Measuring Boundedness for Protein Complex Identification in PPI Networks

Tiantian He, and Keith C.C. Chan

Abstract—The problem of identifying protein complexes in Protein-Protein Interaction (PPI) networks is usually formulated as the problem of identifying dense regions in such networks. In this paper, we present a novel approach, called TBPCI, to identify protein complexes based instead on the concept of a measure of boundedness. Such a measure is defined as an objective function of a Jaccard Index-based connectedness measure which takes into consideration how much two proteins within a network are connected to each other, and an association measure which takes into consideration how much two connecting proteins are associated based on their attributes found in the Gene Ontology database. Based on the above two measures, the objective function is derived to capture how strong the proteins can be considered as bounded together and the objective value is therefore referred as the aggregated degree of boundedness. To identify protein complexes, TBPCI computes the degree of boundedness between all possible pairwise proteins. Then, TBPCI uses a Breadth-First-Search method to determine whether a protein-pair should be incorporated into the same complex. TBPCI has been tested with several real data sets and the experimental results show it is an effective approach for identifying protein complexes in PPI networks.

Index Terms— protein-protein interaction networks, protein complex identification, protein complex discovery

1 INTRODUCTION

A protein complex is a biomolecule that contains a number of proteins connecting to each other to perform different cellular functions [1], such as replication, transcription and the control of gene expression, etc. [2]. Due to its important role in the understanding of cellular organizations and functions, recently much effort has been made to discover protein complexes in the protein-protein interaction (PPI) networks.

To identify protein complexes on a large scale, experiments including affinity purification (AP) followed by mass spectrometry (MS) have to be executed [3], [4]. Though it is an effective method, AP/MS cannot be seen as an efficient approach since many experiments using different baits have to be carried out each time [5].

To speed up the protein complex identification process, some attempts to identify protein complexes using different computational methods have been made in recent years [41]. Many of these methods are developed based on the use of graph topologies to cluster graph nodes to identify graph clusters. A few of the other methods identify graph clusters by taking into consideration the functional information of proteins [45] and extracting binding molecular relationships [46]. Given a PPI network represented as a graph that contains vertices representing proteins and edges representing interactions, these algorithms can discover clusters based on different topological properties. Due to some evidence from known PPI networks showing that proteins within the same protein complex tend to interact more with each other, density-based techniques aiming at identifying densely connected areas are used more when discovering protein complexes in a PPI network [6]. Even though there are other non-density-based approaches, density-related properties are still considered direct or indirectly in discovering the initial graph clusters.

Several density-based graph clustering algorithms that are used for identifying protein complexes have been shown to be rather effective. For example, by taking into consideration local neighborhood density, MCODE [7] has been found to be able to identify graph clusters with densely connected vertices that correspond well to protein complexes in a PPI network. Other than MCODE, MCL [8], another density-based graph clustering algorithm, has also been shown to be effective in identifying protein complexes. MCL is an approach that is based on random walk and it can discover densely connected regions by using expansion and inflation operators [9].

In addition to MCODE and MCL, RNSC [10], has also been shown to be able to effectively identify protein complexes in a given PPI network using a density-based graph-partitioning approach. RNSC finds an optimal set of partitions of a PPI network by employing different cost functions defined in terms of cluster density, cluster size and functional homogeneity. These partitions have been shown to correspond well to known protein complexes.

Similar to RNSC, the algorithm proposed in [11] also attempts to find protein complexes by partitioning a PPI network graph. But it uses a minimum vertex-cut to iden-

- *Tiantian He is with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR. E-mail: csthe@comp.polyu.edu.hk.*
- *Keith C.C. Chan is with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR. E-mail: cskc-chan@comp.polyu.edu.hk.*

Please provide a complete mailing address for each author, as this is the address the 10 complimentary reprints of your paper will be sent

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

tify cluster boundaries and the discovered protein complexes consist of proteins connecting more with the ones in the same complex. In other words, it is also a density-based graph clustering algorithm.

Other than graph density, other graph properties have also been used for protein complex identification. For example, a graph clustering algorithm called DPClus [12], can discover and refine graph clusters by keeping track of cluster periphery. The efficiency of DPClus is improved by another algorithm called IPCA [13]. IPCA finds graph clusters which possess appropriate vertex distance and density.

An example of a non-density-based algorithm for graph clustering is CFinder [14]. It identifies protein complexes using a method of clique percolation. Besides CFinder, the CMC [15] is also a clique-finding algorithm. It identifies cliques by iteratively assigning weights, interpreted as reliability of interactions between proteins, to edges. CMC tries to find a number of cliques which have the highest value of such weights. In [16], an algorithm called COACH is proposed to find protein complexes by making use of a graph property called core-attachment.

Besides the above approaches, some algorithms which can be used to identify overlapping graph clusters have also been used to identify overlapping protein complexes. For example, in [17] an algorithm is proposed to detect overlapping protein complexes based on a generative network model.

As there is some evidence that proteins in the same complex may perform similar functions, there have also been some attempts to identify protein complexes based on what are known about protein functions. In [37], a method that identifies protein complexes based on finding clusters in which proteins perform similar functions is proposed. In [38], another algorithm is proposed to simultaneously consider PPI network data and gene expression data in the protein complex identification process. In [42], an algorithm called PCIA is proposed to identify protein complexes in PPI networks based on network topology and attribute information. It makes use first of a measure of attribute similarity followed by the use of the MCL algorithm to identify densely connected clusters during the process. In [27], an algorithm called GMFTP, is proposed to identify protein complexes using generative model by also considering both network topology and attribute information. The effectiveness of these algorithms shows that protein complexes can be more accurately identified when both topology and attribute information are considered.

In this paper, we propose a novel approach for identification of protein complexes in PPI networks based on information available about both attributes of the proteins and topology of their connections. Unlike traditional approaches, such as those based on graph partitioning, clique percolation, etc., the problem of protein identification is formulated as an optimization problem.

We call the approach as TBPCI (Tightly Bounded Protein Complex Identifier) algorithm. Given a PPI network, TBPCI assigns attribute values to each protein based on information obtained from the Gene Ontology (GO) database [18] which include those that can be classified as *biological processes*, *molecular functions* and *cellular components*. *Biological processes* describe the biological objectives a protein is involved in. *Molecular functions* are concerned with the biochemical activities performed by a protein and *cellular components* are used to describe the location where a protein is active in the cells. In order to avoid bias, attribute information on the cellular components that may reveal the complex affiliation are not considered in our implementation as, knowing what cellular components a protein belongs to may reveal the protein complex it belongs to and knowing about this may lead to biased evaluation.

Given a PPI network represented as a graph, TBPCI performs its tasks by firstly computing the *Degree of Connectedness* (σ) which is defined to be between each pair of proteins in the PPI network. σ quantifies the extent that two proteins are similar based on the topology of the network. A higher value of σ can be interpreted as a higher degree of similarity between the local topology of the network that the proteins are in.

In addition to considering topology, TBPCI also measures how much the attribute values of two proteins are associated with each other. The measure, which we call, the *Degree of Attribute Association* (θ) is used to measure the strength of attribute correlation between each pair of proteins. The higher θ is, the stronger the attribute correlation between two proteins is.

After obtaining σ and θ for each pair of proteins in the PPI network, by representing the network in the form of a graph, TBPCI uses an iterative method to obtain an optimal weighted graph ($\mathbf{W}$), each element of which measures how strong a pair of proteins can be bounded together. We name the entry in $\mathbf{W}$ as *Degree of Boundedness* (w) between pairwise proteins. Using a Breadth-First-Search method in $\mathbf{W}$, TBPCI can find graph clusters that have more tightly bound in a sense. These graph clusters are then considered to represent protein complexes.

In order to evaluate the performance of TBPCI, we have tested TBPCI with different real data sets. The experimental results show that TBPCI can identify protein complexes relatively more accurately. It is also found to be able to identify more protein complexes with shared significant GO terms, compared with other algorithms.

The rest of the paper is organized as follows. The problem of the identification of protein complex using both structure and attribute information is defined mathematically in section 2. In section 3, the details of TBPCI are presented. How TBPCI were tested with different sets of real data and how it was compared with different graph clustering algorithms is described in section 4. In section 5, we give a brief summary of our work and discuss future work based on the existing version of TBPCI.

2 Problem Statement and Notation

Given a PPI network containing n_V proteins and n_E interactions, it is represented as $G = (V, E, \Lambda)$, where 1) V represents as the vertex set of the PPI network; 2) E represents as the edge set of the PPI network; 3) Λ is the set

attribute values for proteins in G and it contains three subsets, Λ_p, Λ_f, Λ_c. They represent the sets of attributes of *biological processes*, *molecular functions* and *cellular components*, respectively.

Then, the possible values of these attributes in different domains can be represent as: $\Lambda_p = \{a_{p,1}, a_{p,2}, \ldots, a_{p,|\Lambda p|}\}$, $\Lambda_f = \{a_{f,1}, a_{f,2}, \ldots, a_{f,|\Lambda f|}\}$, and $\Lambda_c = \{a_{c,1}, a_{c,2}, \ldots, a_{c,|\Lambda c|}\}$, where $|\Lambda_p|$, $|\Lambda_f|$ and $|\Lambda_c|$ are the total number of attributes values in each subset of Λ and the intersection of any two domains is empty.

Based on these domains of attributes, given a vertex, say v_i in G, its corresponding attributes values, which are subsets of are subsets of Λ_p, Λ_f, and Λ_c, can be represented as: $\Lambda^i = \{\Lambda^i_p, \Lambda^i_f, \Lambda^i_c\}$.

