Predicting protein complexes via the integration of multiple biological information

Xiwei Tang1,3, Jianxin Wang∗1, Yi Pan2
1. School of Information Science and Engineering, Central South University, Changsha, 410083, China
2. Department of Computer Science, Georgia State University, Atlanta, GA 30302-4110, USA
3. School of Information Science and Engineering, Hunan First Normal University, Changsha, 410205, China
Email: *corresponding author, jxwang@mail.csu.edu.cn; tangxiwei2010@gmail.com; panj@cs.gsu.edu

Abstract—Protein complexes are a cornerstone of many biological processes and together they form various types of molecular machinery that perform a vast array of biological functions. An increase in the amount of protein—protein interaction (PPI) data enables a number of computational methods for predicting protein complexes. There are a mass of algorithms detecting complexes only consider the PPI data. However, the PPI data from high-throughput techniques is flooded with false interactions. In fact, the insufficiency of the PPI data significantly lowers the accuracy of these methods.

In the current work, we develop a novel method named CMBI to discover protein complexes via the integration of multiple biological resources including gene expression profiles, essential protein information and PPI data. First, CMBI defines the functional similarity of each pair of interacting proteins based on the edge-clustering coefficient (ECC) from the PPI network and the Pearson correlation coefficient (PCC) from the gene expression data. Second, CMBI selects essential proteins as seeds to build the protein complex cores. During the growth process, the seeds' essential protein neighbors and the neighbors whose functional similarity (FS) with the seeds are more than the threshold T will be added to the complex cores. After the complex cores are constructed, CMBI begins to generate protein complexes by attaching their direct neighbors with FS > T to the cores. In addition to the essential proteins, CMBI also uses other proteins as seeds to expand protein complexes. To check the performance of CMBI, we compare the complexes discovered by CMBI with the ones found by other techniques by matching the predicted complexes against the reference complexes. We use subsequently GO::TermFinder to analyze the complexes predicted by various methods. Finally, the effect of parameter T is investigated.

The results from GO functional enrichment and matching analyses show that CMBI performs significantly better than the state-of-the-art methods. It means that it is successful for us to integrate multiple biological information to identify protein complexes in the PPI network.

I. INTRODUCTION

Protein complexes are a cornerstone of many biological processes and together they form various types of molecular machinery that perform a vast array of biological functions. They are assemblies of proteins, which form many interactions with each other and therefore are cohesive and strongly connected to each other in the context of the larger protein interaction network. In the post-genomic era, one of the most challenging tasks is to predict protein complexes from protein—protein interaction network. Many research groups previously used experimental methods to detect protein complexes. However, there are some experimental limitations in these methods. So researchers begin to develop computational approaches for detecting protein complexes.

Recent developments in biotechnology have resulted in an increase in the amount of protein—protein interaction (PPI) data. Pair-wise protein interactions can be modeled as a graph, where vertices are proteins and edges are protein—protein interaction (PPI). Protein complexes correspond to the dense subgraphs of the initial graph [1]. Over the past decade, a wealth of graph clustering algorithms have been proposed and applied to the identification of highly connected nodes in protein interaction graphs [2]–[9]. These methods work well on PPI networks and extracted successfully protein complexes. Nevertheless, it has been noticed that protein interaction data produced by high-throughput experiments are often associated with high false positive due to the limitations of the associated experimental techniques, which may have a negative impact on the complex discovery algorithms.

In order to address that particular question, recent studies concentrate on incorporating gene-expression data to help identify protein complexes in PPI network. In fact, interacting proteins are likely to exhibit similar gene-expression profiles. Studies have shown that genes showing a similar pattern of expression tend to have similar function (guilt by association) [10], [11]. In this direction, many approaches have been proposed [12], [13]. However, there are cases where functionally related genes show dissimilar expression profiles or are inversely co-regulated [14]. Therefore, in order to design a more effective complex discovery algorithm, it is necessary to adopt other biological information.

