Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Predicting overlapping protein complexes in weighted interactome networks*

Wen-yin NI†, Hui-jun XIONG, Bi-hai ZHAO, Sai HU†‡

(*Department of Information and Computing Science, Changsha University, Changsha 410003, China*)

†E-mail: {masonni, husaiccsu}@163.com

Received Apr. 3, 2013; Revision accepted June 8, 2013; Crosschecked Sept. 23, 2013

Abstract: Protein complexes play important roles in integrating individual gene products to perform useful cellular functions. The increasing mount of protein–protein interaction (PPI) data has enabled us to predict protein complexes. In spite of the advances in these computational approaches and experimental techniques, it is impossible to construct an absolutely reliable PPI network. Taking into account the reliability of interactions in the PPI network, we have constructed a weighted protein–protein interaction (WPPI) network, in which the reliability of each interaction is represented as a weight using the topology of the PPI network. As overlaps are likely to have biological importance, we proposed a novel method named WN-PC (weighted network-based method for predicting protein complexes) to predict overlapping protein complexes on the WPPI network. The proposed algorithm predicts neighborhood graphs with an aggregation coefficient over a threshold as candidate complexes, and binds attachment proteins to candidate complexes. Finally, we have filtered redundant complexes which overlap other complexes to a very high extent in comparison to their density and size. A comprehensive comparison between competitive algorithms and our WN-PC method has been made in terms of the *F*-measure, coverage rate, and *P*-value. We have applied WN-PC to two different yeast PPI data sets, one of which is a huge PPI network consisting of over 6000 proteins and 200000 interactions. Experimental results show that WN-PC outperforms the state-of-the-art methods. We think that our research may be helpful for other applications in PPI networks.

Key words: Protein–protein interaction, Weighted network, Overlap
doi:10.1631/jzus.C13b0097 **Document code:** A **CLC number:** TP311; R857.3

1 Introduction

Proteins are elementary building blocks of biological processes occurring within cells. They play their roles via a very broad network of mutual interactions (Ito *et al.*, 2001). High-throughput methods (Uetz *et al.*, 2000; Gavin *et al*., 2002) have led to the emergence of large protein–protein interaction data sets. Protein complexes consisting of molecular aggregations of proteins assembled by multiple protein interactions are fundamental units of macromolecular organization and play crucial roles in integrating individual gene products to perform useful cellular functions.

Currently, proteomics technologies such as tandem affinity purification (Puig *et al.*, 2001) are available for the prediction of protein complexes. However, these methods are not sufficient to deduce satisfactory conclusions. A computational approach is an alternative way of predicting protein complexes from the available PPI data. A wealth of graph clustering algorithms has been used for the identification of highly connected nodes in protein interaction graphs. One of the most commonly used methods is the molecular complex detection (MCODE) algorithm (Bader and Hogue, 2003), which predicts protein complexes via dense protein sub-networks. CFinder (Adamcsek *et al.*, 2006) is also a popular

‡ Corresponding author

* Project supported by the Scientific Research Foundation of Hunan Province (No. 11C0125), the Scientific Planning Project of Hunan Province (No. XJK011CXJ002), and the Science and Technology Foundation of Changsha City (Nos. K1205049-11 and K1205048-11), China

graph clustering algorithm. Based on the clique percolation method (CPM) (Palla *et al.*, 2005), CFinder attempts to locate all *k*-clique clusters that correspond to fully connected subgraphs of *k* vertices. Another common clustering algorithm is the Markov clustering algorithm (MCL) (Enright *et al.*, 2002). Based on the use of a bootstrapping mechanism, MCL simulates random walks (Pearson, 1905) within graphs and partitions the PPI network into many non-overlapping dense clusters. Divisive projected clustering (DPClus) (Altaf-Ul-Amin *et al.*, 2006) is another graph clustering algorithm used to extract densely connected regions from PPI networks. The intuition underlying SPICi (speed and performance in clustering algorithm) (Jiang and Singh, 2010) is similar to that of DPClus. However, SPICi exploits a cluster expansion approach using a different seed selection criterion and incorporates interaction confidences. An algorithm called clustering based on maximal clique (CMC) (Liu *et al.*, 2009) has been proposed to discover complexes in weighted PPI networks. Protein–protein interactions are assigned an iterative scoring weight to indicate the reliability of the interaction. Recently, a clustering with overlapping neighborhood expansion algorithm named ClusterONE (Nepusz *et al.*, 2012) has been introduced for finding overlapping protein complexes in PPI networks.

All the current algorithms are based on the idea of discovering dense subgraphs. To make a breakthrough, we should take into account the inherent organization. Gavin *et al.* (2006) have conducted detailed research on the organization of complexes. This research has shown that a complex should consist of a core component and attachments. Core proteins are highly co-expressed and each attachment protein binds to a core to form a biological complex. Inspired by this discovery, some core-attachment based algorithms have been proposed, such as Core (Leung *et al.*, 2009), COACH (Wu *et al.*, 2009), and Xiaoke (Ma and Gao, 2012). The excellent performance of core-attachment based algorithms demonstrates the significance of structure in predicting protein complexes.

