Genes Genet. Syst. (2015) **90**, p. 317–324

WCOACH: Protein complex prediction in weighted PPI networks

Morteza Kouhsar[1], Fatemeh Zare-Mirakabad[2*] and Yousef Jamali[3,4]

[1]*Department of Computer Science, School of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran*
[2]*Faculty of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran*
[3]*Department of Applied Mathematics, School of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran*
[4]*Computational Physics Research Laboratory, School of Nano-Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

(Received 3 May 2015, accepted 8 August 2015; J-STAGE Advance published date: 15 January 2015)

Protein complexes are aggregates of protein molecules that play important roles in biological processes. Detecting protein complexes from protein-protein interaction (PPI) networks is one of the most challenging problems in computational biology, and many computational methods have been developed to solve this problem. Generally, these methods yield high false positive rates. In this article, a semantic similarity measure between proteins, based on Gene Ontology (GO) structure, is applied to weigh PPI networks. Consequently, one of the well-known methods, COACH, has been improved to be compatible with weighted PPI networks for protein complex detection. The new method, WCOACH, is compared to the COACH, ClusterOne, IPCA, CORE, OH-PIN, HC-PIN and MCODE methods on several PPI networks such as DIP, Krogan, Gavin 2002 and MIPS. WCOACH can be applied as a fast and high-performance algorithm to predict protein complexes in weighted PPI networks. All data and programs are freely available at http://bioinformatics.aut.ac.ir/wcoach.

Key words: Gene Ontology, protein, protein complex, protein interaction network, semantic similarity

INTRODUCTION

Proteins are very important molecules that physically interact with each other during their participation in biological processes (Srihari and Leong, 2013). Many experimental techniques such as yeast two-hybrid assays (Ito et al., 2001) and mass spectrometry (Gavin et al., 2002) have been developed to discover pairwise protein-protein interactions and construct protein-protein interaction (PPI) networks. Although these methods identify if there is an interaction between two proteins, they cannot detect interactions involving more than two protein partners (Li et al., 2010; Moschopoulos et al., 2011). To understand a biological function, it is useful to determine protein complexes, groups of interacting proteins that participate in the process (Pizzuti and Rombo, 2014). However, protein complex detection is one of the most challenging problems in PPI network studies, and computational approaches are an appropriate complement for experimental methods to solve this problem. A protein complex contains multiple proteins which are simultaneously linked by non-covalent protein–protein interactions at the same location to perform a functional process in the cell, such as the transcription of DNA, the translation of mRNA, signal transduction and the cell cycle (Li et al., 2010; Moschopoulos et al., 2011).

Generally, PPI networks are modeled as a graph wherein each vertex and edge represents a protein and an interaction between two proteins, respectively (Srihari and Leong, 2013). There are several computational methods to predict protein complexes in PPI networks. The MCL (Enright et al., 2002) method clusters PPI networks by simulating random walks. Some methods such as MCODE (Bader and Hogue, 2003) predict dense subgraphs as protein complexes. The restricted neighborhood search clustering algorithm (RNSC) (King et al., 2004) is a cost-based algorithm that searches for clusters with minimum cost in the network. The LCMA (Li et al., 2005), CFinder (Adamcsek et al., 2006) and CMC (Liu et al., 2009) algorithms detect protein complexes based on clique merging. The CFA (Habibi et al., 2010) algorithm predicts k-connected sub-graphs. GA-PPI (Pizzuti and

Edited by Ali Masoudi-Nejad
* Corresponding author. E-mail: f.zare@aut.ac.ir
DOI: http://doi.org/10.1266/ggs.15-00032

Rombo, 2012, 2013) applies a genetc algorithm with different topology-based fitness functions to cluster PPI networks. The integrative hierarchical clustering algorithm (Wu et al., 2013) integrates multiple heterogeneous data such as Gene Ontology (GO) terms, gene expression profiles, tandem affinity purification with mass spectrometry data and PPI data to detect protein complexes in a hierarchical clustering approach.

Based on studies in yeast (Dezső et al., 2003; Gavin et al., 2006), each complex is composed of a core and its attachments. The core is the central functional unit of a complex. Attachment proteins are bound to the core to complete its function (Srihari and Leong, 2013). In the core, proteins have relatively more interactions with each other. Accordingly, such methods as CORE (Leung et al., 2009) and COACH (Wu et al., 2009) predict cores and attachments in two separate steps.

Recently, researchers have tried to improve the accuracy of complex prediction methods by integrating biological information with PPI data. Feng et al. (2011) and Chen and Yuan (2006) applied microarray data, while Ozawa et al. (2010) and Ma et al. (2012) combined domain-domain interaction information with PPI data.