The work described in this paper aims to detect protein complexes in the PPI network by considering multiple biological information such as essential proteins, gene expression profiles and protein complex’s inherent organization. To accomplish this goal, a simple algorithm called CMBI (Clustering based on Multiple Biological Information) is developed. Specifically, CMBI first defines the functional similarity of two interacting proteins in the PPI network by combining the edge-clustering coefficient (ECC) [15] between the two proteins and the Pearson correlation coefficient (PCC) [16] for the coexpression profiles of pair of genes coding the two proteins. The essential proteins in the PPI network are subsequently selected as
‘seeds’ to grow the complex cores in terms of their essential proteins neighbors or functionally similar neighbors. After the protein complex cores are produced, CMBI builds the protein complexes by including the cores’ functionally similar neighbors into the complex cores. Moreover, CMBI also uses other proteins as seeds to generate protein complexes by virtue of the protein complexes’ inherent organizations [17].

II. METHOD

A. Preliminaries

The edge clustering coefficient [15] is a measure of degree to which edges in a graph tend to cluster together. It can be defined as

$$ECC(x, y) = \frac{Z_{xy}^{(3)}}{\min(k_x - 1, k_y - 1)}$$

where $Z_{xy}^{(3)}$ denotes the number of triangles that include the edge actually in the network, $k_x$ and $k_y$ are degrees of node $x$ and node $y$, respectively. Then, the meaning of $\min(k_x - 1, k_y - 1)$ is the number of triangles in which the edge $ECC(x, y)$ may possibly participate at most. ECC gives a value between 1 and 0 inclusive. PCC [16] is a frequently used coefficient to express similarity between two gene expression profiles. For two sequences of gene expressions such as $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, PCC is estimated by

$$PCC(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the average expression values of gene $X$ and $Y$, respectively. The value of PCC is always between minus one and plus one. We propose a new functional similarity (FS) between two interacting proteins by means of the combination of ECC and PCC:

$$FS(x, y) = ECC(x, y) + PCC(x, y)$$

It can be found that $FS(x, y)$ is more than -1 and less than 2.

Given a PPI network, the goal of our algorithm is to output a set of dense subgraphs. We model the network as a undirected graph $G = (V, E)$, in which a vertex in vertex set $V$ represents a protein and an edge in edge set $E$ represents an interaction between two distinct proteins. The degree of a vertex $v \in V$ is the number of $v$’s neighbors in $G$, written as $\deg(v)$. We define the second-level degree of a vertex $v \in V$ as follows:

$$sdeg(v) = \deg(v) + \sum_{i=1}^{k} \deg(u_i)$$

where $u$ is the neighbor of vertex $v$, $k$ is the number of $v$’s neighbors. The density of $G$, denoted as $\text{den}(G)$, is defined as follows:

$$\text{den}(G) = \frac{2 \times |E|}{|V| \times (|V| - 1)}$$

For a vertex $v \in V$, the neighborhood graph of $v$ contains $v$, all its neighbors and the edges among them. It is defined as $G_v = (V', E')$, where $V' = \{v\} \cup \{u \in V, (u, v) \in E\}$, and $E' = \{(u, v) \in E, v, u \in V'\}$. Besides, we define the neighborhood of a complex core $C = (V_C, E_C)$ as $N(C) = \{u | (u, v) \in E, v \in V_C, u \in V, v \in V_C\}$ where $V_C$ is the vertexes in the core and $E_C$ is the edges among $V_C$. $N_v$ represents the set of $v$’s neighbors where $v \in N(C)$. $|N_v \cap V_C|$ is the number of vertexes in $C$ connected with $v$. Therefore, we use closeness($v, C$) = $|N_v \cap V_C|$ to quantify the closeness between the vertex $v$ and the core $C$.

B. Data sources

Protein interaction data: The yeast PPI data is downloaded from DIP [19], updated on Feb. 28, 2012. The datasets contain 22,570 interactions between 5,023 proteins.

Gene expression data: We download the data from the NCBi Gene Expression Omnibus website [20]. It is available in the form of a 9,335×36 matrix, includes expression profiles of 9,335 probes under 36 different time points, updated on Apr 14, 2011.

Essential gene data: A list of essential proteins of S. cerevisiae are downloaded from MIPS [21], SGD [22], DEG [23] and SGDp [24], which contain 1,285 essential proteins altogether.