It should be noted that not all the attributes information is used in the proposed algorithm because some of the attributes may bring the bias of indicating a protein belongs to a particular protein complex. Thus, those attributes which might give clues to the affiliation of protein complexes to which proteins belong are removed from the domain of Λ_c. For instance, GO:0005680 shown in Table 1 might allow one to directly conclude Q08683 belongs to Anaphase Promoting complex. Hence, attributes like GO:0005680 will not be used by TBPCI for avoiding the probable bias discussed above.

It is also noted that the number of attributes from different domains that are associated with a vertex, say v_i, might be different. In general, $|\Lambda^i_p| \neq |\Lambda^i_f| \neq |\Lambda^i_c|$ for each vertex in G. Even sometimes the number of attributes from a particular domain might be equal to zero because the attribute information is missing or removed. As for different vertices, say v_i and v_j, the attributes from the same domain are always different, i.e. $|\Lambda^i_p| \neq |\Lambda^j_p|$, $|\Lambda^i_f| \neq |\Lambda^j_f|$ or $|\Lambda^i_c| \neq |\Lambda^j_c|$. TBPCI will perform the task of protein complex discovery using the given PPI network data G.

3 THE DETAILS OF THE APPROACH

Given a PPI network G, TBPCI performs the task of protein complex identification in the following steps. First, the Degree of Connectedness (σ) and the Degree of Attribute Association (θ) between pairwise proteins in G are computed. Second, a weighted graph **W** is generated by TBPCI in which each entry measures the strength that a pair of proteins can be bounded together. At last, a Breadth-First-Search method will be used for further detecting a number of clusters from **W**. These clusters are seen as the protein complexes identified by TBPCI.

TABLE 1
ATTRIBUTES VALUES OF THE PROTEIN WITH UNIPROT ID Q08683

Biological Processes (Λ_p)	{ GO:0007067, GO:0007049, GO:0016567, GO:0031497, GO:0031145, GO:0051301}
Molecular Functions (Λ_f)	{ GO:0004842}
Cellular Components (Λ_c)	{ GO:0005634, GO:0005680#}

#: Attributes like GO:0005680 is excluded from the experiments since it may bring some bias on inferring it belongs to anaphase promoting complex directly.

3.1 Computation of degree of connectedness and attribute association

Given the PPI network G constructed in the last section, TBPCI needs to obtain the Degree of Connectedness and Degree of Attribute Association, σ and θ for pairwise proteins in the PPI network before identifying protein complexes. The information on σ and θ can be easily obtained by TBPCI, based on the data in G.

The *Degree of Connectedness* (σ) quantifies the extent that the common connected proteins are shared by two pairwise proteins, say v_i and v_j. It is defined as a Jaccard index similarity method:

$$\sigma_{ij} = \frac{\left|e_{i+} \cap e_{j+}\right| + e_{ij}}{\left|e_{i+} \cup e_{j+}\right| - e_{ij}} \quad (1)$$

where 1) e_{i+} is the set of edges that connect v_i and others; 2) the symbol $|.|$ means the cardinality of a given set; 3) e_{ij} is equal to 1 if v_i and v_j are connected in G. For each element in σ, say σ_{ij}, it ranges between 0 and 1. A higher value of that means v_i and v_j share more common connected vertices in G. This indicates a higher similarity between v_i and v_j from the structure perspective.

To measure correlation between pairwise proteins from the property of attribute information, TBPCI computes the Degree of Attribute Association (θ). Recently, some algorithms for protein complex identification, such as approaches in [19], [20], [21], [22], are proposed based on identifying protein complexes whose proteins perform the homogeneous functions. Having deeply looked into different sets of real PPI network data, we found that those proteins with functional homogeneity belongs to the same protein complexes was not always the truth. Let the molecular function positive regulation of transcription from RNA polymerase II promoter (GO:0045944) be an example. In some real PPI network data, such as the one described in [23], the number of proteins that perform the above molecular function is 250. Apparently, it is a relatively high number. As a result of using functional similarity as the measure of identifying protein complexes, these 250 proteins might be located into the same complex. Therefore, a probable outcome of such a grouping method, is a lower rate of identification. Moreover, there is another observed fact that some known protein complexes are constituted of proteins which perform very different functions. For example, one protein with ID P50874 of nuclear origin recognition complex (GO: 0005664) performs very different functions, compared with other proteins in that protein complex. As a result, those algorithms based on functional homogeneity probably cannot identify such protein complexes whose proteins perform different functions.

Given such observed facts, a measure that can weight how frequently the attributes of a pair of proteins interact with each other is used by TBPCI when determining whether the two proteins have a higher correlation when considering the perspective of attribute information.

When measuring the correlation of attributes of two proteins, say $a_{fk} \in \Lambda_f$ ($1 \leq k \leq |\Lambda_f|$) and $a_{pm} \in \Lambda_p$ ($1 \leq m \leq |\Lambda_f|$) which are annotated with two vertices, say v_i and v_j, existing measures assigns zero or one based on whether the

two attributes are different. But, such a similarity-based measure might not be effective on indicating the extent of correlation of the attributes, especially when we consider how strongly the two attributes are correlated. Instead of using the existing similarity measure, TBPCI uses the θ measure.

To explain how θ weights the magnitude of correlation between two attributes values from two vertices, let the relationship $a_{fk} \in \Lambda_f$ $(1 \le k \le |\Lambda_f|)$ and $a_{pm} \in \Lambda_p$ $(1 \le m \le |\Lambda_p|)$ be an example. What we would like to know is to what extent the attributes values a_{fk} and a_{pm} can imply each other. If such a relationship can be represented as a numeric value, then whether or not a correlation between a_{fk} and a_{pm} can be easily measured. To measure such a relationship, the difference between the frequency that two attributes values are interacted ($o(a_{fk}, a_{pm})$) and the expected frequency that two attributes values are interacted ($e(a_{fk}, a_{pm})$) can be used. In other words, we would like to know the difference between $o(a_{fk}, a_{pm})$ and $e(a_{fk}, a_{pm})$, that is the difference between the observed frequency and expected frequency of two attributes values are connected.

To determine whether or not $o(a_{fk}, a_{pm})$ and $e(a_{fk}, a_{pm})$ are significantly different, we define a statistical measure $sig(a_{fk}, a_{pm})$ as follows:

$$sig(a_{fk}, a_{pm}) = \frac{o(a_{fk}, a_{pm}) - e(a_{fk}, a_{pm})}{\sqrt{e(a_{fk}, a_{pm})(1 - \frac{o(a_{fk}, +)}{|E|})(1 - \frac{o(a_{pm}, +)}{|E|})}} \quad (2)$$

where $o(a_{fk}, +)$ is the total number of interacting protein pairs that connect with the proteins characterized by a_{fk}, $e(a_{fk}, a_{pm})$, which can be determined by $o(a_{fk}, +)$ $o(a_{pm}, +)/|E|$ with $|E|$ being the total number of interacting protein pairs, is the expected number of protein pairs characterized by a_{fk} and a_{pm} respectively. As a result, whether or not $o(a_{fk}, a_{pm})$ and $e(a_{fk}, a_{pm})$ are significantly different can be represented by the value of *sig*. A higher value of it means a higher probability that a_{fk} and a_{pm} are implied each other. In previous works, *sig* has been shown to follow a Gaussian distribution approximately [24], [25]. Hence, if *sig* is equal or larger than 1.96, we can conclude that $o(a_{fk}, a_{pm})$ and $e(a_{fk}, a_{pm})$ are significantly different and the occurrence that a_{fk} and a_{pm} are connected has a confidence level of 95 percent. Otherwise, we can conclude that the correlation between two attributes values is insufficiently significant.

Having obtained all the values that weight each pairwise combination of attributes values, we can determine whether a pairwise attributes values are significantly correlated as (3) shows:

$$sc(a_{fk}, a_{pm}) = \begin{cases} 1, if\ a_{fk}\ and\ a_{pm} \\ \quad are\ significantly\ correlated \\ 0, otherwise \end{cases} \quad (3)$$

Once whether or not a pairwise combination of attributes values is significantly correlated has been decided, TBPCI can compute the *Degree of Attribute Association* between each pair of proteins (θ_{ij}), based on the attribute values annotated with them. θ is defined as:

$$\theta_{ij} = \begin{cases} \dfrac{\sum_{m}\sum_{k,l} sc(a^i_{mk}, a^j_{ml})}{\left|\Lambda^i_{p+f+c}\right| \times \left|\Lambda^j_{p+f+c}\right|} & if\ \sigma_{ij} > 0 \\ 0 & otherwise \end{cases} \quad (4)$$

where $m \in (p, f, c)$, $|\Lambda^i_{p+f+c}|$ is the total number of attributes values assigned to v_i and a^i_{mk} is the kth attribute value from subset Λ^i_m. θ is a cosine-based measure and it ranges from 0 to 1. A higher value of θ means a higher proportional pairwise combination of significantly correlated attributes values existing between two vertices. And it is concluded that the two vertices have a stronger association when considering their attribute information. It should be noted that, whether or not pairwise attributes are significantly correlated might be different from one PPI network to another because the computation of $sig(a_{fk}, a_{pm})$ is based only on the PPI network which is used with the identification of protein complexes. After obtaining all the values of θ, TBPCI uses values of θ when the corresponded σ is larger than 0. TBPCI makes use of two matrices, $\mathbf{S}$ and $\mathbf{A}$ to store the obtained values of σ and θ between pairwise vertices in G. TBPCI considers to find optimal weights for those pairwise vertices with a larger-than-zero σ when identifying protein complexes in a PPI network. The found weight measures how strong two pairwise proteins can be bounded together.