While great advances have been made in computational approaches for protein complex prediction, the accurate identification of protein complexes is still a bottleneck. A key question is how to measure the reliability of a PPI network and to reduce and tolerate the negative effects of noise. Analysis based on concordance of interaction and expression data suggests that only 30%–50% of the high-throughput interactions are biologically relevant (Deane *et al.*, 2002). Consequently, a crucial step in discovering protein complexes is separating the subset of credible interactions from the background noise. Most current methods generally treat interaction as binary. This means that interaction data are often represented as a network in which edges are either present or absent, without accounting for the quality of each interaction.

In this paper, we develop a novel method named WN-PC for discovering protein complexes based on a weighted PPI network, which can effectively reduce and tolerate the negative effects of noise on protein complex prediction. The existing algorithms generally treat a dense enough subgraph as a complex, but ignore its relationship with its surrounding neighborhood subgraph. We believe that a protein will be joined in a complex if it has high cohesion with subunits in the complex and low coupling with its neighborhood subgraph. We have made a comprehensive comparison between the existing state-of-the-art algorithms and our algorithm. The results show that WN-PC outperforms the competitive methods.

2 Methods

2.1 Constructing a weighted protein–protein interaction network

Let $G=(V, E)$ denote a PPI network, which is an undirected deterministic graph, where V is a set of vertices (proteins), and E is a set of edges (interactions). Current algorithms formulate the problem of identifying protein complexes from a PPI network as that of mining a dense subgraph from the deterministic graph. Taking into account the reliability of interactions, we have constructed a weighted protein–protein interaction (WPPI) network, in which the reliability of each interaction is represented as a weight, using only the topology of the PPI network.

Definition 1 (Weighted protein–protein interaction network) Given an undirected deterministic PPI network $G=(V, E)$, where V is a set of vertices and E is a set of edges, a weighted PPI (WPPI) network is defined as $G=(G, P_E)$, where P_E: $E\rightarrow[0, 1]$ is a probability function that assigns existence probabilities to

edges in E. It is assumed that the existence of an edge $e \in E$ is independent of any other edges.

Considering the reliability of an edge $e(u, v)$, we believe that the greater the number of common neighbors of u and v, the higher the reliability of e, so the existence probability of $e(u, v)$ can be represented as $P_e=N_c/N_{max}$, where N_c is the number of common neighbors of v_i and v_j, and N_{max} is the maximum possible number of common neighbors of the two vertices. N_{max} is represented as $\min(k_u-1, k_v-1)$, where k_u and k_v denote the degrees of u and v, respectively. An example in Fig. 1 illustrates our constructed WPPI network.

Definition 2 (Sample network) A sample network SG=(SV, SE) of a WPPI network $G=(V, E, P_E)$ is a deterministic graph which is a possible outcome of the random variables representing the edges of the WPPI network G (denoted as $G \Rightarrow$SG). In other words, SV=V, SE$\subseteq E$. The sampling probability of SG is given by

$$\Pr(G \Rightarrow \mathrm{SG}) = \prod_{e \in \mathrm{SE}} p(e) \prod_{e \in (E \setminus \mathrm{SE})} (1 - p(e)).$$

The total number of such sample networks is $2^{|E|}$ and $\sum_{i=1}^{n} \Pr(G \Rightarrow \mathrm{SG}_i) = 1$ ($n=2^{|E|}$). According to Definition 2, a WPPI network can be represented as many deterministic networks with sampling probability.

Some concepts are introduced to describe a subgraph or a vertex. The sample degree is used to describe the relationship between a vertex and a subgraph, and the aggregation coefficient is used to describe the degree of coupling of a subgraph.

Definition 3 (Sample degree) Given a WPPI network $G=(V, E, P_E)$ and a vertex u, $u \in V$, SG=$\{\mathrm{SG}_1, \mathrm{SG}_2, \ldots, \mathrm{SG}_n\}$ is the set of all sample networks derived from G, where $\mathrm{SG}_i=(\mathrm{SV}_i, \mathrm{SE}_i)$, $\mathrm{SV}_i=V$, $\mathrm{SE}_i \subseteq E$, $i \in [1, n]$. SD(u, G) denotes the sample degree between u and G.

$$\mathrm{SD}(u, G) = \sum_{i=1}^{n} \Pr(G \Rightarrow \mathrm{SG}_i) \frac{m_i}{|\mathrm{SV}_i|},$$

where m_i is number of edges between u and the vertices of SV_i.

The following theorem gives a simple formula to compute the sample degree:

Theorem 1 Given a WPPI network $G=(V, E, P_E)$ and a vertex u, $u \in V$, the sample degree between u and G can be represented as follows:

$$\mathrm{SD}(u, G) = \sum P(u, v)/|V|, \quad v \in V,$$

where v represents any vertex connected with u.

Proof Let $\{v_1, v_2, ..., v_k\}$ be a set of vertices of G connected with u. The probabilities of (u, v_1), (u, v_2), …, (u, v_k) are $P_1, P_2, \ldots, P_k$, respectively. According to Definition 3,

$$\begin{aligned}\mathrm{SD}(u, G) = {} & \frac{1}{|V|} \sum_{i=1}^{k} (1-P_1)\,(1-P_2)\ldots P_i \ldots (1-P_k) \\ & + \frac{2}{|V|} \sum_{i=1}^{k} \sum_{j>i}^{k} (1-P_1)(1-P_2)\ldots P_i \ldots P_j \ldots (1-P_k) \\ & + \ldots + \frac{k}{|V|} P_1 P_2 \ldots P_i \ldots P_k.\end{aligned}$$

When $k=1$, SD$(u, G)=P_1/|V|$.