C. The CMBI algorithm

CMBI operates in two phases. In the first phase, it uses essential proteins as seeds to grow protein complexes. Specifically, all the essential proteins in the yeast PPI network are curated by comparing protein set $V'$ with the known essential protein set. We acquire a protein set denoted as $\text{Ess}(v)$ which consists of 1156 essential proteins. The protein complex cores are detected from the neighborhood graph $G_v$ of each vertex $v \in \text{Ess}(v)$. Initially, the essential protein seed $v$ is joined in the complex core $C$. Subsequently, $v$’s each direct neighbor $u$ is checked. If $u$ is essential protein or $FS(v, u)$ is more than Threshold $T$, the neighbor $u$ will be add to $C$. After all complex cores from essential protein seeds are built, CMBI begins to detect protein complexes. The neighbors of each complex core $C$ are considered. Every protein $w \in N(C)$ with $FS(w, v) > T \in \text{Ess}(v)$ will be inserted to the complex core $C$ so as to form the complex. Finally, since the number of proteins in the every known complex are more than 1, the predicted complexes only including one protein will be discarded. If a predicted complex consists of all proteins in another predicted complex, the latter will be removed.

At the same time, it can be found that there are a few known protein complexes not containing essential proteins at all. In the second phase, CMBI attempts to discover these
complexes. CMBI constructs a set $H$ by collecting the rest of proteins in the yeast PPI network. The proteins in the set $H$ don’t belong to the complexes expanded from essential proteins. They are sorted descending on their secondary-level degree. CMBI selects the protein with the maximal secondary-level degree in the set $H$ as seed to build the complex core. Next the seed’s neighbors in set $H$ are processed. If the seed’s special neighborhood graph $SG$ only including the seed and its neighbors in set $H$ is dense (with $\text{den}(SG) > 0.7$ [25], [26]), $SG$ is directly predicted as a complex core; otherwise, the seed’s neighbors in set $H$ will be continually removed from $SG$ according to their degrees from low to high until $SG$ is dense. And then the complex core $C$ is generated. The proteins in the core $C$ will be removed from set $H$. Those cores only consist of one protein or two proteins will be discarded. The remaining proteins in $H$ are sorted descending on their secondary-level degree again. Similarly, other cores also are constructed. The core $C$ is grown from its neighborhood $N(C)$. If $\text{closeness}(w, C) > 0.5$ [5] where $w \in N(C)$, $w$ will be added to $C$. After all neighbors in $N(C)$ are handled, the complex is formed. The rest of cores are expanded to complexes in the same way. Algorithm 1 shows the pseudocodes of the CMBI algorithm.

III. RESULTS AND DISCUSSIONS

We have applied the CMBI method on the yeast PPI networks. In this section, we first describe the evaluation methods used in our experiments, and then study the performance of CMBI and the impact of the Threshold $\tau$ on CMBI. CMBI is compared with eight other clustering algorithms: MCODE [3], MCL [2], CFinder [4], CMC [6], COACH [5], SPICI [7], HC-PIN [8] and ClusterONE [9]. The values of the parameters in each algorithm are selected from those recommended by the authors.

Evaluation methods

One evaluation method we use is to match the generated complexes with known complex set [18], and calculate sensitivity ($Sn$), specificity ($Sp$) and F-measure, respectively. We derive 408 typical complexes including two or more proteins from the CYC2008 [18] as the benchmark complex set and use the same scoring scheme used by [3] to determine how effectively a predicted complex matches a reference complex. If two complexes overlap each other, they must share one or more proteins. The Overlap Score ($OS$) of a predicted complex vs. a benchmark complex is then a measure of biological significance of the prediction, assuming that the reference set of complexes is biologically relevant. $OS$ is calculated by using

$$OS = \frac{i^2}{a \times b}$$

where, $i$ refers to the number of proteins shared by a predicted complex and a known complex, $a$ is the number of proteins in the predicted complex and $b$ is the number of proteins in the known complex. If $OS$ is 1, it means that a predicted complex has the same proteins as a known complex. On the contrary, when $OS$ equals to 0, there is not a shared protein between the predicted complex and the known complex [3].