In the last few years, GO annotations have also been applied to improve the accuracy of complex prediction (Mukhopadhyay et al., 2012; Wang et al., 2012b). The GO database is composed of GO terms and their relationships. GO terms are grouped into three domains: biological processes (BP), molecular functions (MF) and cellular components (CC) (Zhang, 2009; Wang et al., 2012b). Each gene or protein is annotated to one or more GO term(s) and each term is related to one or more other terms in a directed acyclic graph (DAG) structure. According to the experimental observation that proteins in a complex tend to carry out common biological functions (Dezső et al., 2003), some methods use GO annotation to measure the reliability of interaction between each pair of proteins. In these methods, low-reliability interactions are first removed to predict complexes in the new unweighted purified network (Wang et al., 2012b). Although PPI networks are purified by GO information, these data are not considered for complex detection in the algorithm. In other words, an unweighted network is given as an input to the algorithm.

In this article, the semantic similarity between protein pairs is calculated by the Lin method (GraSM) to weigh the PPI networks (Couto et al., 2007). We develop the WCOACH (Weighted COACH) method based on the COACH algorithm and predict protein complexes in weighted PPI networks. The proposed algorithm and such well-known algorithms as COACH, CORE, ClusterOne (Nepusz et al., 2012), MCODE, HC-PIN (Wang et al., 2011), OH-PIN (Wang et al., 2012a) and IPCA (Li et al., 2008) are then run on five PPI networks of yeast including DIP (Xenarios et al., 2002), Krogan (Krogan et al., 2006), MIPS (Mewes et al., 2006), Gavin 2002 (Gavin et al., 2002) and Gavin 2006 (Gavin et al., 2006). The predicted complexes are compared to the experimentally detected complexes of CYC2008 (Pu et al., 2009). The results show that WCOACH can predict protein complexes with high performance.

TERMINOLOGY

A weighted PPI network is shown as a weighted graph G = (V, E, W) where *V*, *E* and *W* denote the set of vertices (proteins), edges (interactions), and weight function $W:E\rightarrow R$ mapping an edge to a real number. The weight of each interaction is a criterion to show its reliability. A neighborhood graph of a vertex *v* is defined as $G_v = \left(V^{G_v}, E^{G_v}, W^{G_v}\right)$ where:

$$V^{G_v} = \left\{u \in V \mid (u,v) \in E\right\},$$
$$E^{G_v} = \left\{(u,u') \in E \mid u,u' \in V^{G_v}\right\},$$
$$\forall (u,u') \in E^{G_v}, W^{G_v}(u,u') = W(u,u').$$

The degree of vertex *v*, the average degree and a set of core vertices of graph *G* are defined, respectively, as follows (Yu et al., 2012):

$$d(v) = \sum_{(u,v)\in E} W(u,v),\ \mu(G) = \frac{\sum_{v\in V} d(v)}{|V|},$$
$$\vartheta(G) = \left\{u \in V \mid d(u) > \mu(G)\right\}.$$

The core graph of vertex *v*, $C_v = \left(V^{C_v}, E^{C_v}, W^{C_v}\right)$, is a subgraph of neighborhood graph G_v where:

$$V^{C_v} = \vartheta(G_v),\ E^{C_v} = \left\{(u,u') \in E^{G_v} \mid u,u' \in V^{C_v}\right\},$$
$$\forall (u,u') \in E^{C_v},\ W^{C_v}(u,u') = W^{G_v}(u,u').$$

The density of graph *G* is defined as (Yu et al., 2012):

$$\delta(G) = \frac{2*\sum_{(u,v)\in E} W(u,v)}{|V|(|V|-1)}.$$

The neighborhood affinity between two graphs *G* and *G*' is defined as (Wu et al., 2009):

$$\alpha(G,G') = \frac{|V \cap V'|}{|V|*|V'|},$$

where *V* and *V*' are the vertices sets of graphs *G* and *G*', respectively.

METHOD

Although COACH (COre-AttaCHment based method) is a well-known algorithm for protein complex prediction in unweighted PPI networks, it predicts cores and attachments as complexes without considering any difference between interactions with high or low semantic similarities. Therefore, considering weighted graphs in the COACH method can increasingly enhance the accuracy of complex prediction. In this paper a novel algorithm, called WCOACH, is represented to support weighted networks for complex prediction. The main steps of the proposed algorithm are as follows:

1. Weigh a PPI network based on GO.
2. Detect preliminary cores.
3. Remove redundant cores.
4. Add attachment proteins to each core for constructing a predicted complex.

The details of the algorithm are explained in the next subsections.