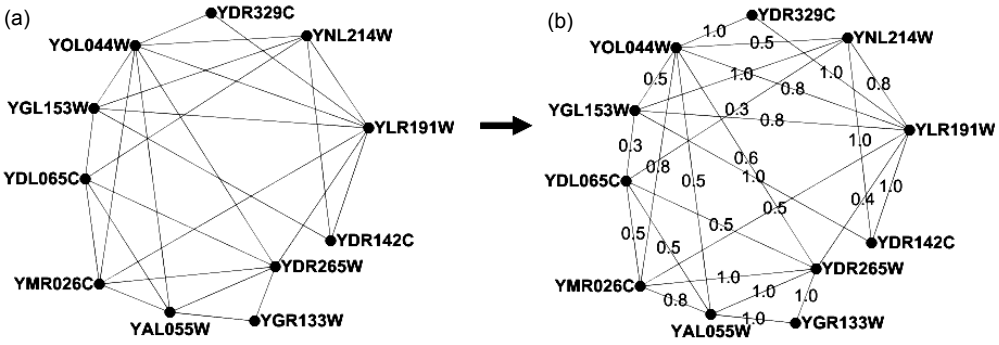

Fig. 1 An example of a constructed weighted protein–protein interaction (PPI) network: (a) original PPI network; (b) weighted PPI network reconstructed using only the topology of the PPI network

The digit in (b) represents the weight of each edge

For k=2,

$$\mathrm{SD}(u,G)=\frac{1}{|V|}[P_1(1-P_2)+(1-P_1)P_2]+\frac{2}{|V|}P_1P_2$$
$$=\frac{1}{|V|}\sum_{i=1}^{2}P_i.$$

For k=3,

$$\mathrm{SD}(u,G)=\frac{1}{|V|}[P_1(1-P_2)(1-P_3)+(1-P_1)P_2(1-P_3)$$
$$+(1-P_1)(1-P_2)P_3]+\frac{2}{|V|}[P_1P_2(1-P_3)$$
$$+P_1P_3(1-P_2)+(1-P_1)P_2P_3]+\frac{3}{|V|}P_1P_2P_3$$
$$=\frac{1}{|V|}\sum_{i=1}^{3}P_i.$$

Assume that when k=n, $\mathrm{SD}(u,G)=\frac{1}{|V|}\sum_{i=1}^{n}P_i$.

For k=n+1,

$$\mathrm{SD}(u,G)$$
$$=\frac{1}{|V|}\sum_{i=1}^{n}(1-P_1)\ (1-P_2)\dots P_i\dots(1-P_n)$$
$$+\frac{P_{n+1}}{|V|}[(1-P_1)(1-P_2)\dots(1-P_i)\dots(1-P_n)$$
$$+\sum_{i=1}^{n}(1-P_1)(1-P_2)\dots P_i\dots(1-P_n)]$$
$$+\frac{2}{|V|}\sum_{i=1}^{n}\sum_{j>i}^{n}(1-P_1)(1-P_2)\dots P_i\dots P_j\dots(1-P_n)$$
$$+\frac{P_{n+1}}{|V|}\sum_{i=1}^{n}\sum_{j>i}^{n}(1-P_1)(1-P_2)\dots P_i\dots P_j\dots(1-P_n)$$
$$+\dots+\frac{n}{|V|}P_1P_2\dots P_n+\frac{1}{|V|}P_1P_2\dots P_nP_{n+1}$$
$$=\frac{1}{|V|}\sum_{i=1}^{n}(1-P_1)(1-P_2)...P_i...(1-P_n)$$
$$+\frac{2}{|V|}\sum_{i=1}^{n}\sum_{j>i}^{n}(1-P_1)(1-P_2)\dots P_i\dots P_j\dots(1-P_n)+...$$
$$+\frac{n}{|V|}P_1P_2...P_n+\frac{P_{n+1}}{|V|}[(1-P_1)(1-P_2)...(1-P_i)...(1-P_n)$$
$$+\sum_{i=1}^{n}(1-P_1)(1-P_2)...P_i...(1-P_n)$$
$$+\sum_{i=1}^{n}\sum_{j>i}^{n}(1-P_1)(1-P_2)\dots P_i\dots P_j\dots(1-P_n)$$
$$+...+P_1P_2...P_n],$$

while

$$\frac{1}{|V|}\sum_{i=1}^{n}(1-P_1)(1-P_2)...P_i...(1-P_n)$$
$$+\frac{2}{|V|}\sum_{i=1}^{n}\sum_{j>i}^{n}(1-P_1)(1-P_2)\dots P_i\dots P_j\dots(1-P_n)$$
$$+...+\frac{n}{|V|}P_1P_2...P_n=\frac{1}{|V|}\sum_{i=1}^{n}P_i.$$

So, when k=n+1,

$$\mathrm{SD}(u,G)=\frac{1}{|V|}\left(\sum_{i=1}^{n}P_i+P_{n+1}\right)=\frac{1}{|V|}\sum_{i=1}^{n+1}P_i.$$

Definition 4 (Aggregation coefficient) Given a WPPI network G=(V, E, P_E) and a vertex u, $u\in V$, NG=(NV, NE, P_{NE}) is a neighbor's subgraph of vertices in G, $G\cap$NG=$\varnothing$. The aggregation coefficient between u and G is defined as

$$\mathrm{AC}(u,G)=\frac{\mathrm{SD}(u,G)|V|}{\mathrm{SD}(u,G)|V|+\mathrm{SD}(u,\mathrm{NG})|\mathrm{NV}|}.$$

SD(u, G)$|V|$ indicates the degree of cohesion between u and the network G, while SD(u, NG)|NV| represents the degree of coupling between u and the neighborhood network NG. A subgraph with high cohesion between its subunits and low coupling with its neighborhood subgraph can be represented as a protein complex. So, the aggregation coefficient is introduced to decide whether an identified cluster can be represented as a protein complex.