The number of true positives ($TP$) is defined as the number of predicted complexes with $OS$ over a threshold value and the number of false positives ($FP$) is the total number of predicted complexes minus $TP$. The number of false negatives ($FN$) equals the number of known complexes not matched by predicted complexes. $Sn$ and $Sp$ are defined as $TP/(TP + FN)$ and $TP/(TP + FP)$, respectively [3]. F-measure, or the harmonic mean of $Sn$ and $Sp$, can then be used to evaluate the overall performance of the clustering algorithms [27]:

$$F\text{-measure} = \frac{2 \times Sn \times Sp}{Sn + Sp}$$

In addition, we also use the functional enrichment of GO terms (p-values) as an evaluation measure to check the performance of each clustering algorithm. A complex is associated with a known function by determining whether the number of proteins known to be annotated with the function is enriched, as judged by the hypergeometric distribution. The p-value can be used to determine the probability that a given set of proteins is enriched by a given functional group by random chance. In [25], it is used as a criterion to assign each cluster to a known function. The smaller the p-value, the more evidence the clustering is not random. In terms of GO annotations, a group of genes with a smaller p-value is more significant than the one with a higher p-value. Consider a cluster of size $c$, with $m$ proteins sharing a particular annotation $A$. Also assume that there are $N$ proteins in the PPI database, and $M$ of them are known to have annotation $A$. Given that, the probability of observing $m$ or more proteins that are annotated with $A$ out of $N$ proteins is:

$$P = 1 - \sum_{i=0}^{m-1} \frac{M^c \times (N-M)^{c-i}}{N^c}$$

Based on above formulation, a p-value is calculated for each of three ontologies. In the case of multiple annotations from the same ontology, the one with the smaller p-value is assigned to the cluster as functional annotation. That being said, the p-value without any restriction is not enough to label clusters as significant. Hence we use the recommended cutoff value of 0.01 [28] in order to select significant complex within each ontology. A popular software package for evaluating the statistical significance of GO terms represented in a set of genes extracted from a population is GO::TermFinder [29], which calculates p-values (with Bonferroni correction) using above formula. GO::TermFinder accepts a list of genes of interest and returns a list of GO terms with which the genes are associated, with corresponding P-values and FDR values (if desired) associated with the enrichment of these terms in the gene list. In our experiments, the direct use of GO::TermFinder
Algorithm 1 CMBI Algorithm

Input:
- PPI network \( G = (V, E) \);
- essential protein set \( \text{Ess}(v) \);
- gene expression profiles;
- similarity threshold \( T \);

Output:
- set of protein complexes \( SC \) discovered from \( G \);

Description:
1: \( SC = \emptyset; // \) initialization
2: for each vertex \( v \in \text{Ess}(v) \) do
3: \( C = (V_c, E_c); // |V| = \{u \in \text{Ess}(v) \ or \ FS(u, v) > T, u \in G_v\} \)
4: for each vertex \( w \in N(C) \) do \( // N(C) \) includes all direct neighbors of \( C \)
5: if \( FS(w, v) > T \) then \( // v \in C \)
6: insert \( w \) into \( C \);
7: if \( C \subset C_v \) then \( SC = SC \cup \{C\}; \)
8: \( H = V - Q; // \) The set \( Q \) contains all proteins in the complexes grown from the essential protein seeds.
9: sort each \( v \in H \) in descending order according to its secondary-level degree;
10: for each vertex \( v \in H \) do
11: construct the core graph of \( G_v, C = (V_c, E_c); // |V| = \{u \in H \ and \ \text{den}(C) > 0.7, u \in G_v\} \)
12: \( H = H - V_c; \)
13: if \( C > 2 \) then \( // \) discard those cores only including one protein or two proteins
14: for each vertex \( w \in N(C) \) do \( // N(C) \) includes all direct neighbors of \( C \)
15: if \( \text{closeness}(w, C) > 0.5 \) then
16: insert \( w \) into \( C \);
17: \( SC = SC \cup \{C\}; \)
18: output the complexes in \( SC \);

is not convenient for analyzing GO enrichment of a vast amount of complexes uncovered by all kinds of algorithms, because this software package can only handle one module at a time. Therefore, combined with the latest version of this toolkit [30], we have used the Perl language to develop a procedure that can automatically process a large number of functional modules in turn.