3.2 Finding optimal weights between pairwise proteins

3.2.1 Objective function and updating method

Given the obtained information on $\mathbf{S}$ and $\mathbf{A}$ and the PPI network data G, TBPCI attempts to seek optimal weights that may quantify the strength that each pair of proteins can be considered as bounded together. Such weight is named as *Degree of Boundedness* (w). To complete the task, TBPCI formulates the following optimization problem:

$$\begin{aligned} &Maximize \\ &O(\mathbf{W}, \mathbf{F}) = \mathbf{tr}\left(\mathbf{W}_{\mathbf{S}}^{\mathrm{T}} \mathbf{W}\right) + \alpha \mathbf{tr}\left(\mathbf{W}_{\mathbf{A}}^{\mathrm{T}} \mathbf{W}\right) - \left|\mathbf{W}\right|_F^2 - \left|\mathbf{F}\right|_F^2 \\ &\mathbf{W}_{\mathbf{S}} = \mathbf{W} \circ \mathbf{S},\ \mathbf{W}_{\mathbf{A}} = \mathbf{F} \circ \mathbf{A} \\ &subject\ \ to\ \ 0 \le \mathbf{W} \le 1, 0 \le \mathbf{F} \le 1 \end{aligned} \quad (5)$$

where 1) $\mathbf{W}$ is the matrix in which the variables represent the Degree of Boundedness between pairwise vertices, 2) $\mathbf{F}$ is the matrix in which the variables represent the weight between pairwise vertices when the attribute association is considered only, 3) the symbol "$\circ$" means the Hadamard product of two matrices, 4) a is a parameter that is used to constrain the effect of attribute association within the optimization process. The objective value of (5) aggregates the Degree of Boundedness that each pair of vertices in G with a larger-than-zero σ. Apparently, those pairwise vertices with relatively higher values of σ and θ should be assigned larger ws. Also, as the trend of overfitting (5) is penalized by the Frobenius norm of $\mathbf{W}$ and each variable in $\mathbf{W}$ is constrained to be no larger than 1, (5) can be optimized only when all the variables in $\mathbf{W}$ are assigned with appropriate values. However, obtaining appropriate Degrees of Boundedness between all possible pairwise vertices cannot be achieved immediately. In or-

der to obtain optimal Degrees of Boundedness, TBPCI uses an iterative method that constantly updates the values in **W** and **F** till the objective function achieves convergence. The updating methods for **W** and **F** used by TBPCI are

$$\begin{aligned}
&\textit{keeping } \mathbf{F}: \\
&w_{ij} \leftarrow \begin{cases} \left[\mathbf{W}+\eta\Delta\mathbf{W}\right]_{ij} & \textit{if } 0 \le \left[\mathbf{W}+\eta\Delta\mathbf{W}\right]_{ij} \le 1 \\ \Omega(\left[\mathbf{W}+\eta\Delta\mathbf{W}\right]_{ij}) & \textit{otherwise} \end{cases} \quad (a) \\
&\Delta\mathbf{W} = 2\mathbf{W}_{\mathbf{S}} + \alpha\mathbf{W}_{\mathbf{A}} - 2\mathbf{W} \\
&\textit{keeping } \mathbf{W} \\
&f_{ij} \leftarrow \begin{cases} \left[\mathbf{F}+\delta\Delta\mathbf{W}_{\mathbf{A}}\right]_{ij} & \textit{if } 0 \le \left[\mathbf{F}+\delta\Delta\mathbf{F}\right]_{ij} \le 1 \\ \Omega(\left[\mathbf{F}+\delta\Delta\mathbf{F}\right]_{ij}) & \textit{otherwise} \end{cases} \quad (b) \\
&\Delta\mathbf{F} = \alpha\left(\mathbf{W}\circ\mathbf{A}\right) - 2\mathbf{F}
\end{aligned} \tag{6}$$

where η and δ are the step lengths for each iteration of updating for **W** and **F**, respectively. By alternatively updating **W** and **F** following the rules (6a) and (6b), the objective function in (5) will be lead to the convergence of the local optima.

3.2.2 Convergence analysis

Whether the proposed objective function is convergent is essential for TBPCI to identify protein complexes in the PPI network. Here we will give a detail proof for the convergence of the objective function (5), using the updating method in (6). First, the proposed objective function is proved to be convergent without constraint. And then, we will proof that the objective function is also convergent when the box constraints, i.e. $0\le\mathbf{W}\le1$ and $0\le\mathbf{F}\le1$ are used by identifying the relationship of convergence between the above two types of constraints.

Convergence proof of the updating rule for ***W***

For the proof of convergence of (5) by updating **W** following the rule in (6a), it is equivalent to show that, keeping **F** unchanged, $O(\mathbf{W}^{t+1}, \mathbf{F}) \ge O(\mathbf{W}^{t}, \mathbf{F})$ after updating **W** following (6a) in each iteration. Since the updating rule is elementwise, it is sufficient to show for any element w_{ij}, $O(w_{ij})$ is non-decreasing, under the updating rule in (6). Therefore, we have

$$\begin{aligned}
&O(w_{ij}^{t+1}) = O(w_{ij}^{t} + \eta\Delta w_{ij}^{t}) \\
&\Delta w_{ij}^{t} = \left[2\mathbf{W}_{\mathbf{S}} + \alpha\mathbf{W}_{\mathbf{A}} - 2\mathbf{W}\right]_{ij}
\end{aligned} \tag{7}$$

Here we assume that the step length η is a small positive scalar which is smaller than 1 and near to zero. Hence, we can conclude that w_{ij}^{t+1} is near to w_{ij}^{t}, after the updating. To investigate the local information near to w_{ij}^{t}, we use the Taylor series expansion to rewrite O with respect to w_{ij} as

$$\begin{aligned}
&O(w_{ij}^{t+1}, w_{ij}^{t}) \\
&= O(w_{ij}^{t}) + \frac{\partial O}{\partial w_{ij}^{t}}(w_{ij}^{t+1} - w_{ij}^{t}) + \frac{1}{2}\frac{\partial^2 O}{\partial\left(w_{ij}^{t}\right)^2}(w_{ij}^{t+1} - w_{ij}^{t})^2
\end{aligned} \tag{8}$$

Since w_{ij}^{t+1} is near to w_{ij}^{t}, the value of (8) can be seen as an approximant of (5) after updating w_{ij} from iteration t to t+1. If the sum of the latter two components in (8) is non-negative, we can verify that O is non-decreasing when updating w_{ij} according to (6a). Because w_{ij} is any element in **W**, it also means O is finally convergent with respect to **W**. It is noted that (8) can also be written as

$$O(w_{ij}^{t+1}, w_{ij}^{t}) = O(w_{ij}^{t}) + \frac{\partial O}{\partial w_{ij}^{t}}\eta\Delta w_{ij}^{t} + \frac{1}{2}\frac{\partial^2 O}{\partial\left(w_{ij}^{t}\right)^2}\eta^2\left(\Delta w_{ij}^{t}\right)^2 \tag{9}$$

By replacing Δw_{ij}, first-order and second-order partial derivative of O, the last two components can be written as

$$\begin{aligned}
&\frac{\partial O}{\partial w_{ij}^{t}}\eta\Delta w_{ij}^{t} + \frac{1}{2}\frac{\partial^2 O}{\partial\left(w_{ij}^{t}\right)^2}\eta^2\left(\Delta w_{ij}^{t}\right)^2 \\
&= \left(\frac{\partial O}{\partial w_{ij}^{t}}\right)^2\left(\frac{1}{2}\eta^2\frac{\partial^2 O}{\partial\left(w_{ij}^{t}\right)^2} + \eta\right) \\
&= \left[2\mathbf{W}_{\mathbf{S}} + \alpha\mathbf{W}_{\mathbf{A}} - 2\mathbf{W}\right]_{ij}^2\left[\left[\mathbf{S}-\mathbf{1}\right]_{ij}\eta^2 + \eta\right]
\end{aligned} \tag{10}$$

Hence the sum of last two components in (8) is equal to the product of two scalar shown in (10). Obviously, the first scalar is non-negative. According to our assumption that the step length η is a small positive scalar which is smaller than 1 and near to zero, and $s_{ij}\le1$, we conclude η^2 is smaller than η and $[\mathbf{S}\text{-}\mathbf{1}]_{ij}\eta^2+\eta>0$. Thus, we have

$$\begin{aligned}
&O(w_{ij}^{t+1}, w_{ij}^{t}) - O(w_{ij}^{t}) \\
&= \frac{\partial O}{\partial w_{ij}^{t}}(w_{ij}^{t+1} - w_{ij}^{t}) + \frac{1}{2}\frac{\partial^2 O}{\partial\left(w_{ij}^{t}\right)^2}(w_{ij}^{t+1} - w_{ij}^{t})^2 \ge 0
\end{aligned} \tag{11}$$

This means $O(w_{ij}^{t+1})$ is non-decreasing and $O(w_{ij}^{t+1}) = O(w_{ij}^{t})$ only when O converges to the local optima. Next, we will show the proposed objective function can also achieve its local optima when the box-constraint $0\le\mathbf{W}\le1$ is active. Based on the KKT condition, when (5) is convergent with respect of **W**, we have the following

$$\begin{cases}
w_{ij} = 1 & \dfrac{\partial O}{\partial w_{ij}^{t}(w_{ij}^{t}=1)} \ge 0 \\
0 < w_{ij} < 1 & \dfrac{\partial O}{\partial w_{ij}^{t}} = 0 \\
w_{ij} = 0 & \dfrac{\partial O}{\partial w_{ij}^{t}(w_{ij}^{t}=0)} \le 0
\end{cases} \tag{12}$$

