JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009

Pp. 133–144
DOI: 10.1089/cmb.2008.01TT

Predicting Protein Complexes from PPI Data: A Core-Attachment Approach

HENRY C.M. LEUNG, QIAN XIANG, S.M. YIU, and FRANCIS Y.L. CHIN

ABSTRACT

Protein complexes play a critical role in many biological processes. Identifying the component proteins in a protein complex is an important step in understanding the complex as well as the related biological activities. This paper addresses the problem of predicting protein complexes from the protein-protein interaction (PPI) network of one species using a computational approach. Most of the previous methods rely on the assumption that proteins within the same complex would have relatively more interactions. This translates into dense subgraphs in the PPI network. However, the existing software tools have limited success. Recently, Gavin et al. (2006) provided a detailed study on the organization of protein complexes and suggested that a complex consists of two parts: a core and an attachment. Based on this core-attachment concept, we developed a novel approach to identify complexes from the PPI network by identifying their cores and attachments separately. We evaluated the effectiveness of our proposed approach using three different datasets and compared the quality of our predicted complexes with three existing tools. The evaluation results show that we can predict many more complexes and with higher accuracy than these tools with an improvement of over 30%. To verify the cores we identified in each complex, we compared our cores with the mediators produced by Andreopoulos et al. (2007), which were claimed to be the cores, based on the benchmark result produced by Gavin et al. (2006). We found that the cores we produced are of much higher quality ranging from 10- to 30-fold more correctly predicted cores and with better accuracy. Availability: *http://alse.cs.hku.hk/complexes/*.

Key words: core protein, PPI network, protein complexes.

1. INTRODUCTION

BIOLOGICAL PROCESSES—such as signal transduction, cell cycle, and replication—involve complicated organization and interactions among protein molecules. Unfortunately, the mechanism for most of these activities is still unknown. Protein complexes are one of the fundamental units of macromolecular organization and play an important role in integrating individual gene products to perform useful cellular functions. For example, $\alpha 3\beta 1$-tetraspanin protein complexes are important in regulating the protrusive activity of the tumor cells (Sugiura et al., 1999); complexes consisting of PDZ proteins have been found

Department of Computer Science, University of Hong Kong, Hong Kong.

to be critical during the establishment of cell-cell adhesions and epithelial cell polarity (Roh et al., 2003). Identifying the component proteins in a protein complex is an essential step towards understanding various biological processes.

Although technologies such as high-pressure liquid chromatography column (HPLC) and tandem mass spectrometry (MS/MS) are available for researchers to analyze protein mixtures, current proteomics technology is still not mature enough to provide an easy and effective method for identifying individual proteins that are present in the protein complexes (McDonald et al., 2006). On the other hand, it is generally believed that proteins inside the same complex usually have more interactions than those not in the same complex. As a large amount of protein interaction data has become available (e.g., DIP [Xenarios et al., 2000], Krogan [Krogan et al., 2006], Gavin [Gavin et al., 2006]), predicting protein complexes from protein-protein interaction data using computational methods provides an alternative approach (e.g., CFinder [Adamcsek et al., 2006], MCODE [Bader et al., 2003], MCL [Dongen, 2000; Enright et al., 2002]).

However, it is not trivial to come up with an effective computational method to predict protein complexes from protein interaction data. Protein interaction data is usually represented by a protein-protein interaction (PPI) network in which the nodes are proteins and an edge between two nodes represents a known interaction between the two proteins. A major difficulty is the absence of an appropriate definition to define which group of proteins should be in the same complex in such a PPI network. Most of the existing approaches rely on the idea that proteins within the same complex would have relatively more interactions. So, these approaches aim at finding dense subgraphs inside the PPI network. They mainly differ on how they define a dense subgraph and the procedure to cluster the nodes into dense subgraphs.

CFinder (Adamcsek et al., 2006) defines a dense subgraph based on the concept of k-clique. A k-clique is a complete subgraph with k nodes. Each dense subgraph is found by merging a series of *adjacent* k-cliques where two k-cliques are said to be adjacent if they share $(k-1)$ nodes. If k is small, say $k = 3$, there is too much noise. We have examined the PPI network for yeast from the Database of Interacting Proteins (DIP) (Xenarios et al., 2000) and found that there are more than 800 3-cliques in the network, but less than 15% are inside known complexes (based on MIPS database [Mewes et al., 2000]). If k is large, the requirement becomes too stringent as it is well-known that there are missing interactions in the PPI network. Thus, some of the complexes, that do not have enough known interactions to form 4-cliques, may be missed. For examples, if k is set to 3, the Golgi Transport Complex (ID: 260.20.40[1]) will be missed as CFinder tries to merge too many noisy 3-cliques into the complex; if k is set to 4, the Transcription Complex (ID: 510.190.40) will be missed as there are not enough 4-cliques inside the complex.