As a forerunner of WN-PC, we construct a WPPI network in which the reliability of each interaction is represented as a weight.

Algorithm 1 illustrates the procedure for constructing the WPPI network.

Algorithm 1 Constructing the WPPI network

Input: the PPI network Q.

Output: the WPPI network G=(V, E, P_E) and the harmonic mean of the vertices' sample degrees HM_SD.

1. Construct the WPPI network G from the original PPI network.

2. Remove edges with P_E=0, and then cut out the isolated vertices.

3. HM_SD=harmonic mean of SD(v_i, NG$_i$), $v_i\in V$, where NG$_i$ is a neighborhood graph of v_i.

2.2 Predicting protein complexes

WN-PC operates in two stages: predicting complexes and processing redundant data.

In WN-PC, a protein is added to a candidate complex if the aggregation coefficient is over a threshold value, and an attachment protein is bound to a candidate complex if the sample degree is over the HM_SD, which is obtained from Algorithm 1.

Each vertex can be drawn as a seed to grow a candidate complex. Initially, a selected seed and its neighbors are put into SC to form a candidate complex. Then neighbors with an aggregation coefficient less than the threshold ACT are removed from SC and marked with the tag AT_CAN. This means that adding these neighbors would lead to SC having a high coupling. The vertices with the tag AT_CAN have then lost the opportunity to be drawn as seeds. We discard candidate complexes that contain less than three proteins. To avoid a candidate complex becoming a subset of others after adding attachment proteins, not all the seeds can be used as attachment proteins to be bound to candidate complexes. In other words, only vertices labeled with AT_CAN would be candidate attachments. For each candidate attachment, if the sample degree within a candidate complex is over the threshold HM_SD, it is inserted into the candidate complex.

Algorithm 2 illustrates the overall framework to predict complexes.

Algorithm 2 Predicting complexes

Input: the WPPI network $G=(V, E, P_E)$, the aggregation coefficient threshold ACT, and the sample degree threshold HM_SD.

Output: PC, which is the set of protein complexes.

```
For each vertex v∈V
  If v is marked with AT_CAN then
    Skip;
  End If
  Insert v and its neighbors into SC;
  For each vertex u∈SC
    If AC(u, SC)<ACT then
      Remove u from SC and mark u with AT_CAN;
    End if
  End for
  If length(SC)>2 then
    Insert SC into PC;
  End if
End for
For each vertex marked with AT_CAN av∈V
  If SD(av, pc_i) > HM_SD(pc_i ⊂ PC)
    Insert av into pc_i;
  End if
End for
```

The last stage is redundant processing. Some overlaps are likely to have biological importance, but overlaps reaching a very high threshold should be dealt with. For a pair of complexes which overlap a threshold, the one with smaller density and size is discarded. Note that the concept of density is redefined as the sum of the sample degrees for the entire vertex. In this study, the overlap threshold value is typically set as 0.8 (Nepusz *et al.*, 2012), where the overlap score of two complexes A and B is defined as follows (Li *et al.*, 2007):

$$\mathrm{NA}(A,B)=\frac{|A\cap B|^2}{|A||B|}. \tag{1}$$

3 Results and discussion

We compared the performance of our method with those of another five state-of-the-art algorithms, MCODE (Bader and Hogue, 2003), MCL (Enright *et al.*, 2002), CMC (Liu *et al.*, 2009), ClusterONE (Nepusz *et al.*, 2012), and COACH (Wu *et al.*, 2009). We have applied WN-PC and the competitive algorithms to two yeast PPI networks, including DIP (the Database of Interacting Proteins) data (Xenarios *et al.*, 2002) updated on 18 Aug. 2012 and BioGRID (Biological General Repository for Interaction Datasets) data (Stark *et al.*, 2006) version 3.19. The DIP data consisted of 4895 proteins and 21 776 interactions among the proteins. To make a comprehensive comparison, we used both the core set (5445 proteins, 49 954 interactions) and the full set (6055 proteins, 201 229 interactions) of BioGRID. To evaluate the protein complexes predicted by WN-PC, we derived 408 typical complexes including two or more proteins from CYC2008 (Pu *et al.*, 2009) as the benchmark complex set.

We will first analyze in detail the results from DIP data using three extensively studied evaluation criteria: the F-measure, the coverage rate, and the functional enrichment of GO terms (P-value). Finally, the results from BioGRID data will also be briefly presented to demonstrate the effectiveness of WN-PC

for large-scale PPI networks. For all those competitive algorithms, the optimal parameters were set based on recommendations by the authors.