It is apparent that (5) is convergent when w_{ij} satisfies the second case in (12) since its first order partial derivative at w_{ij} is zero and we have proved that it is convergent when **W** is updated following (6a). Hence, it is essential to investigate whether (5) is convergent when w_{ij} achieves the upper or lower boundary. Assume w_{ij} is any element in **W** and it equals to 1, the first-order partial derivative of (5) about w_{ij} at 1 is larger than zero. This indicates (5) achieves its local optima when $w_{ij}>1$. We also assume that there is another point w'_{ij}, that is within the constrained area, near to w_{ij} and leads (5) to a larger objective value. Based on the above assumptions, we have

$$\begin{aligned}
&O(w'_{ij}, w_{ij}) - O(w_{ij}) \\
&= \frac{\partial O}{\partial w_{ij}}(w'_{ij} - w_{ij}) + \frac{1}{2}\frac{\partial^2 O}{\partial\left(w_{ij}\right)^2}(w'_{ij} - w_{ij})^2 \\
&= (w'_{ij} - 1)\left[\left(\mathbf{S}-\mathbf{1}\right)_{ij}\left(1 + w'_{ij}\right) + \alpha\mathbf{W}_{\mathbf{A}ij}\right] > 0
\end{aligned} \tag{13}$$

It is noted that $(w'_{ij}\text{-}1) < 0$. Thus, the above inequality holds only when the second component is smaller than zero. But, by comparing the second component of the inequality with the first order partial derivative of (5)

about w_{ij}, we have

$$(\mathbf{S}-\mathbf{1})_{ij}(1+w'_{ij})+\alpha\mathbf{W}_{\mathbf{A}ij} > 2(\mathbf{S}-\mathbf{1})_{ij}+\alpha\mathbf{W}_{\mathbf{A}ij} \quad (14)$$

As we have assumed that the first-order partial derivative of (5) about w_{ij} at 1 is larger than zero, the result obtained in (13) contradicts to the assumption. Therefore, for the first case in (12), (5) converges with respect to w_{ij}=1. As w_{ij} is any element in **W**, (5) is convergent with respect to **W**.

For the case that w_{ij} at the lower boundary, we have the following assumptions before proving its convergence: w_{ij} is any element in **W** and it equals to 0, the first-order partial derivative of (5) about w_{ij} at 0 is smaller than zero. This indicates that (5) achieves its local optima at somewhere w_{ij} <0. We also assume that there is another point w'_{ij}, that is within the constrained area, near to w_{ij} and leads (5) to a larger objective value. Based on the above assumptions, we have the following

$$O(w'_{ij}, w_{ij})-O(w_{ij})=\frac{\partial O}{\partial w_{ij}}w'_{ij}+\frac{1}{2}\frac{\partial^2 O}{\partial(w_{ij})^2}w'^{2}_{ij}>0 \quad (15)$$

As w'_{ij}>0, the second order partial derivative of (5) about w_{ij}=0 is non-positive, (15) holds only when the first-order partial derivative of (5) about w_{ij} at 0 is larger than zero. But this contradicts to the assumption that the first-order partial derivative of (5) about w_{ij} at 0 is smaller than zero. Therefore, for the third case in (12), (5) converges with respect to w_{ij}=0. As w_{ij} is any element in **W**, (5) is convergent with respect to **W**.

In a word, following the updating rule for **W** in (6a), the objective function in (5) will converge to its local optima with respect to **W**.

*Convergence proof of the updating rule for **F***

As the proof of convergence of (5) following the updating rule for **F** is very similar to that of **W**. We don't present the proof in detail due to the space limitation. The updating rule for **F** in (6) ensures the objective function in (5) is non-decreasing and it can achieve its local optima when the box constraint 0≤**F**≤1 is active. Therefore, by keeping one vector of variables unchanged and updating the other, we have

$$O(\mathbf{W}^0,\mathbf{F}^0)\le O(\mathbf{W}^1,\mathbf{F}^0)\le O(\mathbf{W}^1,\mathbf{F}^1)\le\ldots\le O(\mathbf{W}^*,\mathbf{F}^*) \quad (16)$$

This non-decreasing updating will stop till the objective function (5) converges to the local optima $O(\mathbf{W}^*, \mathbf{F}^*)$.

Algorithm 1: Seeking optimal Degree of Boundedness

Input: **S**, **A**, a, η, δ, τ, *MaxIteration*
Output: **W**, **F**

Randomly initialize **W** and **F**;
Compute objective value
t←1;
do
{
oW←**W**;
update **W**←**W**+η**ΔW** according to rule (6a);
update **F**←**F**+δ**ΔF** according to rule (6b);
compute objective value;
t←t+1;
}while((**W**-**oW**)^T(**W**-**oW**) >τ and t< *MaxIteration*);

return **W**, **F**;

Fig. 1. Pseudo codes of the optimization process

3.2.3 Stopping criterion for the optimization process

As the rate of updating becomes lower as **W** and **F** approach to $\mathbf{W}^*$ and $\mathbf{F}^*$, we use the following criterion to determine whether the optimization process stops at after a certain number of iterations:

$$(\mathbf{W}^t-\mathbf{W}^{t-1})^{\mathbf{T}}(\mathbf{W}^t-\mathbf{W}^{t-1})\le\tau \quad (17)$$

where τ is a positive number to tighten or loose the minimum rate of updating **W** that satisfies the stopping condition.

3.2.4 Summary of the optimization process

Based on the above description on finding the optimal Degree of Boundedness between each pair of proteins in the PPI network, the approach can be summarized as the pseudo codes shown in Fig. 1.

3.3 Identifying protein complexes in the weighted graph

Having obtained the optimal **W**, TBPCI sets all those variables in **W** to 0 if the corresponding vertices in G are not connected. After that, **W** can be considered to be a weighted graph with each element (w_{ij}) in it representing the strength that vertex v_i and v_j are bounded together. w_{ij} is larger than zero when there is an edge between v_i and v_j in G and they have a certain degree of boundedness. A higher value of w_{ij} means a higher boundedness that v_i and v_j can be grouped together. TBPCI uses **W** to complete the task of identifying protein complexes in the PPI network.

Given **W**, TBPCI performs a further search of graph clusters as protein complexes. To perform the task of protein complex identification, TBPCI uses a Breadth-First-Search (BFS) method to form protein complexes by selecting each protein in the PPI network as a seed. First, it selects an edge with the highest weight hw_i, that connects the seed in **W**, and incorporates both of the two connected vertices v_i and v_j into a set for forming a protein complex; second, based on the weight of the selected edge, TBPCI searches all the neighboring vertices and incorporates those which satisfy the minimum threshold of w. In TBPCI, this threshold is defined as:

$$PC(seed:v_k)=\begin{cases}PC\cup v_m \ \ if\ \ w_{km}\ge\lambda\times hw_i\\ PC\cup\Phi\ \ otherwise\end{cases} \quad (18)$$

where v_k stands for any vertex in the set and v_m is any vertex connecting to v_k. In other words, λ is used to tighten or loose the minimum w. Only vertices sharing connections with ws that are not lower than $\lambda\times hw_i$ can be incorporated into the complex set so that all the proteins in the complex are tightly bounded to each other. The searching in the second step will be terminated till there is no new vertex added. When the above search is finished, TBPCI forms a protein complex constituted by the connected vertices selected in the searching phase. To reduce the redundancy between the identified protein complexes, TBPCI uses another measure, *Maximum Overlapping Score* to finally determine whether the identified protein complex should be incorporated into the set of identified protein complexes. And the *Maximum Overlapping Score* is defined as:

Algorithm 2: Protein complex identification

```
Input:   G, W, min_size
Output:  Protein complex set

process W according to G;
for each vertex v_i in W{
   create linked list visiting;
   create linked list visited;
   find hw_i;
   if(hw_i>0){
   add v_i to visiting;
      }
   while(visiting ≠ Φ){
      v_j←head of visiting;
      delete v_j from visiting;
      add v_j to visited;
      search v_k: neighbors of v_j;
      if(w_jk ≥ hw_i){
       add v_k to visiting;
      }
      }
      if(|visited|≥min_size){
       create protein complex PC_i;
       if(Max_OSPCi<Max_OS){
        add PC_i to Protein complex set;
       }
      }
}
return Protein complex set;
```

Fig. 2. Pseudo codes of identifying protein complexes

$$Max_{OS} = \max \frac{|PC_I \cap PC_S|}{|PC_I \cup PC_S|} \quad (19)$$

where PC_I and PC_S stand for the identified protein complex and any protein complex in the complex set, respectively. When Max_{OS} is larger than some threshold, TBPCI will not incorporate the identified protein complex into the identified complex set. A lower threshold of Max_{OS} used by TBPCI means there are fewer same proteins formulating each protein complex in the identified set. TBPCI will stop forming protein complexes till it traverses all vertices in **W**. In order to explain this BFS method in detail, we give its pseudo codes in Fig. 2.

4 Experimental Results

For performance testing, TBPCI has been tested with five sets of real world PPI network data: Collins [40], Gavin [2], Krogan [28], DIP-Scere [23] and DIP-Hsapi [23]. The data sets of Collins, Gavin and Krogan can be collected from the BioGRID database [29] and the version of 3.2.118 was used in our experiments. For the data sets of DIP-Scere and DIP-Hsapi, they were both collected from the DIP database [23] in April 2013. In particular, the first four datasets are all related to *yeast Saccharomyces cerevisiae*. All the data sets are cleaned by the removal of duplicated and self-connecting interactions. The properties of the data sets after data cleaning are shown in Table 2.

