less than 4 are built from hand-curated complexes, one is from Wodaklab CYC2008 (Pu et al., 2009), the other is from MIPS (Mewes et al., 2008). The PPI network about yeast is from DIP (Xenarios et al., 2002) (the Database of Interacting Proteins) data. Gene ontology data includes an ontology file and a Saccharomyces (Genome Database SGD) of the yeast gene database.

3.2. Evaluation criteria

Given a set of real complexes R and a set of predicted clusters P, the detected clusters are expected to match with the known complexes in the benchmark dataset by the similarity score, which is calculated between a detected cluster and a known complex in (9):

$$\omega = \frac{|C|^2}{|P_i|*|R_j|} \tag{9}$$

Here, $|C|$ is the size of the interaction set between the detected cluster P_i and the known complex R_j, $|P_i|$ is the size of detected cluster and $|R_j|$ is the size of known complex. In (Chen et al., 2014; Wu et al., 2009), a detected cluster is assumed to match a known complex if its overlapping score is at least 0.2, which is also adopted in this study.

For comparison, recall, precision and f-measure are adopted, recall and precision (Li et al., 2010; Lin, 1998; project; Wu et al., 2009) are defined in (10) and (11):

$$recall = \frac{|\{R_j | R_j \in R \wedge \exists \ P_i \in P, \ P_i matches R_j\}|}{|R|} \tag{10}$$

$$precision = \frac{|\{P_i | P_i \in P \wedge \exists \ R_j \in R, \ R_j matches P_i\}|}{|P|} \tag{11}$$

F-measure (Chen et al., 2014; Li et al., 2010), as the harmonic mean of precision and recall, can be used to evaluate the overall performance in (12):

$$f - measure = \frac{2*recall*precison}{recall + precison} \tag{12}$$

3.3. The effect of $\cos\theta$ on clustering

In terms of molecular functions, biological process and cellular location, three types of the weighted graphs are constructed, namely BMA_CC, BMA_BP and BMA_MF. In addition, in order to show the efficiency of our method, three kinds of weighted graphs, BMA/CC, BMA/PP, BMA/MF, are built with the IPI-excluded in GO terms. To understand how the value of $\cos\theta$ influences the outcome of the clustering, we generate 20 sets of clusters by using *density* ≥ 0.7 and diameter$\le$2 with $\cos\theta$ =0.1, 0.2,..., 1.0 from the BMA/BP, *BMABP*, BMA/CC, *BMACC*, BMA/MF and *BMAMF* weighted graph and the effect on recall, with different $\cos\theta$ is given in Fig. 3 and Fig. 4.

Fig. 3 and Fig. 4 shows that the performances of recall are increasing as $\cos\theta$ increases and that there is almost no difference for the recall performance by *density* ≥ 0.7 and diameter$\le$2 when $\cos\theta$ =0.8,0.9,1.0. Moreover, we can observe that the number of the matched known complexes keeps almost the same when $\cos\theta$ =0.8, 0.9, 1.0 and the probability of neighbors added to the cluster is decreasing. More

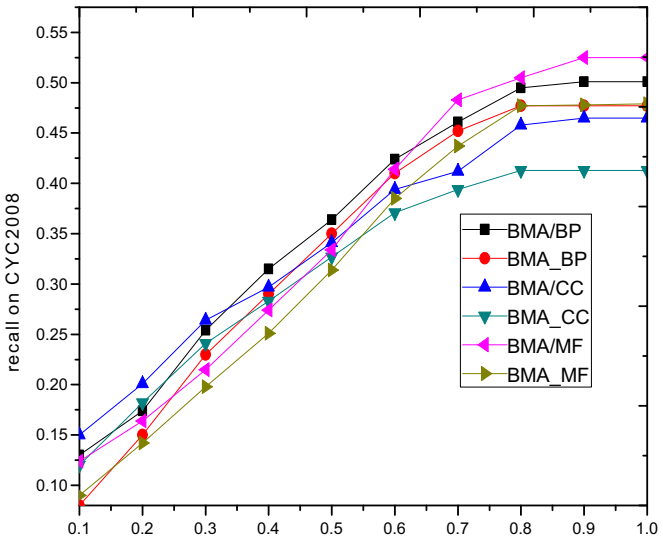

Fig. 3. Recall performance on CYC2008.