3.1 *F*-measure and coverage rate

One evaluation method we used was to match the predicted complexes with the benchmark complex set and calculate the *F*-measure and coverage rate.

Given a predicted complex pc∈PC and a benchmark complex bc∈BC, if NA(pc, bc)≥*t*, then pc and bc are considered to be matching, where *t* is a predefined threshold. In general, *t* is set as 0.2 (Bader and Hogue, 2003; Wu *et al*., 2009). TP is defined as the number of complexes matched by the benchmark complex set, while TN is the number of predicted benchmark complexes. The *F*-measure is used to evaluate the overall performance of the clustering methods.

$$F\text{-measure}=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}, \qquad (2)$$

where Precision=TP/|PC|, Recall=TN/|BC|.

The coverage rate shows how many proteins in the benchmark complex set can be covered by the predicted complexes (Brohée and van Helden, 2006; Friedel *et al.*, 2008). Given a benchmark complex set BC and a predicted complex set PC, T_{ij} is the number of proteins in common between the *i*th benchmark complex and the *j*th predicted complex. The coverage rate is then defined as

$$\text{CR}=\sum_{i=1}^{|\text{BC}|}\max_{1\le j\le|\text{PC}|}T_{ij}\Big/\sum_{i=1}^{|\text{BC}|}N_i, \qquad (3)$$

where N_i is the number of proteins in the *i*th benchmark complex.

The basic information on predicted complexes produced by various algorithms using DIP data is presented in Table 1.

WN-PC matched 328 complexes of predicted complexes, which is the highest number compared with competing algorithms, and contained the second-highest number of matched benchmark complexes after COACH, while the number of complexes identified by WN-PC was far lower than that identified by COACH. The Precision of WN-PC, MCODE, MCL, CMC, COACH, and ClusterONE was 59%, 51%, 19%, 47%, 35%, and 41%, respectively, and the Recall was 53%, 7%, 49%, 31%, 54%, and 32%, respectively. WN-PC achieved the highest precision and comparable recall.

Table 1 Results of various algorithms using DIP data

Algorithm	#PC	MBC	MPC
WN-PC	557	215	328
ClusterONE	345	131	142
MCODE	55	30	28
MCL	934	198	179
COACH	896	221	318
CMC	339	126	159

#PC is the number of complexes identified by each algorithm, MBC represents the number of benchmark complexes that match at least a predicted complex, and MPC is the number of predicted complexes which match at least a benchmark complex

Fig. 2 shows a comprehensive comparison among the selected methods in terms of the *F*-measure and the coverage rate. Using DIP data, the *F*-measure of WN-PC was 56%, which was 330.77%, 107.41%, 51.35%, 30.23%, and 55.56% higher than those of MCODE, MCL, CMC, COACH, and ClusterONE, respectively; WN-PC produced the highest coverage rate of 60%, which was 106.9%, 13.21%, 87.5%, 5.26%, and 66.67% higher than those of MCODE, MCL, CMC, COACH, and ClusterONE, respectively.

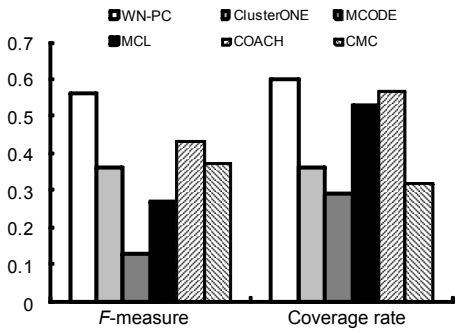

Fig. 2 Performance comparison for various algorithms using DIP data

From Table 1 and Fig. 2, we can draw the conclusion that WN-PC achieves the best performance in terms of the *F*-measure and the coverage rate using DIP data. In other words, WN-PC achieves higher accuracy in predicting complexes than competitive algorithms.

Fig. 3 illustrates the example of the real SAGA complex, where our predicted complex can cover more proteins than competitive algorithms.

The real SAGA complex in the benchmark consists of 20 proteins, of which 3 have been isolated. Fig. 3a is the benchmark SAGA complex, and Figs. 3b–3f are the complexes predicted by WN-PC, COACH, ClusterONE, CMC, and MCL, respectively. MCODE cannot generate the matched SAGA complex. The complex predicted by WN-PC had 20 proteins and covered 16 proteins (circled). According to Eq. (1), the NA was 61%. However, COACH, ClusterONE, CMC, and MCL covered only 11, 8, 6, and 3 proteins of the real SAGA complex, respectively. The NA was 46.5%, 35.6%, 30%, and 11.3%, respectively.