Molecular Complex Detection (MCODE) (Bader et al., 2003) assumes that a protein v should be part of a complex if it has a subset of neighbors with high degree, say k, and there are a lot of interactions (more precisely, the number of known interactions over the total number of possible interactions) among these proteins (including v). It starts by assigning a weight to each protein based on local neighbor density. Then, starting from the protein with the highest weight, it recursively adds neighbors with weights larger than a threshold into the cluster. However, MCODE has a similar problem as CFinder. The assumption that v should have neighbors of high degree and that these neighbors have lots of interactions may not be valid if k is large. In practice, the default value for k is recommended to be 2 in the software. This tends to include noise into the complexes. The COPI complex (ID: 260.30.10) is an example missed by MCODE because too much noise is included in the predicted cluster.

Markov Cluster Algorithm (MCL) (Dongen, 2000; Enright et al., 2002) uses quite a different idea to identify the dense regions. They use the concept of random walks as follows. Assume that a walker starts at an arbitrary protein v, the walker will visit a neighbor with equal probability. If he walks into a dense region, it will be harder to get out of the region. So, by simulating a large amount of random walks, called *flow*, for a long enough period, an underlying cluster structure will be identified. Note that MCL considers the notion of dense subgraph in a relative sense. In other words, the clusters identified may not necessarily have an absolute high density. MCL is able to predict complexes with low density provided the proteins inside the complex have relatively fewer interactions with proteins outside the complex. However, they may miss some complexes with very high density, for example, the Exosome complex (ID: 440.12.10).

[1]The complex ID refers to the complex category used in MIPS database.

Brohee et al. (2006) compared the performance of MCL, MCODE, and the other two approaches, RNSC (King et al., 2004) and SPC (Blatt et al., 1996), and found that MCL performed the best in extracting complexes from PPI networks. However, we found that MCL can only identify 24 out of 81 known complexes (in MIPS) of size greater than or equal to 5 based on the DIP dataset (a match of the known complex is defined to have an accuracy of at least 0.6. There should be room for further improvement. Note that for complexes of size 4 or less, if only protein-protein interaction data is considered, a good method to locate them is unlikely without including a lot of false positives.

All these existing approaches are based on the idea of finding dense subgraphs. To achieve a breakthrough, we may need a deeper understanding of the organization of complexes. Recently, Gavin et al. (2006) have taken a further step to study the organization of protein complexes. They suggested that a complex should consist of two parts: *a core component* and *attachments*. Core proteins are the center of the protein complex that have relatively more interactions among themselves. Each attachment protein binds to a subset of core proteins to form a complex (see Section 2.1). They also found that among the attachment proteins, there should be some proteins, called *modules*, which interact with the core proteins quite a lot (Fig. 1). The proposed concept has been used in Andreopoulos et al. (2007) to identify cores, called *mediators*, which are locally significant proteins that mediate the function of modules. The way they formulate a mediator captures some of the properties of a core in Gavin et al. (2006), so a few of their predicted mediators overlap with the cores found in Gavin et al. (2006). However, not all properties of a core component, in particular, the number of interactions among the proteins inside the core, have been considered. The overall accuracy of using mediators to predict cores is not satisfactory.

In this paper, we try to use the new findings in Gavin et al. (2006) directly to predict protein complexes from the PPI network. The key idea behind our approach consists of three main steps: (1) predict core components; (2) identify attachments for the cores and eliminate insignificant cores; and (3) compute and rank the significance of predicted complexes. Basically, for each disjoint potential core, we define a p-score to evaluate how likely it would be the core component of a complex based on the number of interactions among proteins inside a potential core and the number of interactions between proteins inside the core and their common neighbors (potential attachment proteins). Then, for each identified core, we add all proteins that have interactions with the majority of core proteins in the complex. At the same time, some of insignificant core candidates will be eliminated based on the assumption that only one core is assumed for each complex. Finally, we derive a measure to evaluate the significance of each predicted complex. None of the previous work gives an effective method to rank the predicted complexes. On the other hand, we make use of the concept that proteins inside the same complex should have relatively more interactions than proteins not in the same complex and propose an effective ranking score for the predicted complexes.