Calculation of semantic similarity between protein pairs The GO database, created from GO terms, forms a DAG based on their relationships to each other. In this paper, we considered three kinds of relationships between terms, called "is-a", "part-of" and "regulates", representing specific-to-general, part-to-whole and regulation relationships, respectively. The semantic similarity between terms is defined based on DAG structure (Zhang, 2009).

Several methods have been developed to measure the semantic similarity between GO terms. Structure-based methods (Wu and Palmer, 1994; Leacock and Chodorow, 1998) measure the semantic similarity based on path length and common parentage. Information content-based methods such as those of Lin (1998) and Resnik (1995) measure the semantic similarity based on a priori probabilities or information content of GO terms (Zhang, 2009). In this article, the *csbl.go* package (http://csbi.ltdk.helsinki.fi/csbl.go) is employed to measure the semantic similarity between protein pairs based on the Lin (GraSM) method. It is clear that the interaction between each pair of proteins unannotated in GO is not weighed. Therefore, the Acceptance-Rejection method is applied to generate values based on weight distribution of other interactions. These values are considered as the weight of unannotated interactions.

We downloaded the GO annotation file from http://www.yeastgenome.org/download-data/curation with version 2, dated 01/18/2014.

Protein complex prediction Based on the COACH method, a new method called WCOACH is developed to predict protein complexes in weighted PPI networks. A weighted graph $G = (V, E, W)$ is given as an input to the algorithm. The set Φ is defined as follows:

$$\Phi = \left\{ \left(C_v, T_v \right) \middle| v \in V \right\},$$

where $C_v = \left(V^{C_v}, E^{C_v}, W^{C_v} \right)$ is a core graph of vertex v (see TERMINOLOGY) and T_v is a set of vertices. The set Φ is given as an input to Algorithm 1 and the preliminary cores set $\prod$ is generated as an output. In this algorithm, $\delta_{av}(C_v)$ is defined as:

$$\delta_{av}\left(C_v\right) = \frac{2 * \left| E^{C_v} \right| * W_{av}}{\left| V^{C_v} \right| \left(\left| V^{C_v} \right| - 1 \right)},$$

where W_{av} is the average weight of all edges in the input network.

Algorithm 1: Predict preliminary cores

Inputs: The set Φ.

Output: The set of preliminary cores Π

1. $\Pi = \varnothing$, $T_v = \varnothing$.
2. **For each** $(C_v, T_v) \in \Phi$,
3. **If** $(\delta(C_v) \geq \delta_{av}(C_v))$
4. Each vertex of T_v is added to C_v and then C_v is appended to Π.
5. **Else**
6. Sub-graph $C_v = \left(V^{C_v}, E^{C_v} \right)$ is decomposed to C_1, C_2, …, C_t by removing the vertices of $\vartheta(C_v)$ from V^{C_v}.
8. **For each** $1 \leq i \leq t$
9. $T_v = \vartheta(C_v)$.
10. (C_i, T_v) is added to Φ.
11. **Return**(Π)

To make the set of final cores Γ from Π the "Redundancy-filtering" procedure of the COACH method is utilized (see Algorithm 2), wherein τ is a threshold to control the overlap between predicted cores (see section 4.3).

Algorithm 2: Remove redundant cores

Input: The set of preliminary cores Π

The neighborhood affinity threshold τ.

Output: The set of final cores Γ

1. $\Gamma = \varnothing$
2. **For each** core graph $C \in \prod$
3. $C' = \underset{C' \in \Gamma}{argmax}\, \alpha(C, C')$.
4. **If** $(\alpha(C,C') < \tau)$
5. add C to Γ.
6. **Else**
7. **If** $\left(\delta(C) * \left| V^C \right| \geq \delta(C') * \left| V^{C'} \right| \right)$
8. Replace C' with C in set Γ.
9. **Return**(Γ)

In the COACH method, a protein is considered as an attachment if it interacts with more than half of the proteins in the core. In real complexes (Gavin et al., 2006), many attachment proteins interact with fewer than half of the proteins in the core. The COACH method cannot detect these attachments. For example, as shown in Fig. 1, the core consists of five proteins in the Coatomer COPII complex. There are six attachments in this complex; however, only one of them interacts with more than two proteins of the core. Our goal is to improve the COACH method for detecting these attachments through defining weighted closeness.