3.2 GO analysis

To test the biological significance of the complexes predicted by WN-PC, we adopted the functional enrichment of GO terms (*P*-value). The *P*-value is considered as a measure of the possibility that a predicted complex is a real protein complex. A low *P*-value of a predicted protein complex achieves high statistical significance. A cut-off parameter is used to differentiate significant groups from insignificant ones. In this work, we used the recommended cut-off of 0.01 (Hu *et al.*, 2005). Table 2 compares *P*-values obtained by various algorithms using DIP data. In general, a higher proportion and *P*-score indicate that proteins in the same protein complexes tend to share

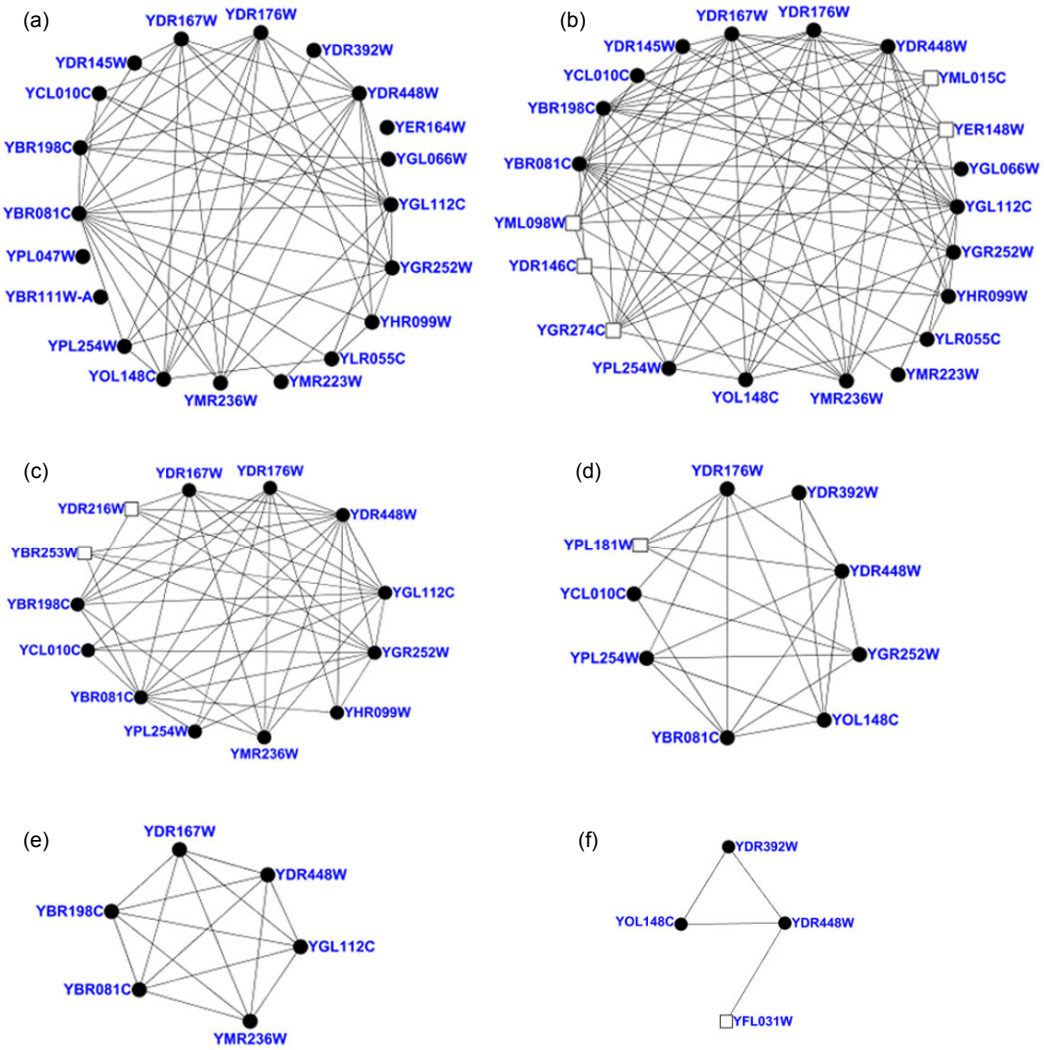

Fig. 3 SAGA complexes predicted by various algorithms
(a) Benchmark; (b) WN-PC; (c) COACH; (d) ClusterONE; (e) CMC; (f) MCL

Table 2 Statistical significance of complexes predicted by various algorithms

Algorithm	#PC	#SC	Proportion	*P*-score
WN-PC	557	520	93.36%	12.70
ClusterONE	345	242	70.14%	8.01
MCODE	55	49	89.09%	8.26
MCL	934	407	43.58%	5.68
COACH	896	715	79.80%	8.07
CMC	339	279	82.30%	6.96

#PC is the number of predicted complexes, #SC is the number of significant complexes, and *P*-score is the average −lg(*P*-value) of all the significance complexes

higher functional similarity, so they can be used to evaluate the overall quality of predicted protein complexes.

Table 2 shows that the vast majority of our predicted complexes (93.36%) were significant. WN-PC also obtained a higher *P*-score of significant complexes than the other five algorithms. In other words, the complexes predicted by WN-PC had the most biological significance.

For further comparison between WN-PC and competitive methods, we employed piecewise statistics of the significant complexes according to their *P*-value. Fig. 4 shows the percentage of the predicted complexes whose *P*-values fell within $(0, 10^{-15})$, $[10^{-15}, 10^{-10})$, and $[10^{-10}, 0.01)$. Higher proportions of significant complexes were generated by WN-PC than by the other algorithms with *P*-values falling within $(0, 10^{-15})$ and $[10^{-15}, 10^{-10})$. In particular, the advantages are more obvious with *P*-value less than 10^{-15}. Comparison of all the results in Table 2 and Fig. 4 shows that WN-PC outperformed the other methods, in terms of the number of significant complexes and the high statistical significance.