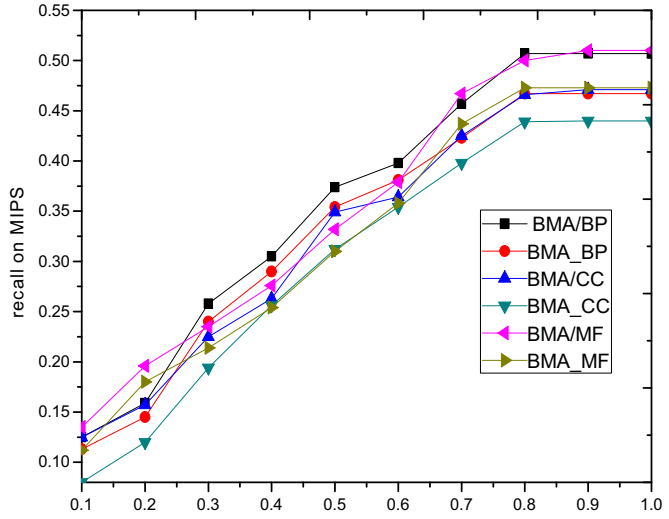

Fig. 4. Recall performance on MIPS.

known complexes are matched by using the weighted graph BMA/BP than by using BMA_BP with the same $\cos\theta$ when it is larger than 0.8. The same trends also exist between BMA/CC and BMA_CC, and between BMA/MF and BMA_MF. In fact, we focus on the actual usefulness of the algorithm in detecting clusters which match real complexes reasonably well. Therefore, the results indicate that our method is helpful to mine the real complexes.

3.4. Evaluation among the methods

Based on the algorithm mentioned above, this part mainly focuses on BMA strategy for the experimental analysis by using *BMACC*, *BMABP*, *BMAMF*. BMA/CC, BMA/PP, and BMA/MF to present the six kinds of weighted graph. In this part, $u = 0.8$ is used for performance comparison. CFinder (Adamcsek et al., 2006), MCODE (Bader and Hogue, 2003), MCL (Dongen, 2001), COACH (Wu et al., 2009) and L (Ahn et al., 2010) are compared with our proposed method and two groups of identification results are tested among in the two complex datasets in Table 1, Fig. 5 and Fig. 6. Taking into account the actual biological meaning for the distribution of the real complexes, the size of the protein complex between 4 and 80 is used. Based on the results from Table 1, Fig. 5 and Fig. 6, three advantages are showed as follows:

Firstly, in Table 1, the performance comparison between the three weighted graphs and the three IPI-excluded weighted graphs shows that the IPI information is moved off and the number of the located cluster in BMA/BP is less than that in BMA_BP. Moreover, we can find that both $|P_m|$ and $|K_m|$ in BMP/BP are more than those in BMP_BP and the same trend can be found when we compare between BMA/CC and BMA_CC, and between BMA/MF and BMA_MF.

Table 1
Performances on benchmark MIPS and CYC2008.

	ON MIPS			ON CYC2008		
	$\|C\|$	$\|P_m\|$	$\|K_m\|$	$\|C\|$	$\|P_m\|$	$\|K_m\|$
BMA/BP	218	80	54	218	115	75
BMA_BP	236	73	52	236	106	69
BMA/CC	207	74	50	207	110	69
BMA_CC	214	68	45	214	102	65
BMA/MF	190	78	54	190	114	94
BMA_MF	210	74	52	210	103	70
CFinder	112	26	28	112	42	41
MCODE	40	14	19	40	16	23
MCL	265	35	39	265	55	60
COACH	639	149	73	639	196	100
L	545	145	67	545	187	91

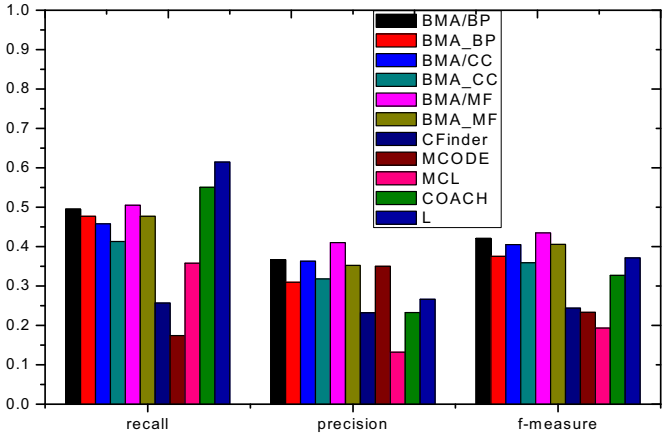

Fig. 5. Comparisons on MIPS.