A. Comparison with the known complexes

After researching the effect of Overlap Score threshold on number of predicted and matched known complexes, Badger et al. [3] find that the average and maximum number of matched known complexes drops more quickly from zero until an \( OS \) threshold of 0.2 than from 0.2 to 0.9. It indicates that many predicted complexes only have one or a few proteins that overlap with known complexes. An \( OS \) threshold value which falls within the region from 0.2 to 0.3 thus seems to filter out most predicted complexes that have insignificant overlap with known complexes. Table I shows the basic information of predicted complexes by various methods when \( OS \) is set as 0.2. \#PC is the number of complexes identified by each algorithm. \#S is the average size of the complexes detected by each algorithm. MS is the number of complexes in the maximal complex predicted by each algorithm. MKC represents the number of real complexes that match at least a predicted one and MPC is the number of correct predictions which match at least a real complex. PMC is the number of correct predictions perfectly matching the known complexes. In other words, a prediction has the same proteins with the known complex matched by it. As shown in Table I, CMBI detects 793 protein complexes, of which 351 match 161 real complexes. The maximal complex predicted by CMBI contains 114 proteins and the average size of these complexes is 15.80. Besides, CMBI identifies 10 complexes overlapping fully with the known complexes. The properties of complexes discovered by other algorithms also are shown in Table I. Table II shows the matching comparison in terms of \( Sn \), \( Sp \) and F-measure.

On DIP data of \( S. cerevisiae \), the F-measure of CMBI is 0.50, which is higher than 316.67%, 92.31%, 100.00%, 35.14%, 11.11%, 110.00%, 38.89% and 28.21% higher than MCODE, MCL, CFinder, CMC, COACH, SPICi, HC-PIN and ClusterONE, respectively. Our CMBI algorithm can achieve the highest F-measure by providing the second-highest sensitivity and comparable specificity, which shows that CMBI can predict protein complexes very accurately.

B. GO analysis

In many studies, the GO has been used as the 'gold standard' to validate the functional relevance of the obtained network modules. In this subsection, as described by Evaluation methods, we used the GO biological process (BP) annotation to take GO enrichment analysis with our developed analytical tool based on GO::TermFinder software package [30].

TABLE III
BP ANALYSIS OF DISCOVERED COMPLEXES.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>CMBO</th>
<th>MCODE</th>
<th>MCL</th>
<th>CFinder</th>
<th>CMC</th>
<th>COACH</th>
<th>SPICl</th>
<th>HC-PIN</th>
<th>ClusterONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>#SC</td>
<td>713</td>
<td>55</td>
<td>414</td>
<td>122</td>
<td>196</td>
<td>736</td>
<td>297</td>
<td>176</td>
<td>253</td>
</tr>
<tr>
<td>#PC</td>
<td>793</td>
<td>59</td>
<td>928</td>
<td>197</td>
<td>235</td>
<td>902</td>
<td>574</td>
<td>277</td>
<td>371</td>
</tr>
<tr>
<td>Proportion</td>
<td>89.91%</td>
<td>93.22%</td>
<td>44.41%</td>
<td>61.93%</td>
<td>83.40%</td>
<td>81.60%</td>
<td>51.74%</td>
<td>63.54%</td>
<td>68.19%</td>
</tr>
<tr>
<td>P-score</td>
<td>13.95</td>
<td>7.75</td>
<td>5.66</td>
<td>7.19</td>
<td>8.77</td>
<td>7.99</td>
<td>6.13</td>
<td>9.04</td>
<td>8.18</td>
</tr>
</tbody>
</table>

TABLE II
THE MATCHING RESULTS OF EACH ALGORITHM.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Sn</th>
<th>Sp</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMBO</td>
<td>0.57</td>
<td>0.44</td>
<td>0.59</td>
</tr>
<tr>
<td>MCODE</td>
<td>0.07</td>
<td>0.47</td>
<td>0.12</td>
</tr>
<tr>
<td>MCL</td>
<td>0.45</td>
<td>0.38</td>
<td>0.26</td>
</tr>
<tr>
<td>CFinder</td>
<td>0.19</td>
<td>0.38</td>
<td>0.25</td>
</tr>
<tr>
<td>CMC</td>
<td>0.30</td>
<td>0.51</td>
<td>0.37</td>
</tr>
<tr>
<td>COACH</td>
<td>0.63</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>SPICl</td>
<td>0.31</td>
<td>0.21</td>
<td>0.25</td>
</tr>
<tr>
<td>HC-PIN</td>
<td>0.31</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>ClusterONE</td>
<td>0.36</td>
<td>0.42</td>
<td>0.39</td>
</tr>
</tbody>
</table>