We have compared our approach with MCL, MCODE, and CFinder. The results based on three PPI networks (DIP, Krogan, Gavin) show that our approach is better than all these existing approaches and can identify significantly many more known complexes than other approaches (30% more than MCL and CFinder, and double that of MCODE). We have also evaluated the quality of our predicted cores using the cores found by Gavin et al. (2006) as the benchmark. We compared our result with the mediators given by Andreopoulos et al. (2007). We can predict a lot more cores (ranging from 10- to 30-fold more), also

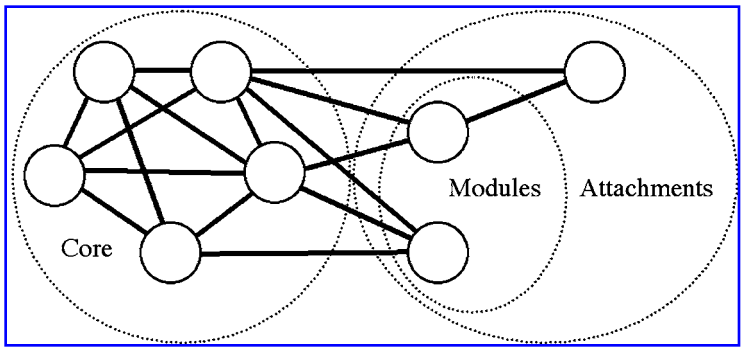

FIG. 1. Example of a protein complex.

with much higher accuracy. We have examined some of the high-ranked complexes we identified but not found in the list of known complexes. From the evidence of the known functions of the involved proteins, we found that they should be closely related and may be novel complexes not yet verified. We hope that this new core-attachment approach can provide a new and effective method for finding novel complexes from PPI networks.

2. METHODS

In this section, we describe how to discover a protein complex and estimate how likely a particular set of proteins can form a complex based on a given protein-protein interaction database.

The interactions between proteins can be represented by a graph (*protein-protein interaction* [PPI] *network*) defined as follows: Each protein is represented by a distinct node, and there is an edge (interaction) joining node u and node v if and only if some biological experiments (Ito et al., 2001; Krogan et al., 2006; Uetz et al., 2000) have shown that the corresponding proteins of these two nodes can bind to each other, i.e., we do not consider transient protein-protein interactions. In this case, we call u a *neighbor* of v, and vice versa. A protein complex is a set of proteins bound together to form a stable structure. The proteins in the complex and their known interactions form a subgraph of the PPI network. However, not every subgraph of the PPI network is a protein complex. In this work, we aim to find the list of distinct proteins that form a complex, so we ignore the interactions between the same protein, i.e., we do not consider self-loop in the PPI network.

Our method is based on a new finding for the organization of a protein complex (Gavin et al., 2006) in which proteins in each complex can be divided into two types: *core* and *attachment*. Each protein complex has a unique set of core proteins with sizes varying from 1 to 23 which do not appear in other protein complexes. Different from the core proteins, attachment proteins may appear in several complexes. Based on this organization, we discover protein complexes by first discovering the disjoint sets of core proteins and then identifying the attachments for each complex from its core proteins (see Section 2.2). The methods of predicting core proteins and their attachments are described in Sections 2.1 and 2.2, respectively, with their corresponding p-value. A ranking of the protein complexes based on p-value can also be derived.

2.1. Core proteins

Core proteins have three main properties (Gavin et al., 2006):

1. They have relatively more interactions among themselves.
2. The attachment proteins bind to the core proteins to form protein complex.
3. Each protein complex has a unique set of core proteins.

When considering property 1, the core proteins of a complex should be a subgraph of the PPI network with relatively (when compared with the degree of the nodes) more edges among themselves. For property 2, since each attachment protein usually binds to 2 or more core proteins to form a protein complex, if an attachment protein binds to a subset of core proteins, this attachment protein will be a common neighbor of the subset of core proteins it binds to in the PPI network. The core proteins in the same protein complex should have relatively more common neighbors than other proteins in the PPI network. For property 3, each set of core proteins should be disjoint.