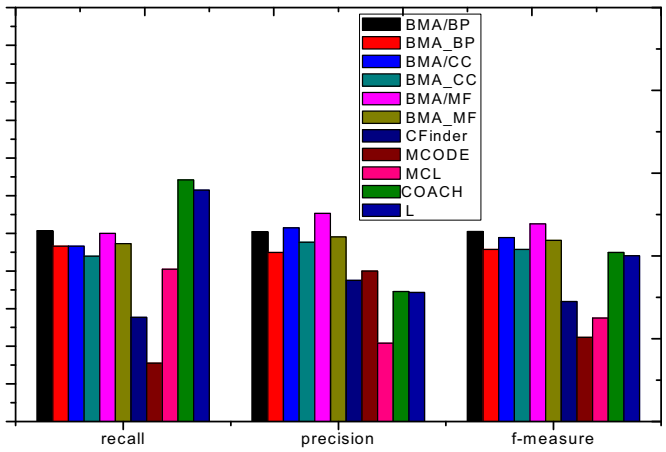

Fig. 6. Comparisons on CYC2008.

Secondly, it can be found that the method based on MF (molecular function) and the method based on BP (biological processes) has similar performances in the aspects of the recall rate. In precision and f-measure metrics: the performance of the weighted graph algorithm based on the molecular function is significantly better than the other two types of the weighted graph methods.

Lastly, our method can match more real protein complexes and has better performances than CFinder, MCODE and MCL, except methods COACH and L, which may be caused that COACH detects protein complex in two stages: core complex and attachment complex, which is different from our method and L merges cliques based upon their topological features, which may contains overlap complexes. It is illustrated in Fig. 5 and in Fig. 6 that our method can achieve mostly better performance on precision, which means that the some noise data maybe be removed by combining sequence and gene ontology information. Comparative results can be seen from the performance of f-measure and show that our method has better overall performance than other methods. This seems reasonable by combining ontology information and sequence information and our method can find more biological protein complexes.

3.5. The robustness of the algorithm CSeq-GO to the different thresholds

In order to show the robustness of the combination of sequence and gene ontology information to identify the protein complexes, performances of f-measure are illustrated among the different methods with nine different threshold $t = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

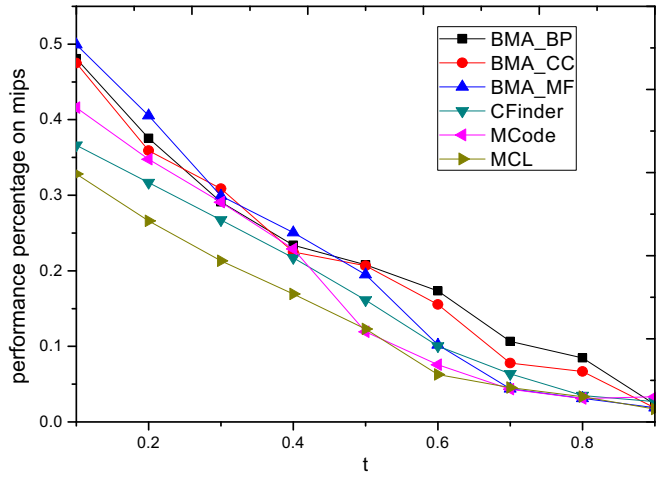

Fig. 7. The f-measure comparisons on MIPS.