The attachments set of core $C = (V^C, E^C, W^C)$ is defined as:

$$A(C) = \left\{ v \in N(C) \mid \exists u \in V^C, W(u,v) \geq W_{av}^C \right\},$$

where W_{av}^C is the average weight of all edges in C and $N(C)$ (the neighbors set of C) is defined as follows:

$$N(C) = \bigcup_{v \in V^C} \left\{ u \in V \mid (u,v) \in E \right\}.$$

Finally, the set of complexes predicted by our method is defined as follows:

$$\wp = \left\{ C \cup A(C) \mid C \in \Gamma \right\}.$$

RESULTS AND DISCUSSION

In this section, we introduce the PPI networks applied to test our method. Some evaluation criteria are then represented to compare the WCOACH method to the other methods. After that, we show how to compute the neighborhood affinity threshold (τ). Finally, we compare our method to the others on the PPI networks based on defined evaluation criteria.

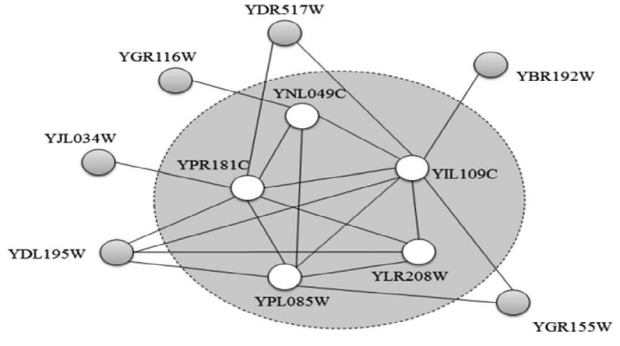

Fig. 1. Coatomer COPII complex. Proteins shown in the gray circle form the core and other proteins are attachments. The COACH method cannot detect YJL034W, YGR116W, YDR517W, YBR192W or YGR155W proteins as attachments because they interact with fewer than half of the core proteins.

Datasets In this article, five PPI networks of yeast are applied to compare our method with others: DIP (Xenarios et al., 2002), Krogan (Krogan et al., 2006), MIPS (Mewes et al., 2006), Gavin 2002 (Gavin et al., 2002) and Gavin 2006 (Gavin et al., 2006). The properties of these networks are shown in Table 1. The real complexes dataset CYC2008, with 428 protein complexes, is used as a benchmark for evaluation (Pu et al., 2009).

Evaluation criteria To compare WCOACH with the other methods, precision, recall, F-measure, coverage rate and P-value are introduced as evaluation criteria.

Let $\wp = \{P_1, P_2, \ldots, P_k\}$ and $\Re = \{R_1, R_2, \ldots R_m\}$ be two sets of predicted and real complexes, respectively. Precision and recall are represented as follows (Ahn et al., 2013; Zaki et al., 2013):

$$Precision = \frac{|\{P \mid P \in \wp, \exists R \in \Re, \alpha(R,P) \geq \theta\}|}{|\wp|},$$

$$Recall = \frac{|\{R \mid R \in \Re, \exists P \in \wp, \alpha(R,P) \geq \theta\}|}{|\Re|},$$

where θ shows an overlapping threshold between predicted and real complexes. By default, θ is considered 0.5 (Habibi et al., 2010). Based on precision and recall, the F-measure is defined as (Chiam and Cho, 2012; Ma and Gao, 2012):

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}.$$

For the predicted and real complex sets $\wp = \{P_1, P_2, \ldots, P_k\}$ and $\Re = \{R_1, R_2, \ldots R_m\}$, coverage rate is defined as follows (Ma and Gao, 2012):

$$CR = \frac{\sum_{i=1}^{m} \max_j \{T_{ij}\}}{\sum_{i=1}^{m} N_i},$$

where T_{ij} is the number of common proteins between the ith real complex and the jth predicted complex and N_i is the number of proteins in the ith real complex.

The biological significance of the predicted complexes is evaluated by P-value. Given a predicted complex P =

Table 1. PPI network properties

Network name	Number of proteins	Number of interactions
DIP	4930	17201
Krogan	2675	7084
Gavin 2002	1352	3210
Gavin 2006	1430	6531
MIPS	4564	15175

(V^P, E^P), that includes k proteins from a functional group (such as a GO term) with size m, the P-value of P is defined as (Ma and Gao, 2012):

$$P-value(P)=1-\sum_{i=0}^{k-1}\frac{\binom{m}{i}\binom{N-m}{|V^P|-i}}{\binom{N}{|V^P|}},$$

where N denotes the total number of proteins in the network.

The P-value is the probability that a given set of proteins is enriched by a given functional group merely by chance (Ma and Gao, 2012). Accordingly, a low P-value shows high statistical significance. Here, the *GO Term Finder* tool (Boyle et al., 2004), available on http://go.princeton.edu/cgi-bin/GOTermFinder, is used to calculate P-values of predicted protein complexes.