TABLE 2
STATISTICS ON THE USED DATA SETS OF PPI NETWORKS

Data sets	N_V	N_E	N_A	AvgSize (PC)
Collins	1620	9064	2042	
Gavin	1430	6531	2107	6.76
Krogan	2674	7075	3064	
DIP-Scere	4584	20845	4237	
DIP-Hsapi	2523	3053	7031	5.508

N_V, N_E and N_A: number of proteins, interactions and attribute values in the PPI network.
AvgSize (PC): The average number of proteins in each ground truth protein complex.

For the experiments, the attributes information on the proteins were obtained from GO database [30]. As what has been mentioned, all GO terms in *cellular components* which may infer a particular protein belongs to some complex (es) have been excluded. Specifically, there are originally 497, 467, 577, 664, 787 distinct GO terms of *cellular components* in Collins, Gavin, Krogan, DIP-Scere, and DIP-Hsapi respectively. Since some of the GO terms may provide information as to whether or not a protein belongs to a particular protein complex, all these GO terms are removed from the data set. Correspondingly, a total of 226, 223, 261, 316, 456 distinct GO terms of *cellular components*, remain for TBPCI to be used in Collins, Gavin, Krogan, DIP-Scere, and DIP-Hsapi respectively. We report the experimental results when TBPCI uses the remained GO terms in the manuscript, and the results obtained when TBPCI uses all GO terms are presented in the additional file.

For performance evaluation using the four sets of data related to *yeast Saccharomyces cerevisiae*, we compared the protein complexes identified by different algorithms with the known protein complexes as contained in the CYC2008 [31] and the MIP/CYGD [32] databases. The known protein complexes in these two databases were both collected in March 2013. Altogether, there are 408 known protein complexes in CYC2008 and 255 known protein complexes in MIP/CYGD database. After merging the data in the two databases, a total of 296 different known protein complexes containing more than 2 proteins were available for our use for performance analysis. The mean number of proteins in each of these 296 protein complexes is 6.76.

For the DIP-Hsapi data set, the protein complexes identified in it by each algorithm were compared with known protein complexes in the MIPS/CORUM [34] database which contain a total of 1466 different protein complexes. The mean number of proteins in each of these

TABLE 3
PARAMETER SETTINGS OF DIFFERENT APPROACHES

Approach	Parameter	Approach	Parameter
TBPCI	Experiment Trials	PCIA	inflation=1.8, μ=0.7 (default setting)
MCL	inflation = 1.8 (default setting)	GMFTP	K=1000 (default setting)
RNSC	N/A	CFinder	K=3
MCODE	VWP=0.2 (default setting)	COACH	Experiment Trials
CMC	Experiment Trials		

N/A: There is no parameters needed to be set.

TABLE 4
RESULTS OF PRECISION, RECALL, F-MEASURE AND MMR

Data Set	*Approach*	#	*Coverage*	*Precision*	*Recall*	*f-measure*	*MMR*
Collins	TBPCI	392	1194	0.599	0.642	**0.62**	**0.32**
	PCIA	416	1607	0.368	0.591	0.453	0.303
	GMFTP	203	1088	0.596	0.473	0.527	0.283
	MCL	282	1620	0.401	0.493	0.442	0.259
	MCODE	111	862	0.775	0.357	0.489	0.208
	RNSC	353	1480	0.365	0.544	0.437	0.309
	CFinder	114	1160	0.702	0.319	0.438	0.196
	CMC	200	1067	0.545	0.443	0.489	0.265
	COACH	245	1114	0.51	0.481	0.495	0.176
Gavin	TBPCI	404	1191	0.619	0.607	**0.613**	**0.247**
	PCIA	378	1417	0.315	0.438	0.366	0.216
	GMFTP	172	843	0.643	0.421	0.509	0.228
	MCL	177	1430	0.395	0.278	0.336	0.145
	MCODE	66	602	0.682	0.167	0.268	0.099
	RNSC	307	1258	0.322	0.402	0.358	0.221
	CFinder	98	1124	0.561	0.2	0.295	0.115
	CMC	391	946	0.263	0.392	0.315	0.203
	COACH	324	1052	0.43	0.46	0.445	0.219
Krogan	TBPCI	863	2083	0.453	0.762	**0.547**	**0.342**
	PCIA	1022	2633	0.158	0.583	0.248	0.332
	GMFTP	299	1376	0.413	0.492	0.449	0.282
	MCL	514	2674	0.2	0.422	0.272	0.224
	MCODE	75	552	0.693	0.2	0.31	0.109
	RNSC	751	2142	0.169	0.5	0.253	0.298
	CFinder	115	1140	0.496	0.213	0.298	0.139
	CMC	869	1100	0.236	0.592	0.337	0.287
	COACH	349	1056	0.494	0.536	0.514	0.275
Dip-Scere	TBPCI	1405	3307	0.348	0.836	**0.491**	0.349
	PCIA	1497	4445	0.117	0.636	0.198	0.328
	GMFTP	514	2509	0.286	0.605	0.388	0.305
	MCL	691	4579	0.124	0.327	0.18	0.184
	MCODE	60	756	0.383	0.082	0.135	0.046
	RNSC	1376	3772	0.094	0.522	0.16	0.284
	CFinder	192	2143	0.302	0.205	0.244	0.129
	CMC	1389	1763	0.184	0.736	0.295	**0.353**
	COACH	854	1952	0.275	0.697	0.395	0.32
Dip-Hsapi	TBPCI	754	1822	0.336	0.184	**0.237**	**0.067**
	PCIA	706	2179	0.189	0.108	0.138	0.063
	GMFTP	187	806	0.31	0.044	0.077	0.024
	MCL	549	2434	0.171	0.074	0.103	0.044
	MCODE	61	273	0.459	0.019	0.038	0.008
	RNSC	730	1847	0.148	0.086	0.109	0.059
	CFinder	111	515	0.505	0.042	0.077	0.019
	CMC	149	403	0.537	0.061	0.109	0.027
	COACH	151	492	0.629	0.073	0.131	0.029

#: The number of protein complexes identified. Coverage: The number of distinct proteins in the identified protein complexes.

ground truth protein complexes is 5.508.

4.1 Setting up of experiments and performance evaluation

For performance evaluation, TBPCI were compared with a number of different algorithms, including PCIA, MCL, GMFTP, RNSC, MCODE, CFinder, CMC and COACH. When testing these algorithms, we either used the default parameter settings recommended by the authors or tried as many different settings as we can to obtain the best performance with different algorithms. For the case of TBPCI, we set a= 0.9, η=δ=0.05 τ=1e-6, *MaxIter* = 100 and min_size = 3. For λ and Max_{OS}, they are set within the range from 0.1 to 1 with a 0.1 increment while the performance was being tuned. The detailed settings of parameters for all different algorithms are listed in Table 3.

For performance comparison, we mainly used two different measures, *f-measure*, and *Maximum Matching Rate*

(*MMR*). The *f-measure* can be taken as a measure that determines the overall accuracy of the identified protein complexes when each algorithm performs the task of protein complex identification in a data set. The *f-measure* is defined as the follows:

$$f-measure = 2\times\frac{precision\times recall}{precision+recall}$$
$$precision=\frac{TP}{TP+FP},\quad recall=\frac{TP}{TP+FN} \tag{20}$$

The *f-measure* is defined in terms of *precision* and *recall*. In (20), TP is the number of protein complexes whose matching rates (*mr*) are equal to or larger than a particular threshold. FP represents the number of identified protein complexes that do not match any known protein complex or whose matching rates are lower than the threshold. And FN is the number of known protein complexes in the benchmarking set that are not matched with any identified protein complex. To determine the matching rates, *mr*, between an identified protein complex and the benchmarking set, we follow the method used in [42] and [44], and define it as:

$$mr=\max\left(\frac{|r\cap PC_i|^2}{|r|\times|PC_i|}\right) \tag{21}$$

where r stands for the protein complex identified by TBPCI or other algorithms and PC_i stands for protein complexes in the benchmarking database. The symbol | | stands for the function that determines the size of the protein complex that it encloses.

Given this definition, *mr* can be considered as measuring the best matching between a protein complex identified with computational methods and any of the known ones in the benchmarking set. For example, *mr* of an identified protein complex might be 0.625 if its size is 5, and all these 5 proteins are in one known protein complex containing 8 proteins, in the benchmarking set. In our case, the benchmarking set is either the database obtained by merging MIP/CYGD and CYC2008 or the database MIPS/CORUM. As for the minimum threshold of *mr* for determining whether or not an identified protein complex can be counted as one that is accurately predicted, we typically set it to 0.2 [42] [44]. Given this setting, those approaches attempting to partition all proteins into a number of clusters, e.g., MCL, may also obtain satisfying experimental results. For the experimental results using other settings of *mr*, they are included in the additional file.

Unlike the *f-measure* which needs a predetermined matching rate, the *Maximum Matching Rate* (*MMR*) [43] offers a more natural way to compare the identified protein complexes with the known ones. The *MMR* measures how accurately the predicted protein complexes match that with the known complexes. Thus, the magnitude of the measure also reflects the average percentage of proteins in the identified protein complexes that match with those in the known protein complexes. Based on their definitions, *f-measure* and *MMR* can be seen to be complement each other.

Besides evaluating the experimental results by *f-measure* and *MMR*, we also used GO::TermFinder [39] to perform functional enrichment analysis on the protein complexes identified by TBPCI. This analysis helps to evaluate the biological significance of the protein complex identified. Some of the results of the analysis are presented in the following sections.