3.3 Effect of parameter ACT

WN-PC employs a user-defined parameter ACT in Algorithm 2 to filter low aggregation coefficient neighbors of seeds. In this subsection, we study the effect of the threshold ACT on the performance of WN-PC. Fig. 5 shows how the *F*-measure and the coverage rate of WN-PC fluctuated under various values of ACT based on the DIP data. The coverage rate and the value of ACT had an inverse relationship and the *F*-measure reached the top value when ACT was assigned 0.2. The predicted complexes were 1154, 825, and 535 when ACT was set as 0, 0.1, and 0.2, respectively. The predicted complexes then decreased with the increase of ACT. To obtain a good balance of both the *F*-measure and the coverage rate, we set ACT as 0.19 in our experiment.

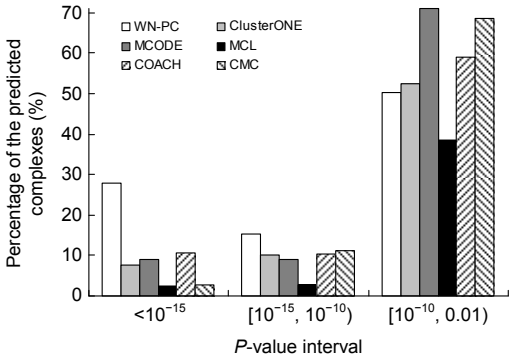

Fig. 4 Comparison of the proportions of significant complexes within *P*-value intervals using various algorithms

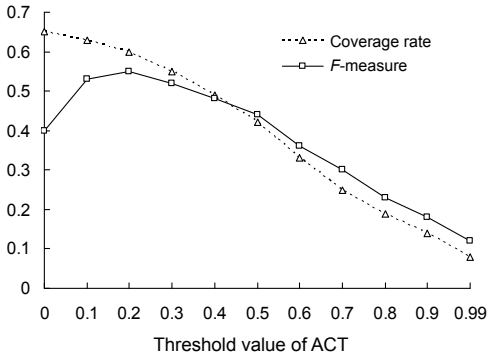

Fig. 5 Effect of parameter ACT on the coverage rate and *F*-measure

3.4 Results using BioGRID data

To test the robustness of WN-PC for large-scale PPI networks, we also performed WN-PC on BioGRID data, including a core set and a full set. BioGRID is a public database generated by various experimental technologies, e.g., affinity purification, two-hybrid, dosage, and synthetic. As affinity purification methods (Gavin *et al.*, 2006) are well suited for studying complexes under near-physiological conditions (Edwards *et al.*, 2002; Kemmeren *et al.*, 2002), we chose protein–protein interactions generated by affinity purification from the full set to form a core set.

Fig. 6 shows the *F*-measure of each method using BioGRID data. WN-PC achieved the best performance on the *F*-measure for both BioGRID data sets. In detail, on the core set, the *F*-measure of WN-PC was 59%, which was 20.41%, 247.06%, 321.43%, 40.48%, and 63.89% higher than those of ClusterONE, MCODE, MCL, COACH, and CMC, respectively; on the full set, the *F*-measure of WN-PC was 36%, which was 33.33%, 3500%, 1100%, 300%, and 414.29% higher than those of ClusterONE, MCODE, MCL, COACH, and CMC, respectively. Note that the *F*-measure of WN-PC on the full set was higher than those of MCODE and MCL on the core set, even though the core set is more reliable than the full set.

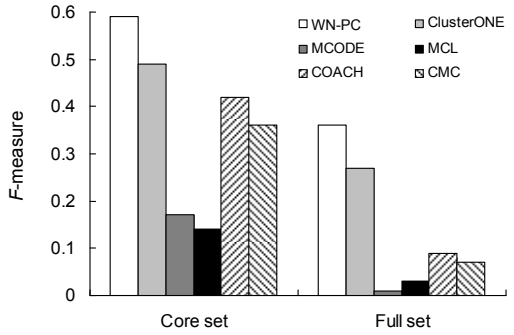

Fig. 6 *F*-measures of various methods using BioGRID data

Table 3 shows the coverage rate and maximum number of complexes predicted by various algorithms using BioGRID data. Using the BioGRID data, the coverage rate of WN-PC was higher than those of ClusterONE and MCODE, and lower than those of COACH and MCL. On the full set of BioGRID data, almost all the proteins in the benchmark complex set were covered by the complexes predicted by MCL and CMC, due to their super complex, whose MS values were over 4000. For the benchmark complex set, the total number of proteins was 1920, and the number of proteins in the maximal complex was 81. From this point of view, we think that the coverage rate cannot be used as an effective criterion to assess the quality of the complexes predicted by various algorithms, because of the influence of the predicted maximal complex.

In conclusion, WN-PC outperformed other competitive algorithms and had better robustness than others in large-scale PPI networks.

Table 3 MS and CR values of various methods using BioGRID data

Algorithm	MS		CR	
	Core set	Full set	Core set	Full set
WN-PC	138	175	0.77	0.77
ClusterONE	86	108	0.71	0.59
MCODE	137	266	0.46	0.42
MCL	3454	5424	0.88	0.99
COACH	504	552	0.84	0.88
CMC	10	4293	0.32	0.98

MS is the number of proteins in the maximal complex predicted by each algorithm and CR is the coverage rate

4 Conclusions

Protein complexes play important roles in many molecular processes and functions. High-throughput methods have led to the emergence of a large amount of protein–protein interaction (PPI) data, which has enabled us to detect protein complexes. Current computational methods regard the PPI network as a reliable network and ignore the negative effects of noise. However, research shows that only 30%–50% of the high-throughput interactions are biologically relevant. Assessing the quality of protein–protein interactions and designing a more effective and fault-tolerant approach are highly desirable.