Neighborhood affinity threshold The neighborhood affinity threshold ($\tau \in [0,1]$), used in Algorithm 2, controls the overlap between predicted cores. A zero (one) value of τ shows there is no overlap (there is overlap) between cores. To find the appropriate threshold, WCOACH is run with various values of τ on the Gavin 2006 network. Fig. 2 shows that an increase in τ causes higher F-measure and coverage rate values. According to this experiment, τ is set to 0.85 as default.

Comparison with other methods In this section, the method is compared to the COACH (Wu et al., 2009), ClusterOne (Nepusz et al., 2012), CORE (Leung et al., 2009), MCODE (Bader and Hogue, 2003), HC-PIN (Wang et al., 2011), OH-PIN (Wang et al., 2012a) and IPCA (Li et al., 2008) algorithms. In this experiment, all com-

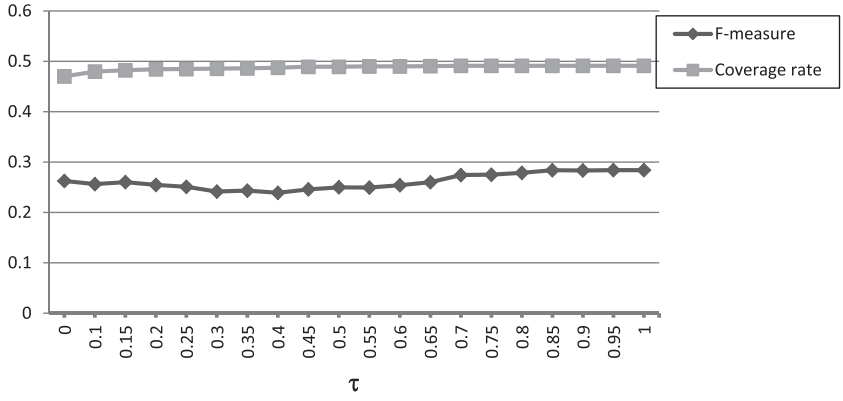

Fig. 2. The F-measure and coverage rate values of WCOACH for various values of neighborhood affinity threshold (τ) on the Gavin 2006 network.

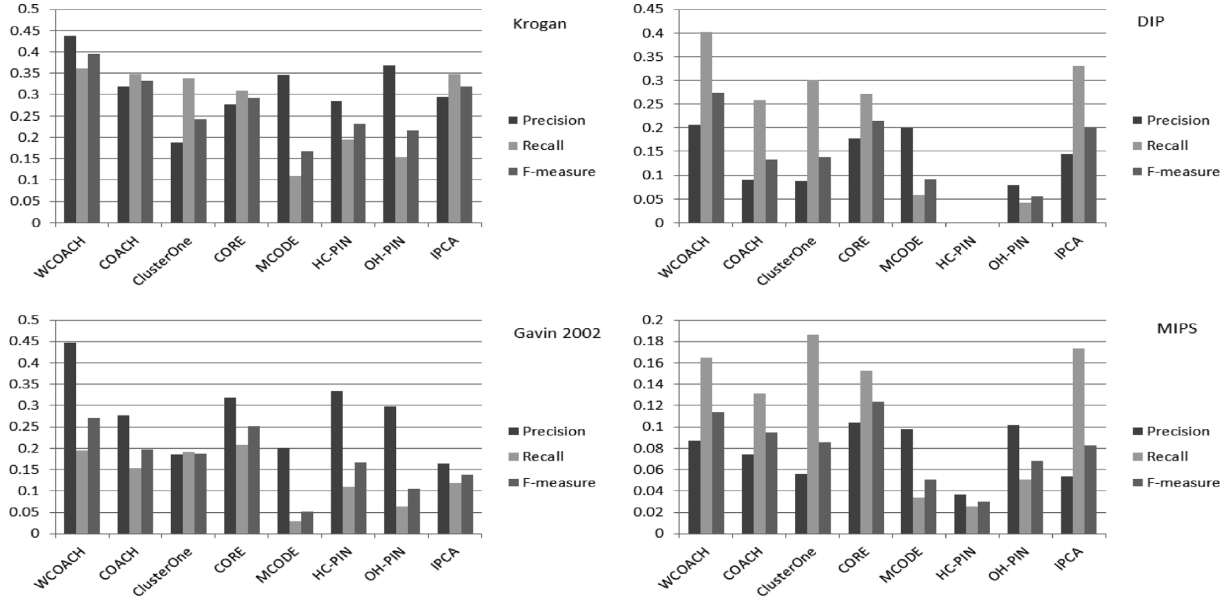

Fig. 3. Precision, recall and F-measure values of various algorithms on the Krogan, DIP, Gavin 2002 and MIPS networks.

plexes with fewer than three proteins are ignored.