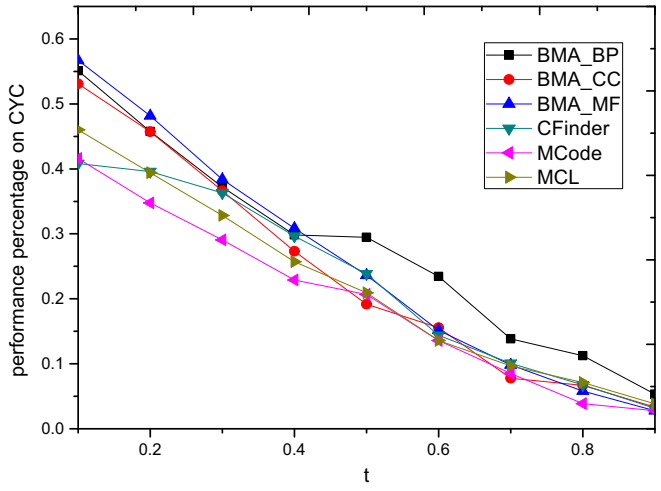

Fig. 8. The f-measure comparisons in CYC2008.

Fig. 7 shows the relative performances of the compared methods on the MIPS benchmark and it illustrates that the six methods follow the same trend: the performance of f-measure decreases with t increasing. The performances of BMA_BP, BMA_CC, BMA_MF are better than CFinder, MCODE and MCL in terms of f-measure with the exception CFinder by t=0.7. Fig. 8 also shows the similar trend as Fig. 7. The results demonstrate the advantages of our method on the improvement of the identification performances. Going by ontology information from protein functional information is helpful and the combination of the two biological features can improve the detection of the clusters efficiently.

3.6. The functional analysis of our method

High functional homogeneity often is exhibited in known protein complexes (Bu et al., 2003; Przulj et al., 2004). To prove the biological functional significance of the detected clusters, p-value is calculated, which represents the probability of co-occurrence of proteins with common function (Maraziotis et al., 2007). P-value is based on hyper geometric distribution and has been used to assign each detected cluster to a main functional group. Low p-values indicate that proteins in the complex do not occur only by chance and the complex has statistical significance. In our paper, p-values are calculated by the tool Go::termFinder (Boyle et al., 2004) which is defined in Eq. (13), where N is the size of the whole network, C is the size of the detected cluster, F is the size of the functional group and k is the number of the proteins of

Table 2
Some detected protein complexes on MIPS PPI data and their p-values.

Gene name	Corrected P-values	Main Functional group
YBR234C YDL029W YIL062C YJR065C YJR091C YKL013C YLR370C YNR035C	5.28e−18	Arp2/3 complex-mediated actin nucleation
YDL145C YER122C YFR051C YGL137W YIL004C YIL076W YLR078C YLR268W YPL010W	2.63e−17	retrograde vesicle-mediated transport, Golgi to ER
YBR087W YJR068W YMR078C YNL290W YOL094C YOR144C YOR217W	9.94e−12	leading strand elongation
YAL021C YCR093W YDL165W YER068W YIL128W YNL288W YNR052C YPR072W	2.76e−13	positive regulation of DNA-templated transcription, elongation
YAR003W YBR175W YBR258C YDR140W YDR469W YHR119W YKL018W YLR015W YPL138C	2.33e−18	protein methylation
Q0085 YBL099W YDR298C YDR377W YJR121W YKL016C YML081C-A YPL078C	4.66e−20	ATP synthesis coupled proton transport

the functional group in the detected clusters. The smallest p-value corresponding to each complex with size ≥7 is shown in Table 2. All of these predicted clusters match well with known functional categories with corrected p-value ≤0.01.

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i}\binom{N-F}{C-i}}{\binom{N}{C}} \tag{13}$$

Our method is different from the traditional clustering algorithms, which mainly employ topological information to partition PPI networks into clusters and fail to consider protein functional and sequence information. In this background, we get insight into the clustering approach by combining amino acid background frequency sequence information and the topological properties in the weighted graph based on the gene ontology information. For instance, some detected clusters shows low P-values and the performance of precision is improved by reducing the noise data of protein interaction data. Hence, it is possible to predict the function of uncharacterized proteins from the prediction of the complexes because proteins in the same complex are likely to share the same function. Meanwhile, based on the theory, it is helpful to infer protein functional type. As shown in Fig. 9, the complex with size 5 is detected by our method, 4 proteins belong to this protein folding type and match the MIPS complex and an uncharacterized protein YMR186W is found. We suggest that YMR186W has a protein folding function because the other proteins in the complex have the same function.