In this paper, we propose a weighted network-based method to predict protein complexes from PPI networks. We first reconstructed a weighted PPI network, where edges are assigned existence probabilities and unreliable edges and vertices are removed. In WN-PC, the aggregation coefficient, not density, is the key criterion to measure whether a subgraph can be represented as a candidate complex. The sample degree is used to decide whether an attachment protein can be bound to a candidate complex.

Experiments showed that WN-PC has higher accuracy, more significant biological relevance, and stronger robustness than other algorithms for large-scale PPI networks with noise. The experimental results also demonstrated that taking into account the reliability of interactions can effectively reduce the negative effects of false-positive interactions for complex prediction.

References

Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T., 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**(8):1021-1023. [doi:10.1093/bioinformatics/btl039]

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. [doi:10.1186/1471-2105-7-207]

Bader, G.D., Hogue, C.W.V., 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.*, **4**:2. [doi:10.1186/1471-2105-4-2]

Brohée, S., van Helden, J., 2006. Evaluation of clustering algorithms for protein-protein interaction network. *BMC Bioinf.*, **7**:488. [doi:10.1186/1471-2105-7-488]

Deane, C., Salwinski, L., Xenarios, I., Eisenberg, D., 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteom.*, **1**(5):349-356. [doi:10.1074/mcp.M100037-MCP200]

Edwards, A., Kus, B., Jansen, R., Creenbaum, D., Greenblatt, J., Gerstein, M., 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**(10):529-536. [doi:10.1016/S0168-9525(02)02763-4]

Enright, A., Dongen, S., Ouzounis, C., 2002. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.*, **30**(7):1575-1584. [doi:10.1093/nar/30.7.1575]

Friedel, C., Krumsiek, J., Zimmer, R., Vingron, M., Wong, L., 2008. Boostrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. Proc. 12th Annual Conf. on Research in Computational Molecular Biology (RECOMB), p.3-16. [doi:10.1007/978-3-540-78839-3_2]

Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**(6868):141-147. [doi:10.1038/415141a]

Gavin, A., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., *et al.*, 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084):631-636. [doi:10.1038/nature04532]

Hu, H., Yan, X., Huang, Y., Han, J., Zhou, X., 2005. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21**(Suppl 1):i213-i221. [doi:10.1093/bioinformatics/bti1049]

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS*, **98**(8):4569-4574. [doi:10.1073/pnas.061034498]

Jiang, P., Singh, M., 2010. A fast clustering algorithm for large biological networks. *Bioinformatics*, **26**(8):1105-1111. [doi:10.1093/bioinformatics/btq078]

Kemmeren, P., Berkum, N., Vilo, J., Bijma, T., Donders, R., Brazma, A., Holstege, F., 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, **9**(5):1133-1143. [doi:10.1016/S1097-2765(02)00531-2]

Leung, H., Xiang, Q., Yiu, S., Chin, F., 2009. Predicting protein complexes from PPI data: a core-attachment approach. *J. Comput. Biol.*, **16**(2):133-144. [doi:10.1089/cmb.2008.01TT]

Li, X.L., Foo, C.S., Ng, S.K., 2007. Discovering Protein Complexes in Dense Reliable Neighborhoods of Protein Interaction Networks. IEEE Computational Systems Bioinformatics Conf., **6**:157-168.

Liu, G., Wong, L., Chua, H.N., 2009. Complex discovery from weighted PPI networks. *Bioinformatics*, **25**(15):1891-1897. [doi:10.1093/bioinformatics/btp311]

Ma, X., Gao, L., 2012. Discovering protein complexes in protein interaction networks via exploring the weak ties effect. *BMC Syst. Biol.*, **6**(Suppl 1):S6. [doi:10.1186/1752-0509-6-S1-S6]

Nepusz, T., Yu, H., Paccanaro, A., 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**(5):471-475. [doi:10.1038/nmeth.1938]

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043):814-818. [doi:10.1038/nature03607]

Pearson, K., 1905. The problem of random walk. *Nature*, **72**(1858):139-144. [doi:10.1038/072139a0]

Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S., 2009. Up-to-date catalogues of yeast protein complexes. *Nucl. Acids Res.*, **37**(3):825-831. [doi:10.1093/nar/gkn1005]

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Séraphin, B., 2001. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**(3):218-229. [doi:10.1006/meth.2001.1183]

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., 2006. BioGRID: a general repository for interaction datasets. *Nucl. Acids Res.*, **34**:D535-D539. [doi:10.1093/nar/gkj109]

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.*, 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**(6770):623-627. [doi:10.1038/35001009]

Wu, M., Li, X., Kwoh, C., Ng, S., 2009. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinf.*, **10**:169. [doi:10.1186/1471-2105-10-169]

Xenarios, I., Salwnski, L., Duan, X., Higney, P., Kim, S., Eisenberg, D., 2002. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.*, **30**(1):303-305. [doi:10.1093/nar/30.1.303]