All algorithms were run on the DIP, Gavin 2002, MIPS and Krogan networks. Their precision, recall and F-measure values are shown in Fig. 3. It can be seen that WCOACH has a high performance on the DIP, Krogan and Gavin 2002 networks and its precision, recall and F-

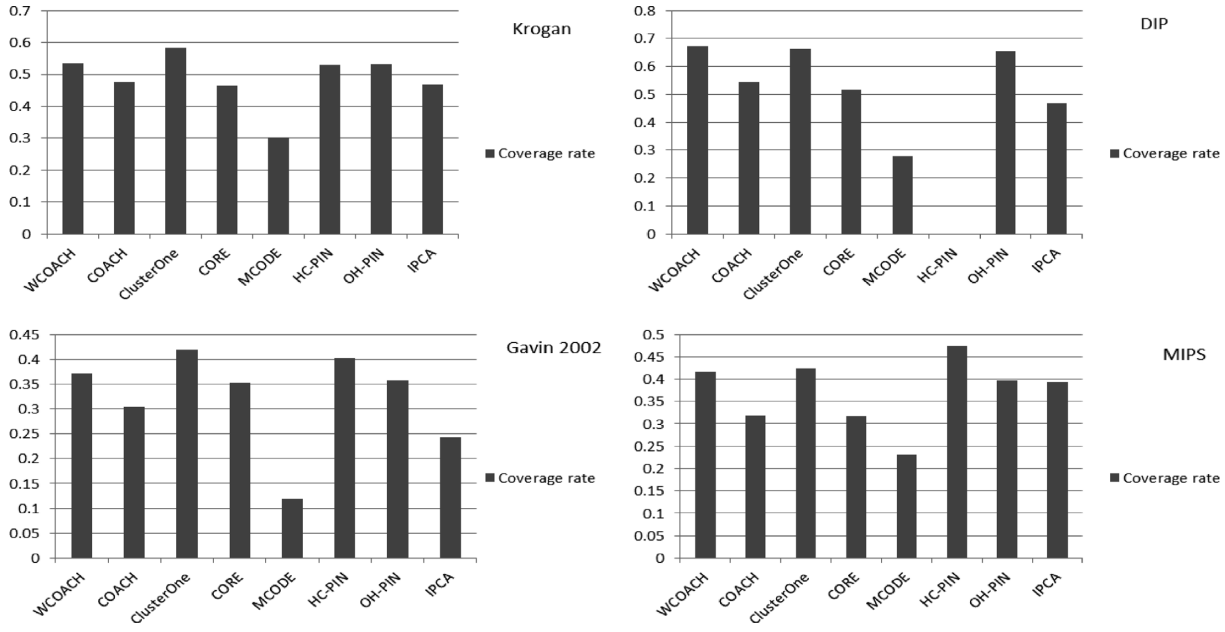

Fig. 4. Coverage rate values of various algorithms on the Krogan, DIP, Gavin 2002 and MIPS networks.

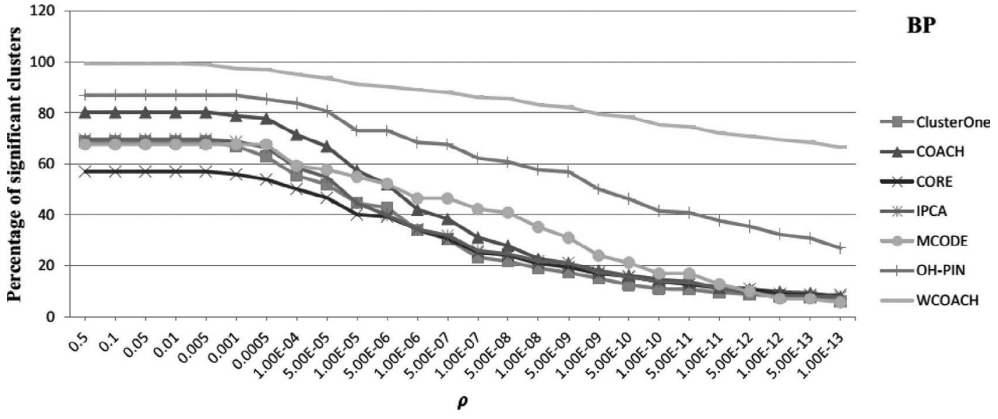

Fig. 5. Percentage of significant clusters for various values of P-value cut-off calculated based on the BP domain.