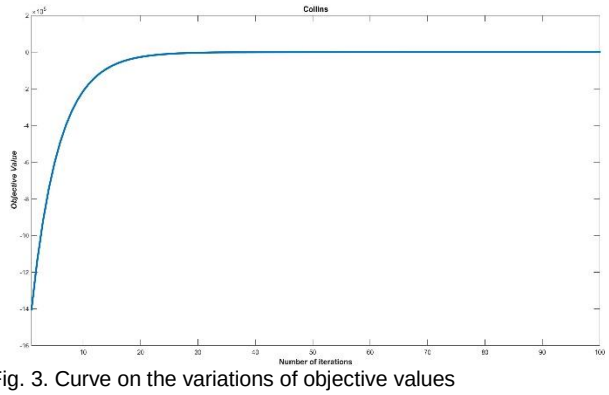

Fig. 3. Curve on the variations of objective values

4.2 Performance analysis

The results of the different experiments performed using the different data sets and algorithms are presented in Table 4 in terms of the values of the *f-measure* and *MMR*. As shown in the Table, according to the *f-measure*, TBPCI performed the best in all the five data sets. When evaluated with *MMR*, TBPCI also obtains a consistently good performance. In terms of the number of identified protein complexes and coverage, it is noted that TBPCI records a relatively higher coverage while maintains a smaller number of identified protein complexes when compared with the other algorithms. Through investigating the structure of **W**, we find that all the interactions in the five testing datasets are assigned with larger-than-zero weights by TBPCI so that TBPCI is able to take into the consideration all the connected proteins when formulating the protein complexes. This is the reseason why TBPCI obtains the mentioned results related to the coverage.

4.3 Convergence of optimization process and sensitivity test on parameters

It is shown earlier that the objective function (5), that TBPCI uses may achieve local optima by iteratively updating **W** and **F** based on (6a) and (6b). In performing the experiments, the variations of the objective values were tracked. In all experiments using different data sets, it was noted that the objective values exhibit similar changing trends. In Fig. 3, the curve showing the changes in

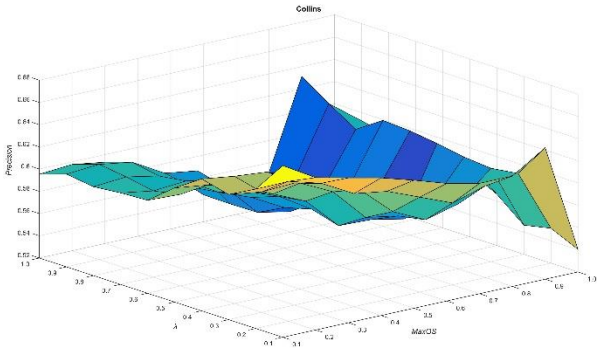

Fig. 4. *Precision* under different combinations of λ and Max_{OS}

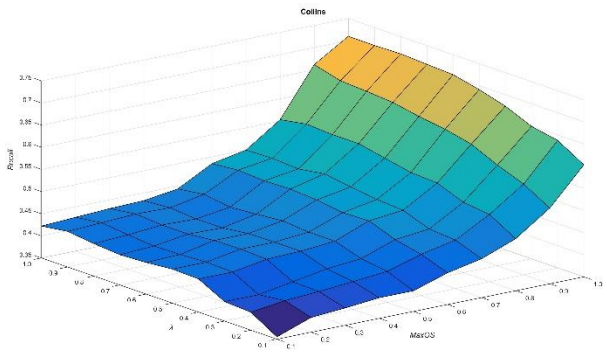

Fig. 5. *Recall* under different combinations of λ and Max_{OS}

objective value which obtained from the data set Collins is displayed. As seen in the figure, TBPCI could approach a convergent objective value within a number of iterations. This also means that the updating rate of **W** becomes less evident as the number of iterations increases. When the updating rate is less than the predefined threshold or TBPCI achieves the maximum number of iterations, the **W** obtained can be considered as optimal Degrees of Boundedness between pairs of proteins in a PPI network.

As mentioned above, the two parameters λ and Max_{OS} used in the BFS search might have some impact on the results of the identified protein complexes. In order to investigate how these two parameters may affect the performance of TBPCI, it is run with each data set using different combination settings of λ and Max_{OS} from 0.1 to 1 with a 0.1 increment every time. The number of distinct protein complexes identified are recorded and the protein complexes identified are evaluated using the measures of *Precision*, *Recall*, *f-measure* and *MMR*. The variations of *Precision*, *Recall*, *f-measure*, *MMR* and the number of identified protein complexes obtained using different λ and Max_{OS} are shown in Fig. 4, 5, 6, 7 and 8.

As it is seen in Fig. 4, the best performance on *Precision* is not obtained when both λ and Max_{OS} are set relatively lower or when they are set at very high values. We can obtain better performance, with respect to *Precision*, if we use an appropriate setting for λ and Max_{OS}. As for the surface describing the variation of *Recall*, it is found that the magnitude of *Recall* becomes higher as λ and Max_{OS} are set at higher values. From Fig. 5 and Fig. 8, it is noted that the number of identified protein complexes shows an incremental trend as λ and Max_{OS} are set at larger values. Under such settings, TBPCI might identify more protein

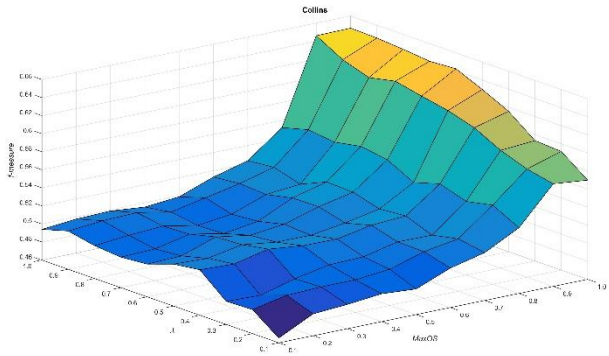

Fig. 6. *f-measure* under different combinations of λ and Max_{OS}

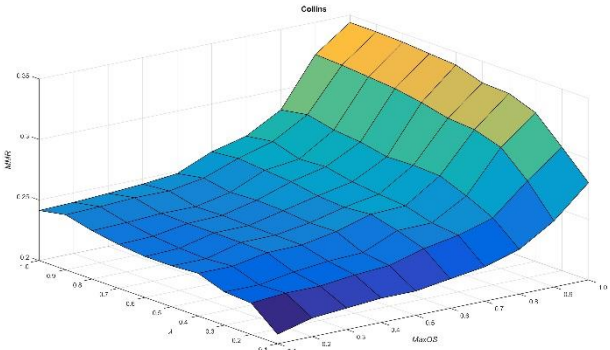

Fig. 7. *MMR* under different combinations of λ and Max_{OS}

complexes that match well with what are known so that a higher *Recall* can be obtained.

As for the surface of *f-measure* shown in Fig. 6, it shows similar trends as with that of *Precision*. Setting λ and Max_{OS} near 0 or 1 might not imply better performance. The surface of *MMR* shown in Fig. 7 shares similar variations with that of *Recall*. This is because more protein complexes identified increases the chance of them matching those of the known complexes. Since we prefer a robust approach based on such measures as the *f-measure* and *MMR*, it is essential to find better settings of λ and Max_{OS} for TBPCI to best perform its task. It also should be noted that the values of λ and Max_{OS} have little impact on the performance of TBPCI (e.g., variation is less than 0.05 in case of the *f-measure*), when they are configured appropriately. This is an indication of the effectiveness of the optimization process used by TBPCI.

4.4 Function enrichment analysis between TBPCI and PCIA

Besides evaluating TBPCI using the *f-measure* and *MMR*, we have also tried to find out whether there was something biologically significant in the identified protein complexes. To do so, we used GO::TermFinder [39] to make a functional enrichment analysis. Provided by SGD [26], GO::TermFinder is a web-based service that can be used for searching significant shared GO terms in the proteins of an identified protein complex. To perform more detailed analysis, we used different thresholds of *p-value*s when analyzing the identified protein complexes. In other words, those GO terms whose *p-values* are equal to or lower than the threshold may be identified as significant ones. We used *Bonferroni correction* to adjust *p-value* of each identified protein complex to make the *p-value* test more accurate. Compared with FDR control, *Bonferroni correction* is a more sensitive and stricter measure for adjusting the obtained *p-value*s [47], and it is widely used in previous works, such as [27] and [42]. Moreover, we also investigated the percentage of genes involved in the detected significant enriched functions for all the identified protein complexes.

To compare the performance between the approaches that make use of attributes information when they identify protein complexes in the PPI network, we also performed similar analysis on the protein complexes identified by PCIA, which has been proved to be a very effective approach to protein complex identification. It should be noted that not all the proteins in these protein com-

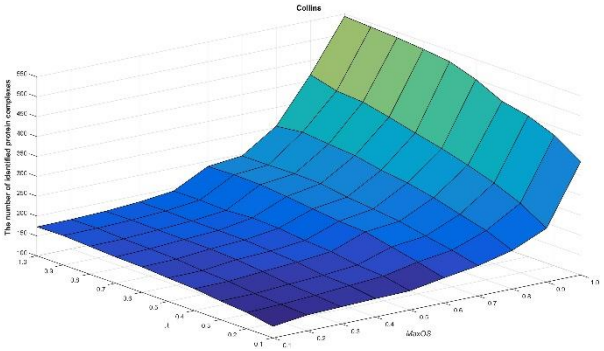

Fig. 8. The number of protein complexes identified by TBPCI using different combinations of λ and Max_{OS}

plexes share significant GO terms that are known and can be found in databases such as MIPS/CYGD, CYC2008 and MIPS/CORUM. In such cases, they can be considered as candidates of real protein complexes due to their statistical significance revealed by the function enrichment analysis. The results of functional enrichment analysis on the protein complexes that are identified by TBPCI and PCIA are summarized in Table 5.