When considering whether two proteins p_1 and p_2 of degree d_1 and d_2, respectively, are core proteins for the same protein complex, we consider the number of interactions i ($i = 0$ or 1) between these two proteins and the number of their common neighbors m. We calculate the probability that p_1 and p_2 have $\geq i$ interactions and $\geq m$ common neighbors under the null hypothesis that the d_1 edges connecting p_1 and d_2 edges connecting p_2 are randomly assigned in the PPI network according to a uniform distribution. In fact we only study two situations: when $i = 1$, p_1 and p_2 must have an interaction between them; when $i = 0$, p_1 and p_2 might or might not have an interaction between them.

Let N be the number of proteins in the PPI network, the probability that p_1 and p_2 have exactly i interactions between themselves can be calculated by considering the number of combinations to assign

$d_1 - i$ and $d_2 - i$ edges connecting p_1 and p_2 respectively to the rest of the $N - 2$ proteins as follows:

$$p_{interact}(i \mid N, d_1, d_2) = \frac{\binom{N-2}{d_1-i}\binom{N-2}{d_2-i}}{\binom{N-2}{d_1-0}\binom{N-2}{d_2-0} + \binom{N-2}{d_1-1}\binom{N-2}{d_2-1}} \tag{1}$$

Given that p_1 and p_2 have i interactions, the probability that p_1 and p_2 have exactly m common neighbors can be calculated by dividing the proteins into three groups: common neighbors of p_1 and p_2, neighbors of p_1, and neighbors of p_2:

$$p_{common}(m \mid N, d_1, d_2, i) = \frac{\binom{N-2}{m}\binom{(N-2)-m}{d_1-i-m}\binom{(N-2)-m-(d_1-i-m)}{d_2-i-m}}{\binom{N-2}{d_1-i}\binom{N-2}{d_2-i}} \tag{2}$$

The probability that p_1 and p_2 have $\geq i$ interactions and $\geq m$ common neighbors can be calculated by the product of Equations (1) and (2)

$$\begin{aligned} p\text{-value}(p_1, p_2) &= \Pr(\geq i \text{ interactions and } \geq m \text{ common neighbors}) \\ &= \sum_{i \leq j \leq 1, m \leq k \leq \min\{d_1, d_2\} - j} \left[\begin{array}{l} \Pr(= j \text{ interactions}) \cdot \\ \Pr(= k \text{ common neighbors} \mid = j \text{ interactions}) \end{array} \right] \\ &= \sum_{i \leq j \leq 1, m \leq k \leq \min\{d_1, d_2\} - j} [P_{interact}(j \mid N, d_1, d_2) P_{common}(k \mid N, d_1, d_2, j)] \end{aligned} \tag{3}$$

Small p-value(p_1, p_2) means that the null hypothesis is likely to be wrong for proteins p_1 and p_2, i.e., proteins p_1 and p_2 have higher chance of being a pair of core proteins in the same complex.

Complexes (as well as cores) may not have an absolute high density. Also, the size of a core may vary from 1 to 23. The degrees and the numbers of common neighbors of core proteins may vary a lot in different protein complexes. Thus, we cannot simply set one single threshold (i.e., p-value$(p_1, p_2) \leq$ threshold) to determine whether two proteins are a pair of core proteins in the same complex. Instead, we compare p-value(p_1, p_2) with other p-values involving protein p_1 and p_2. If p-value(p_1, p_2) is the smallest p-value among p-value(p_1, p_k) and p-value(p_2, p_k) for all possible protein p_k, i.e.,

$$p\text{-value}(p_1, p_2) < \min_{\substack{p_j = p_1 \text{ or } p_2 \\ p_k \notin \{p_1, p_2\}}} \{p\text{-value}(p_j, p_k)\}, \tag{4}$$

then we will conclude that proteins p_1 and p_2 are a pair of core proteins in the same complex. Similarly, we say a protein p_j is in a set of proteins C in the same complex if the p-value of p_j and any other protein in the core is smaller than the p-value of p_j and any other protein not in the core, i.e.,

$$\forall p_j \in C \max_{p_l \in C - \{p_j\}} \{p\text{-value}(p_j, p_l)\} < \min_{p_k \notin C} \{p\text{-value}(p_j, p_k)\} \tag{5}$$

In our algorithm, we progressively merge some proteins to a set of core proteins of size 2, 3, etc., until we cannot further increase the size of a core protein set with the condition defined by Equations (4) and (5). It is easy to see that the core protein sets are disjoint because each protein can only associate with a unique set of proteins with the lowest p-values. After discovering core proteins of size at least 2, we pick each of the rest of the proteins as single core protein. As a remark, Equation (1) can be extended to compute the p-value for cores of more than two proteins. However, the computational requirement will be increased substantially. So, we use the progressive approach to identify cores of more than two proteins.