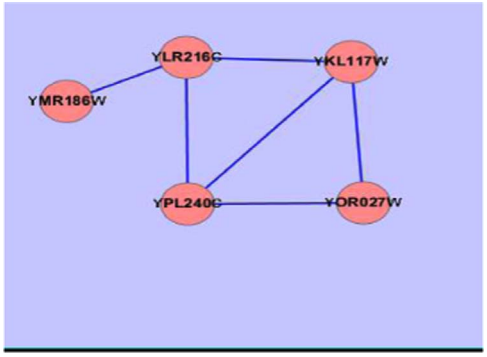

Fig. 9. One predicted example.

4. Conclusion

In this study, a new approach was proposed to detect protein modules in the weighted graph based on the semantic similarity of GO terms. After the term-based similarity measures were used to measure the semantic similarities of protein pairs, six kinds of the weighted graphs were constructed and were applied to the detection algorithm to cluster proteins. Density, diameter and the included angle cosine was adopted as the update conditions and the evaluations were conducted on both yeast complexes datasets. The experimental results show that the combination of the two biological information can generally perform better than the five traditional methods in forms of recall and f-measure, which provides a framework to predict protein complex and also can be helpful to other network community studies. What is more valuable is that this method can be extended to other types of biological information to study the biological network. Future work will also include investigating other biological information, such as structural properties.

Acknowledgments

We wish to thank the authors of the toolkits used in this paper and the reviewers of this paper. This work is supported by Science Research Project of Liaoning Province Education Department (No. L2015496 and No. L201605) and the One-hundred Talent Program of the Chinese Academy of Sciences (Y5AA100A01).