As the table shows, the protein complexes identified by TBPCI obtain a much better performance on the functional enrichment analysis. And, the proportion of significant genes involved in each protein complex identified by TBPCI is also higher than that of PCIA in most datasets. This means there are many more protein complexes identified by TBPCI that have higher possibility to be real protein complexes that are yet to be confirmed, although TBPCI doesn't perform better with the *f-measure* or *MMR*, just like *MMR* in the case of DIP-Scere.

Based on the results of functional enrichment analysis, we can conclude that TBPCI is a very promising approach for protein complex identification.

4.5 Examples of protein complexes identified by TBPCI

We select several protein complexes identified by TBPCI to determine how the consideration of both topology and attribute information may allow interesting protein complexes to be identified.

One example of protein complex that is identified by TBPCI is Kornberg's mediator (SRB) complex (MIPS ID 510.40.20). This protein complex was identified by TBPCI in the data set Krogan. Its structure is shown in Fig. 9. 16 proteins out of 20 that are the same as known protein complexes were successfully identified. For other approaches which also identified protein complexes similar to the SRB complex, such as CMC and COACH, missed some proteins have fewer interactions, such as Q99278, with other proteins. MCODE was not able to successfully identify this protein complex. Although proteins like Q99278 may share fewer interactions with other proteins in the SRB complex, we find that it has a relatively high Degree of Attribute Association with other proteins. For example, there are 10 and 14 GO terms that are associated with Q99278 and P34162, respectively. In Table 6, a number of significantly associated attribute-value pairs belonging to Q99278 and P34162 are listed. As shown in the table, most of attribute values are significantly associated with each other except those that are represented with the GO terms of GO:0003676, and the degree of Attribute Association between these two proteins are determined to be 0.9429. Hence, the Degree of Boundedness assigned by TBPCI is appropriate for Q99278 to be identified as a member of the SRB complex. While the degree of Attribute Association between P19659 and Q02821 (which is not a member of the SRB complex) is only 0.1905. Given this relatively low degree of Attribute Association between these two proteins, it is clear that TBPCI is able to correctly exclude Q02821 from the SRB complex, despite its interaction with some members such as P19659 in the SRB complex. Given the fact that TBPCI may correctly identify proteins like Q99278 and exclude protein like Q02821 from the protein complex that it identifies, there is good potential that TBPCI can be an effective tool for protein complex discovery.

It should be noted that two proteins, Q08278 and P19263 were not correctly identified by TBPCI as they share relatively lower degree of Connectedness and Attribute Association with other identified proteins. Hence, they are assigned with lower degrees of Boundedness by TBPCI and may therefore not satisfy the threshold of λ to be incorporated into the complex. As for the other two proteins, P25453 and P35189, they are not connected to any protein in the SRB complex and are therefore not identified by TBPCI.

TABLE 5
RESULTS OF FUNCTIONAL ENRICHMENT ANALYSIS

Data Set	Approach	Significant Gene	≤1.0E-10	≤1.0E-5	≤1.0E-2
Collins	TBPCI	99.58%	162(42.3%)	273(71.28%)	350(91.38%)
	PCIA	99.84%	88(21.57%)	191(46.81%)	307(75.25%)
Gavin	TBPCI	99.79%	236(58.42%)	336(83.17%)	389(96.29%)
	PCIA	99.79%	80(21.33%)	147(39.2%)	228(60.8%)
Krogan	TBPCI	99.97%	250(29.04%)	529(61.44%)	723(83.97%)
	PCIA	99.91%	87(8.56%)	219(21.56%)	412(40.55%)
DIP-Scere	TBPCI	99.95%	498(35.44%)	905(66.41%)	1214(86.41%)
	PCIA	99.91%	122(8.19%)	274(18.4%)	554(37.21%)
DIP-Hsapi	TBPCI	88.72%	183(24.27%)	541(71.75%)	699(92.71%)
	PCIA	90.66%	86(12.18%)	318(45.04%)	557(78.89%)

X (Y%): the number of identified protein complexes that share significant GO terms under a particular threshold and the proportion of protein complexes sharing significant GO terms in the total protein complexes identified by each algorithm.

TABLE 6
SIGNIFICANTLY ASSOCIATED ATTRIBUTE VALUES BETWEEN P34612 AND Q99278

P34162 / Q99278	GO:0006355	GO:0006357	GO:0045944	…	GO:0003676	GO:0004072
GO:0006355	Y	Y	Y	…	N	Y
GO:0006351	Y	Y	Y	…	Y	Y
…	…	…	…	…	…	…
GO:0001104	Y	Y	Y	…	Y	Y
GO:0001135	Y	Y	Y	…	N	Y

Y (N): two GO terms are (not) significantly associated.

When considering both topology and attribute information of a PPI network, TBPCI may find an appropriate degree of Boundedness that can be used to quantify the relationship and identify those proteins with fewer interactions with other proteins such as Q99278 as a member of some protein complex.

Another example of an identified protein complex is discovered in DIP-Hsapi. In that data set, TBPCI successfully identified the PBAF complex (CORUM ID: 1238). The structure of the PBAF complex is shown in Fig. 10. Seven out of nine proteins in the PBAF complex were detected by TBPCI. What is interesting about the structure of the PBAF complex is O96019. Based on the interactions obtained from the dataset DIP-Hsapi, it should be noted that it is disconnected from all other proteins in PBAF. Given such a reason, TBPCI did not treat it as a member of the PBAF complex.

The protein Q9P2D1 is also worthy of notice as it is not a member of the known PBAF complex, but is grouped into it by TBPCI. In fact, it shares relatively high degrees of Connectedness and Attribute Association with other proteins in the PBAF complex and it is the only protein that interacts with Q92922 which is a member of the PBAF complex but shares no interactions with other proteins. Given the structure and Attribute between Q9P2D1 and other proteins in PBAF complex, we can conclude that Q9P2D1 is possible to be verified as a member of the complex in future laboratory-based experiments.

As for the other missing protein, Q8TAQ2, we find that it is very different from other proteins in the PBAF complex from both the structural and attributional perspectives. This indicates that Q8TAQ2 has fewer interactions with other proteins and its attributes are not well-related to those of the other proteins. Hence, the Degrees of Boundedness computed by TBPCI are smaller than the others in the PBAF complex and Q8TAQ2 was excluded from the complex by TBPCI.

Based on the identified protein complex and the known PBAF complex, it can be said that TBPCI might identify the two missing proteins if there is more evidence showing that Q8TAQ2 and O96019 are related to other proteins in PBAF, either from the topological or the attributional perspectives.

5 CONCLUSION

In this paper, a novel approach to protein complex identification, TBPCI is proposed. TBPCI considers to seek appropriate Degrees of Boundedness between pairwise proteins in the PPI network by taking both topology and attribute information into the consideration. Based on the optimization strategy that can find the optimal Degrees of Boundedness that may quantify the strength that two pairwise proteins can be bounded together, a Breadth-First-Search method is then used to search protein complexes in the weighted graph that is constructed based on the found Degrees of Boundedness. The experimental performance proves that TBPCI using such an optimization strategy can obtain better results when identifying known protein complexes. In future, we will attempt to

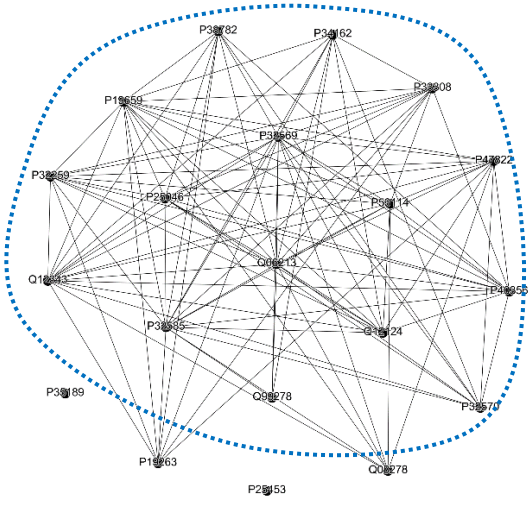

Fig. 9. The structure of Kornberg's mediator (SRB) complex (MIPS ID: 510.40.20) in the MIPS/CYGD database. The proteins successfully identified by TBPCI are circled in the dashed line.

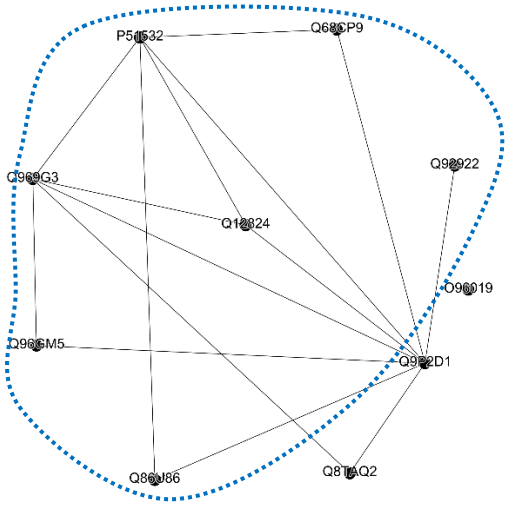

Fig. 10. The protein structure of PBAF complex (CORUM ID: 1238) in MIPS/CORUM database. Proteins successfully identified by TBPCI are in the dashed line.

improve the efficiency of TBPCI, develop some approaches discovering overlapping protein complexes in the PPI network.

REFERENCES

[1] V. Spirin and L. A. Mirny, "Protein Complexes and Functional Modules in Molecular Network," *Proc. Nat'l Academy of Sciences, USA*, vol. 100, no. 21, pp. 12123-12128, 2003.