2.2. Attachment proteins

Each attachment protein usually binds to two or more core proteins depending on the size of core protein set. Given a core protein set of a complex, we pick those proteins that are common neighbor of over half of the core proteins as the attachment proteins. When there is only one core protein, we pick all its neighbors as the attachment proteins. We do not consider the degree of protein when determining whether it is an attachment of a complex because an attachment protein may appear in more than one protein complex. In this case, the attachment protein can have many interactions with core proteins in other complexes.

2.3. Noisy cores

It is possible that some of the cores identified in the first step are noise because some of the attachment proteins which have relatively more interactions may also be identified as core proteins. After finding the attachment proteins for each core, some of these noisy cores may appear as the attachment of other cores. Since core proteins usually should not appear as attachments to other cores (Gavin et al., 2006), we can make use of this concept to filter out some of these noisy cores: the cores that overlap with the attachment of other cores or the cores whose attachment proteins overlap with other cores. We define a significant score for each core as follows.

Given a set C of c core proteins with total degree d_c, i_c interactions in between the core proteins and i_a interactions with their corresponding m attachments, we define $p_{exact}(i_c, m, i_a \mid c, d_c)$ to be the probability that C has i_c interactions in between and i_a interactions with its m attachments under the null hypothesis that the d_c interactions involved with proteins in C and the N proteins in the PPI network under a uniform distribution (Equation (6)). The first term of the numerator calculates the number of ways to allocate the i_c interactions between core proteins. The second term calculates the number of ways to select m attachments. The third term calculates the number of ways to select the rest of the proteins interacting with the core proteins. The fourth term approximates the number of ways to allocate the interactions between core proteins and attachments where $d_{\min}$ is the minimum number of core proteins an attachment protein must interact with. This is an approximation because of the restriction that an attachment cannot have two interactions with the same core proteins.

$$p_{exact}(i_c, m, i_a \mid c, d_c) \approx \frac{\binom{\binom{c}{2}}{i_c}\binom{N-c}{m}\binom{N-c-m}{d_c-2i_c-i_a}\binom{i_a-d_{\min}m+m-1}{i_a-d_{\min}m}}{\sum_{i'_c=0}^{\binom{c}{2}}\sum_{m'=0}^{\lfloor\frac{d_c-2i'_c}{d_{\min}}\rfloor}\sum_{i'_c=d_{\min}m'}^{d_c-2i'_c}\binom{\binom{c}{2}}{i'_c}\binom{N-c}{m'}\binom{N-c-m'}{d_c-2i'_c-i'_a}\binom{i'_a-d_{\min}m'+m'-1}{i'_a-d_{\min}m'}} \tag{6}$$

We then calculate the probability $p_{core}(i_c, m, i_a \mid c, d_c)$ that C has $\geq i_c$ interactions in between, $\geq m$ attachments and $\geq i_a$ interactions with the core proteins by summing up all possible cases:

$$p_{core}(i_c, m, i_a \mid c, d_c) = \sum_{i''_c=i_c}^{\binom{c}{2}}\sum_{m''=m}^{\lfloor\frac{d_c-2i''_c}{d_{\min}}\rfloor}\sum_{i''_a=i_a}^{d_c-2i''_c} p_{exact}(i''_c, m, i''_a \mid c, d''_c) \tag{7}$$

A set of core proteins with low p_{core} value means that the numbers of interactions within this set of proteins and with its attachments are unexpectedly large and it is more likely to be a real core. Therefore, we should select core protein sets with low p_{core} values. We filter the core proteins by repeatedly select core protein set C with the lowest p_{core} value and remove those core proteins that overlap with the attachments of C. Finally, we would get a set of protein complexes such that the core proteins of each complex do not appear in other complexes.