References

Adamcsek, B., Palla, G., Farkas, I.J., I, D., Vicsek, T., 2006. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22, 1021–1023.
Ahn, Y.Y., Bagrow, J.P., Lehmann, S., 2010. Link communities reveal multiscale complexity in networks. Nature 466 (7307), 761–764.
Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S., 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinf. 7, 207–219.
Andreopoulos, B., Winter, C., Labudde, D., Schroeder, M., 2009. Triangle network motifs predict complexes by complementing high-error interactomes with structural information. BMC Bioinforma. 10, 196–215.
Arnau, V., Mars, S., Marin, I., 2005. Iterative cluster analysis of protein interaction data. Bioinformatics 21, 364–378.
Bader, G.D., Hogue, C.W., 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf. 4, 2–28.
Barabási, A.-L., Gulbahce, N., Loscalzo, J., 2011. Network medicine: a Network-based approach to human disease. Nat. Rev. Genet. 12, 56–68.
Boyle, E.I., Weng, S.A., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G., 2004. 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, 3710–3715.
Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., Chen, R., 2003. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Res 31, 2443–2450.
Chakrabarti, D., 2005. Tools For Large Graph Mining (Ph.D thesis). Carnegie Mellon University.
Chen, B., Fan, W., Liu, J., Wu, F.X., 2014. Identifying protein complexes and functional modules–from static PPI networks to dynamic PPI networks. Brief Bioinf. 15, 177–194.
Cho, Y.R., Hwang, W., Ramanathan, M., Zhang, A., 2007. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinf. 8, 1–13.
Consortium, G.O., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261.
Dongen, S.M.V., 2001. Graph Clustering by Flow Simulation (Ph.D thesis). University of Utrecht.
Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.
Feng, J., Jiang, R., Jiang, T., 2010. A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. Comput. Biol. Bioinf. IEEE/ACM Trans. 8 (3), 621–634.
Holme, P., Huss, M., Jeong, H., 2003. Subnetwork hierarchies of biochemical pathways. Bioinformatics 19, 532–538.
Hwang, W., Cho, Y.-R., Zhang, A., Ramanathan, M., 2008. CASCADE: a novel quasi all paths-based network analysis algorithm for clustering biological interactions. BMC Bioinf. 9, 64–77.
Inoue, K., Li, W., Kurata, H., 2010. Diffusion model based spectral clustering for protein-protein interaction networks. PLoS One 5 (e12623), 1–20.
King, A.D., Przulj, N., Jurisica, I., 2004. Protein complex prediction via cost-based clustering. Bioinformatics 20, 3013–3020.
Kouhsar, M., Zaremirakabad, F., Jamali, Y., 2016. WCOACH: protein complex prediction in weighted PPI networks. Genes Genet. Syst. 9 (1), 317–324.
Lakizadeh, A., Jalili, S., Marashi, S.A., 2015. PCD-GED: protein complex detection considering PPI dynamics based on time series gene expression data. J. Theor. Biol. 378, 31–38.
Lecca, P., Re, A., 2015. Detecting modules in biological networks by edge weight clustering and entropy significance. Front. Genet. 6 (265), 1–12.
Li, H., Tong, P., Gallegos, J., Dimmer, E., Cai, G., Molldrem, J.J., Liang, S., 2015. PAND: a distribution to identify functional linkage from networks with preferential attachment property. PLoS One 10 (7), 1–19.
Li, M., Chen, J.E., Wang, J.X., Hu, B., Chen, G., 2008. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinf. 9 (1), 1–16.
Li, X.L., Wu, M., Kwoh, C.K., Ng, S.K., 2010. Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genom. (Suppl 1), S1–S19.
Lin, D., 1998. An information-theoretic definition of similarity. Fifteen Int. Conf. Mach. Learn., 296–304.
Lorenz, D.M., Jeng, A., Deem, M.W., 2011. The emergence of modularity in biological systems. Phys. Life Rev. 8, 129–160.
Maraziotis, I.A., Dimitrakopoulou, K., Bezerianos, A., 2007. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. Bmc Bioinforma. 8 (1), 1–15.
Mewes, H.W., Dietmann, S., Frishman, D., Gregory, R., Mannhaupt, G., Mayer, K.F.X., Munsterkotter, M., Ruepp, A., Spannagl, M., Stuempflen, V., Rattei, T., 2008. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res. 36, 196–201.
Nepusz, T., Yu, H., Paccanaro, A., 2012. Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods 9, 471–472.
Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A., 2002. Detection of functional modules from protein interaction networks. Anal. Biochem. 306, 55–62.
Pfeiffer, M., Pfeiffer, M., 2007. Clustering by passing messages between data points. Science 315, 972–976.
project, F., ⟨https://sourceforge.net/projects/fastsemsim/⟩.
Przulj, N., Wigle, D.A., Jurisica, I., 2004. Functional topology in a network of protein interactions. Bioinformatics 20, 340–348.
Pu, S.Y., Wong, J., Turner, B., Cho, E., Wodak, S.J., 2009. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. 37, 825–831.
Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabasi, A.L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.
Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. Int. Jt. Conf. Artif. Intell., 448–453.
Segal, E., Friedman, N., Koller, D., Regev, A., 2004. A module map showing conditional activity of expression modules in cancer. Nat. Genet. 36, 1090–1098.
Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34, 166–176.
Sharma, P., Ahmed, H.A., Roy, S., Bhattacharyya, D.K., 2015. Unsupervised methods for finding protein complexes from PPI networks. Netw. Model. Anal. Health Inform. Bioinforma. 4, 1–15.
Thiagalingam, S., 2006. A cascade of modules of a network defines cancer progression. Cancer Res. 66, 7379–7385.
Wang, Z., Liu, J., Yu, Y., Chen, Y., Wang, Y., 2012. Modular pharmacology: the next paradigm in drug discovery. Expert Opin. Drug Discov. 7, 667–677.
Wu, M., Li, X.L., Kwoh, C.K., Ng, S.K., 2009. A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinf. 10 (1), 1–16.
Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30, 303–305.
Yu, F.Y., Yang, Z.H., Hu, X.H., Sun, Y.Y., Lin, H.F., Wang, J., 2015. Protein complex detection in PPI networks based on data integration and supervised learning method. BMC Bioinf. 16 (Suppl 12), S1–S9.
Yu, Y., Lin, L., Sun, C., Wang, X., Wang, X., 2011. Complex Detection Based on Integrated Properties. Neural Information Processing. Springer Berlin Heidelberg, 121–128.
Yu, Y., Wang, X., Lin, L., Sun, C., Wang, X., 2013. A supervised approach to detect protein complex by combining biological and topological properties. Int. J. Data Min. Bioinforma. 8, 105–121.
Zhang, S.B., Tang, Q.R., 2016. Protein-Protein interaction inference based on semantic similarity of Gene Ontology terms. J. Theor. Biol. 401, 30–37.