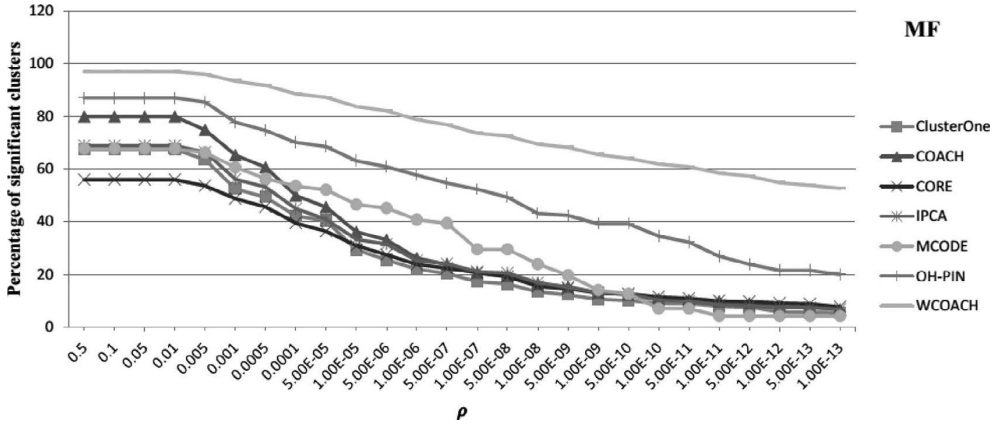

Fig. 6. Percentage of significant clusters for various values of P-value cut-off calculated based on the MF domain.

measure values are higher than those of the other algorithms. Note that in our experiment, the HC-PIN algorithm detected only eight clusters on the DIP network and its precision, recall and F-measure values are very low; for this reason, this algorithm is ignored on the DIP network. On the MIPS network, WCOACH displays the highest F-measure value after CORE and the recall value of our algorithm is higher than COACH, CORE, OH-PIN, HC-PIN and MCODE.

The coverage rates of the algorithms are shown in Fig. 4. WCOACH displays the highest coverage rate on the DIP network. Our method also has acceptable coverage rates on three other networks.

To evaluate the biological significance of this method, the P-values of all predicted clusters on the DIP network were calculated. The *GO Term Finder* tool was applied to calculate P-values based on the BP and MF domains of GO. In this experiment, a predicted cluster P is significant if $P - value\ (P) < \rho$ where ρ is a cut-off for the P-value. The percentage of significant clusters for several values of ρ is shown in Figs. 5 and 6, where the P-values are calculated based on BP and MF, respectively. With respect to using GO for protein complex prediction, the WCOACH method detected more clusters with low P-values.

CONCLUSION

In this paper, the COACH method is modified and a new weighted core attachment method, WCOACH, is proposed to detect protein complexes in weighted PPI networks. The semantic similarity measure between protein pairs based on GO structure is applied to estimate the reliability of protein interactions. This measure, calculated by the Lin (GraSM) method, is considered as the weight of each interaction. The COACH method cannot detect all attachment proteins in complexes. In the WCOACH method, the weight of interactions is applied to detect attachment proteins.

We would like to thank Dr. C. N. Moschopoulos for providing PPI datasets.

REFERENCES

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

Ahn, J., Lee, D. H., Yoon, Y., Yeu, Y., and Park, S. (2013) Improved method for protein complex detection using bottleneck proteins. BMC Med. Inform. Decis. Mak. **13 (Suppl. 1)**, S5.

Bader, G. D., and Hogue, C. W. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics **4**, 2.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., et al. (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.

Chen, J., and Yuan, B. (2006) Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics **22**, 2283–2290.

Chiam, T. C., and Cho, Y.-R. (2012) Accuracy improvement in protein complex prediction from protein interaction networks by refining cluster overlaps. Proteome Sci. **10 (Suppl. 1)**, S3.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2007) Measuring semantic similarity between Gene Ontology terms. Data Knowl. Eng. **61**, 137–152.

Dezső, Z., Oltvai, Z. N., and Barabási, A.-L. (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. Genome Res. **13**, 2450–2454.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30**, 1575–1584.

Feng, J., Jiang, R., and Jiang, T. (2011) A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 621–634.

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

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature **440**, 631–636.

Habibi, M., Eslahchi, C., and Wong, L. (2010) Protein complex prediction based on k-connected subgraphs in protein interaction network. BMC Syst. Biol. **4**, 129.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**, 4569–4574.

King, A. D., Pržulj, N., and Jurisica, I. (2004) Protein complex prediction via cost-based clustering. Bioinformatics **20**, 3013–3020.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature **440**, 637–643.