The p-values (with Bonferroni correction) of protein complexes predicted by each algorithm are calculated. The detected complexes with corrected p-value < 0.01 [28] are considered to be significant. Maraziotis et al. [13] conclude that the proportion of significant complexes over all identified ones can be used to evaluate the overall performance of various algorithms. In addition to this measure, we also use the average −\( \log(p - value) \) of predicted complexes to check the performance of the prediction algorithms. Table III shows the GO enrichment results of BP. In Table III, SC is the number of significant complexes with \( p \)-value<0.01. PC is the number of protein complexes predicted by each method. P-score is the average \( -\log(p - value) \) of identified complexes. As shown in Table III, there are 713 significant complexes in all 793 complexes discovered by CMBO. CMBO predicts higher proportion (89.91%) of significant complexes than other algorithms excepting MCODE. However, MCODE predicts 59 complexes and correctly matches only 30 known complexes as shown in table I. Moreover, table I also shows that #PC and MKC of MCODE is far fewer than those of other algorithms. More importantly, the P-score of complexes identified by CMBO reaches up to 13.95, which is 80.0%, 146.5%, 94.0%, 59.1%, 74.6%, 127.6%, 54.3% and 70.5% higher than that of complexes detected by MCODE, MCL, CFinder, CMC, COACH, SPICl, HC-PIN and ClusterONE, respectively.

We find a interesting fact that many complexes mined by CMBO don’t match any known complexes but they have very low p-values. Due to the incompleteness of the reference complexes, these complexes may provide potential candidate complexes for biologists to validate.

The results from GO analysis demonstrate that the biological significance of complexes identified by CMBO is much stronger than that of complexes discovered by other algorithms.

C. Effect of the parameter \( T \)

In this experiment, we study the effect of the threshold \( T \) on the performance of CMBO. Figure 1 shows the F-measure of CMBO under different values of \( T \). As shown in figure 1, the F-measure of CMBO increases with the increase of \( T \) when \( T < 1.4 \). The size of each complex decreases as \( T \) increases in the interval since the number of proteins captured by the essential protein seed drops as \( T \) increase. But the F-measure of CMBO is 0.47 and remains unchanged when \( T \geq 1.8 \). It means that the number of proteins belonging to a complex rarely remains the same. In this case, CMBO doesn’t actually consider gene expression information. But even then, it can be seen that the F-measure of CMBO is still higher than that of other algorithms. It demonstrates that it is successful for us to incorporate essential proteins information into CMBO. It can also be found that the consideration of gene expression profiles further improves the performance of CMBO when \( 1.0 \leq T \leq 1.8 \). We recommend that the suitable setting of \( T \) would be in the range i.e., \( T \in [1.2, 1.4] \). In fact, the performance of CMBO doesn’t change significantly in this interval. The parameter \( T \) is set 1.4 in our experiment.

IV. CONCLUSION

In this research, we develop a new technique called CMBO to identify protein complexes in yeast PPI network by integrating other biological resources into the PPI data. CMBO constitutes protein complexes that originate from a kernel protein set (protein complex core) built up from a seed protein. More
specifically, CMBI selects the essential proteins as seeds to generate protein complexes by using gene expression and essential protein information. In addition to the essential proteins, CMBI also chooses other proteins as seeds to construct protein complexes based on the inherent organization of protein complexes. In order to characterize these clusters as protein complexes we check their biological relevance. This is achieved through some criteria such as matching analysis between predicted complexes and known complexes and the functional enrichment analysis of the derived complexes in GO terms. The evaluation and analysis of our predictions demonstrate that CMBI performs very well and outperforms other algorithms.

Since the integration of multiple data sources shows great superiority in predicting protein complexes, we will apply the idea to other research areas such as the identification of functional modules including one or more complexes in future.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grant Nos.61073036, 61003124, the Ph.D. Programs Foundation of Ministry of Education of China No.20090161210073, the Freedom Explore Program of Central South University No.201012200124, High-tech Program of China Hunan Provincial Science and Technology Department No.2010GK3049, 2011GK3138, Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province No.2010212, Hunan Provincial Department of education Science Foundation under Grant No.11C0281, the U.S. National Science Foundation under Grants CCF-0514750, CCF-0646102 and CNS-0831634.

REFERENCES