[2] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Nouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, "Proteome survey reveals modularity of the yeast of the yeast cell machinery," *Nature*, vol. 440, pp. 631-636, 2006.

[3] A. Gavin et al., "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes," *Nature*, vol. 415, no.6868, pp. 141-147, 2002.

[4] Y. Ho et al., "Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry," *Nature*, vol. 415, no. 6868, pp. 180-183, 2002.

[5] T. Deisboeck and J. Y. Kresh, *Complex systems Science in BioMedicine*. Springer, 2006.

[6] A.H.Y. Tong, B. Drees, G. Nardelli, G.D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C.W.V. Hogue, S. Fields, C. Boone, and G. Cesareni, "A combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules," *Science*, vol. 295, no. 5553, pp. 321-324, 2002.

[7] G. Bader and C. Hogue, "An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks," *BMC Bioinformatics*, vol. 4, article 2, 2003.

[8] S. v. Dongen, "Graph Clustering by Flow Simulation," PhD thesis, Univ. of Utrecht, The Netherlands, 2000.

[9] S. v. Dongen, "A Cluster Algorithm for Graphs," Technical Report, R 0010, CWI, 2000.

[10] A.D. King, N. Przulj, and I. Jurisica, "Protein Complex Prediction via Cost-Based Clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013-3020, 2004.

[11] X. Ding, W. Wang, X. Peng, and J. Wang, "Mining Protein Complexes from PPI Networks Using the Minimum Vertex Cut," *Tsinghua Science and Technology*, vol. 17, no. 6, pp. 674-681, 2012.

[12] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks," *BMC Bioinformatics*, vol. 7, no. 1, article 207, 2006.

[13] M. Li, J. Chen, J. Wang, B. Hu, and G. Chen, "Modifying the DPClus Algorithm for Identifying Protein Complexes Based on New Topological Structures," *BMC Bioinformatics*, vol. 9, no. 1, article 398, 2008.

[14] B. Adamcsek, G. Palla, I.J. Farkas, I. Dere´nyi, and T. Vicsek,"CFinder: Locating Cliques and Overlapping Modules in Biological Networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021-1023, 2006.

[15] G. Liu, L. Wong, and H.N. Chua, "Complex Discovery from Weighted PPI Networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891-1897, 2009.

[16] M. Wu, X. Li, C. Kwoh, and S. Ng, "A Core-Attachment Based Method to Detect Protein Complexes in PPI Networks," *BMC Bioinformatics*, vol. 10, no. 1, article 169, 2009.

[17] X. Zhang, D. Dai, and X. Li, "Protein Complexes Discovery Based on Protein-Protein Interaction Data via a Regularized Sparse Generative Network Model," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 857-870, May/June 2012.

[18] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, no. 1, pp. 25-29, 2000.

[19] F.M. Couto, M.J. Silva, and P.M. Coutinho, "Measuring Semantic Similarity between Gene Ontology Terms," *Data Knowledge Eng.*, vol. 61, no. 1, pp. 137-152, 2007.

[20] F.M. Couto, M. Silva, and P.M. Coutinho, "Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors," *Proc. ACM Int'l Conf. Information and Knowledge Management* (*CIKM '05*), pp. 343-344, 2005.

[21] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, "Semantic Similarity Measures as Tools for Exploring the Gene Ontology," *Proc. Pacific Symp. Biocomputing.*, pp. 601, 2003

[22] Z. Lei and Y. Dai, "Assessing Protein Similarity with Gene Ontology and Its Use in Subnuclear Localization Prediction," *BMC Bioinformatics*, vol. 7, no. 1, article 491, 2006.

[23] I. Xenarios, L. Salwı´nski, X.J. Duan, P. Higney, S. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303-305, 2002.

[24] S.J. Haberman, "The Analysis of Residuals in Cross-Classified Tables," *Biometrics*, vol. 29, pp. 205-220, 1973.

[25] K.C.C. Chan, A.K.C. Wong, and D.K.Y. Chiu, "Learning Sequential Patterns for Probabilistic Inductive Prediction," *IEEE Trans. Systems, Man and Cybernetics*, vol. 24, no. 10, pp. 1532-1547, Oct. 1994.

[26] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein, "SGD: Saccharomyces Genome Database," *Nucleic Acids Research*, vol. 26, no. 1, pp. 73-79, 1998.

[27] X. Zhang, D. Dai, L. Ouyang, and H. Yan, "Detecting overlapping protein complexes based on a generative model with functional and topological properties," *BMC Bioinformatics*, vol. 15, article 186, 2014.

[28] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St Onge, S. Ghanny, M.H.Y. Lam, G. Butland, A. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili,

and J.F. Greenblatt, "Global Landscape of Protein Complexes in the Yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637-643, 2006.

[29] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A General Repository for Interaction Datasets," *Nucleic Acids Research*, vol. 34, no. Suppl. 1, pp. D535-D539, 2006.

[30] E. Camon, M., E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler," The Gene Ontology Annotation (GOA) Database: Sharin Knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, no. Suppl. 1, pp. D262-D266, 2003.

[31] S. Pu, J. Wong, B. Turner, E. Cho, and S.J. Wodak, "Up-to-Date Catalogues of Yeast Protein Complexes," *Nucleic Acids Research*, vol. 37, no. 3, pp. 825-831, 2009.

[32] H.W. Mewes, D. Frishman, U. Gu¨ ldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Mu¨ nsterko¨ tter, S. Rudd, and B. Weil, "MIPS: A Database for Genomes and Protein Sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 31-34, 2002.

[33] U. Güldener, M. Mu¨ nsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S.J. Wodak, J. Garcı´a-Martı´nez, J.E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J.L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H.W. Mewes, "CYGD: The Comprehensive Yeast Genome Database," *Nucleic Acids Research*, vol. 33, no. suppl. 1, pp. D364-D368, 2005.

[34] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O.N. Doudieu, V. Stümpflen, and H.W. Mewes, "CORUM: The Comprehensive Resource of Mammalian Protein Complexes," *Nucleic Acids Research*, vol. 36, no. suppl. 1, pp. D646-D650, 2008.

[35] J. Wang, G. Chen, B. Liu, M. Li, and Y. Pan, "Identifying Protein Complexes from Interactome Based on Essential Proteins and Local Fitness Method," *IEEE Trans. NanoBioscience*, vol. 11, no. 4, pp. 324-335, Dec. 2012.

[36] G. Liu, C.H. Yong, H.N. Chua, and L. Wong, "Decomposing PPI Networks for Complex Discovery," *Proteome Science*, vol. 9, no. Suppl. 1, article S15, 2011.

[37] W.W.M. Lam and K.C.C. Chan, "Discovering Functional Interdependence Relationship in PPI Networks for Protein Complex Identification," *IEEE Trans. Biomedical Eng.*, vol. 59, no. 4, pp. 899-908, Apr. 2012.

[38] M. Li, X. Wu, J. Wang, and Y. Pan, "Towards the Identification of Protein Complexes and Functional Modules by Integrating PPI Network and Gene Expression Data," *BMC Bioinformatics*, vol. 13, no. 1, article 109, 2012.

[39] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, and G. Sherlock, "GO::TermFinder—Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710-3715, 2004.

[40] S. R. Collins, P. Kemmeren, X. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, and N. J. Krogan, "Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae," *Mol. Cell. Proteomics*, vol. 6(3), pp. 439-450, Mar. 2007.

[41] J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, "Survey: Functional module detection from protein-protein interaction networks," *IEEE Trans. Kowl. Data Eng.*, vol. 26(2), pp. 261-277, Nov. 2012.

[42] L. Hu, and K. C. C. Chan, "Utilizing both topological and attribute information for protein complex identification in PPI networks," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 10(3), pp 780-792, May 2013.

[43] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks", *Nat. Methods*, vol. 9, pp. 471-472, 2012.

[44] M. Li, T. Yu, X. Wu, J. Wang, F. Wu, and Y. Pan, "C-DEVA: detection, evaluation, visualization and annotation of clusters from biological networks," *BioSystems*, vol. 150, pp. 78-86, 2016.

[45] T. Z. Sen, A. Kloczkowski, and R. L. Jernigan, "Functional clustering of yeast proteins from the protein-protein interaction network," *BMC Bioinformatics*, vol. 7, no. 1, article 466, 2006.

[46] T. C. Rindflesch, J. V. Rajan, and L. Hunter, "Extracting molecular binding relationships from biomedical text," in *Proc. 6th Conf. Applied natural language processing*, 2000, pp.188-195.

[47] J. H. McDonald, *Handbook of biological statistics*, vol. 2, Baltimore, MD: Sparky House Publishing, 2009.

Tiantian He received the BEng degree in computer science from North China University of Technology, Beijing, China in 2008, the MSc degree in information systems, and the PhD degree in computer science from the Hong Kong Polytechnic University in 2012 and 2017, respectively. Currently he is a postdoctoral fellow in Department of Computing, the Hong Kong Polytechnic University. His research interests include machine learning, data mining, and bioinformatics.

Keith C.C. Chan received the BMath (Hons.) degree in computer science and statistics in 1984 and the MASc and PhD degrees in systems design engineering in 1985 and 1989, respectively, from the University of Waterloo, Ontario, Canada. Soon after graduation, he worked as a software analyst for the development of multimedia and software engineering tools at the IBM Canada Laboratory in Toronto, Canada. He joined the Hong Kong Polytechnic University in 1994, where he is currently a professor in the Department of Computing. His research interests include bioinformatics, data mining, and software engineering. He has over 200 publications in these areas, and his research is supported both by government research funding agencies and the industry. Chan serves on the editorial board of five journals and has also been serving on the program committees of numerous conferences.