Leacock, C., and Chodorow, M. (1998) Combining local context and WordNet similarity for word sense identification. In: WordNet: An Electronic Lexical Database (ed.: C. Fellbaum), pp. 265–283. MIT Press, Cambridge, MA.

Leung, H. C., Xiang, Q., Yiu, S., and Chin, F. Y. (2009) Predicting protein complexes from PPI data: a core-attachment approach. J. Comput. Biol. **16**, 133–144.

Li, M., Chen, J.-e., Wang, J.-x., Hu, B., and Chen, G. (2008) Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinformatics **9**, 398.

Li, X.-L., Tan, S.-H., Foo, C.-S., and Ng, S.-K. (2005) Interaction graph mining for protein complexes using local clique merging. Genome Inform. **16**, 260–269.

Li, X., Wu, M., Kwoh, C.-K., and Ng, S.-K. (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics **11 (Suppl. 1)**, S3.

Lin, D. (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on

Machine Learning (ed.: J. W. Shavlik), pp. 296–304. Morgan Kaufmann, San Francisco.

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

Ma, W., McAnulla, C., and Wang, L. (2012) Protein complex prediction based on maximum matching with domain–domain interaction. BBA, Proteins and Proteomics **1824**, 1418–1424.

Ma, X., and Gao, L. (2012) Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability. Inform. Sciences **189**, 233–254.

Mewes, H.-W., Frishman, D., Mayer, K. F., Münsterkötter, M., Noubibou, O., et al. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. **34**, D169–D172.

Moschopoulos, C. N., Pavlopoulos, G. A., Iacucci, E., Aerts, J., Likothanassis, S., et al. (2011) Which clustering algorithm is better for predicting protein complexes? BMC Res. Notes **4**, 549.

Mukhopadhyay, A., Ray, S., and De, M. (2012) Detecting protein complexes in a PPI network: a gene ontology based multiobjective evolutionary approach. Mol. BioSyst. **8**, 3036–3048.

Nepusz, T., Yu, H., and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods **9**, 471–472.

Ozawa, Y., Saito, R., Fujimori, S., Kashima, H., Ishizaka, M., et al. (2010) Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions. BMC Bioinformatics **11**, 350.

Pizzuti, C., and Rombo, S. (2012) Experimental evaluation of topological-based fitness functions to detect complexes in PPI networks. In: Proceedings of the 14th annual conference on Genetic and evolutionary computation (ed.: T. Soule), pp. 193–200. ACM, New York.

Pizzuti, C., and Rombo, S. E. (2013) Restricted neighborhood search clustering revisited: an evolutionary computation perspective. In: Pattern Recognition in Bioinformatics (eds: A. Ngoln, E. Formenti, J.-K. Hao, X.-M. Zhao and T. van Laarhoven), pp. 59–68. Springer, Berlin.

Pizzuti, C., and Rombo, S. E. (2014) Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics **30**, 1343–1352.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009) Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. **37**, 825–831.

Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of 14th International Joint Conference on Artificial Intelligence, pp. 448–453. Morgan Kaufmann, San Fransisco.

Srihari, S., and Leong, H. W. (2013) A survey of computational methods for protein complex prediction from protein interaction networks. J. Bioinform. Comput. Biol. **11**, 1230002.

Wang, J., Li, M., Chen, J., and Pan, Y. (2011) A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 607–620.

Wang, J., Ren, J., Li, M., and Wu, F.-X. (2012a) Identification of hierarchical and overlapping functional modules in PPI networks. IEEE Trans. NanoBioscience **11**, 386–393.

Wang, J., Xie, D., Lin, H., Yang, Z., and Zhang, Y. (2012b) Filtering Gene Ontology semantic similarity for identifying protein complexes in large protein interaction networks. Proteome Sci. **10 (Suppl 1)**, S18.

Wu, M., Li, X., Kwoh, C.-K., and Ng, S.-K. (2009) A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics **10**, 169.

Wu, M., Xie, Z., Li, X., Kwoh, C. K., and Zheng, J. (2013) Identifying protein complexes from heterogeneous biological data. Proteins **81**, 2023–2033.

Wu, Z., and Palmer, M. (1994) Verbs semantics and lexical selection. In: Proceedings ofthe 32nd annual meeting on Association for Computational Linguistics, pp. 133–138. Morgan Kaufmann, San Francisco.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**, 303–305.

Yu, Y., Wang, X., Lin, L., Sun, C., and Wang, X. (2012) Detecting protein complexes based on sequence information in the weighted protein-protein interaction network. J. Comput. Theor. Nanoscience **9**, 1565–1570.

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

Zhang, A., (2009) Protein Interaction Networks: Computational Analysis. Cambridge University Press, New York.