2.4. Ranking protein complexes

After predicting a set of protein complexes, we have to evaluate which predicted protein complex is more likely to be a real protein complex. Many researchers (Adamcsek et al., 2006; Bader et al., 2003) have suggested that proteins in a real protein complexes usually have more interactions in between. However, when the sizes of the protein complexes are different, it is difficult to determine which protein complex is more likely to be real. For example, protein complex A has five proteins and 10 interactions in between (there are at most 10 interactions between five proteins) and protein complex B has 10 proteins and 40 interactions in between (there are at most 45 interactions between 10 proteins), we cannot determine whether A or B is more likely to be a real complex easily as we also have to consider the number of interactions complex A and complex B have with proteins outside the complexes. Therefore, most existing algorithms (Adamcsek et al., 2006; Bader et al., 2003; Dongen, 2000; Enright et al., 2002) predict sets of protein complexes without ranking them.

In this paper, we evaluate each predicted protein complex by comparing the total number of interactions within the complex and the total number of interactions with proteins outside the complex. Suppose a protein complex has q proteins, i interactions between proteins in the complex and deg_o interactions with proteins not in the complex, i.e., the total degree of all proteins in the complex is $2i + deg_o$. As in Section 2.1, we calculate the probability that there are $\geq i$ edges in the subgraph representing the protein complex under the null hypothesis that the $2i + deg_o$ edges connecting proteins in the complex are randomly assigned in the PPI network according to a uniform distribution.

$$p\text{-score} = \frac{\sum_{j=i,\ldots,I} \binom{\binom{q}{2}}{j} \binom{q(N-q)}{2i+deg_o-2j}}{\sum_{k=0,\ldots,I} \binom{\binom{q}{2}}{k} \binom{q(N-q)}{2i+deg_o-2k}} \tag{8}$$

where $I = \min\{(2i + deg_o)/2, \binom{q}{2}\}$ is the maximum number of interactions in the protein complex.

A protein complex will have a smaller p-score when it has unexpectedly large number of interactions in the complex and has a higher probability to be a real protein complex. Therefore, we calculate the p-score for each predicted complex and rank them according to their p-scores in increasing order.

3. RESULTS

We have implemented our approach and have evaluated the quality of our predicted complexes as well as the quality of the cores. We used three protein-protein interaction datasets for yeast, including DIP[2] (Xenarios et al., 2000) and two others obtained from high-throughput methods[3] (Krogan [Krogan et al., 2006] and Gavin [Gavin et al., 2006]). The details of the datasets are shown in Table 1.

3.1. Evaluation of predicted complexes

We compared the quality of our predicted complexes with those produced by MCL (Brohee et al., 2006), MCODE (Bader et al., 2003), and CFinder (Adamcsek et al., 2006) using their default parameters. The known complexes found in the MIPS (Mewes et al., 2000) database[4] are used as the benchmark. We only consider the complexes that were manually annotated independently from the DIP interaction data, and exclude all complexes predicted based on high-throughput experiments. We consider complexes of sizes at least 5, and there are 81 such complexes.

[2]Download from *http://dip.doe-mbi.ucla.edu/* [version yeast20071104].

[3]Download from the GRID database *www.thebiogrid.org/* [BioGRID version 2.0.33].

[4]Download from *http://mips.gsf.de/* [version complexcat_data_18052006].

TABLE 1. DETAILS OF THE DATASETS

Datasets	*Number of proteins*	*Number of interactions*	*Average number of interactions per proteins*
DIP	4928	17,201	5.29
Krogan	2675	7080	6.98
Gavin	1430	6531	10.62

3.2. Evaluation method

Each software tool based on their recommended settings reports a list of predicted clusters (complexes). Each cluster is compared with every known complex and given an accuracy $Acc = N_c^2/N_pN_t$ (Bader et al., 2003), where N_c is the number of common proteins shared by the predicted and known complexes, and N_p and N_t are the numbers of proteins in the predicted and true complexes, respectively. Then, for each complex, select the cluster with the maximum accuracy. If this accuracy is larger than a threshold $t = \{0.6, 0.7, 0.8\}$, we assume that the complex has been found. The complexes not found by the clusters are regarded as false negative. Sensitivity is defined as the percentage of known complexes being found. Clusters that cannot match any known complex with accuracy more than t are regarded as false positives. Specificity is defined as the percentage of false positive clusters. Here, we do not use the average accuracy for comparison as in Brohee et al. (2006), because the average accuracy cannot indicate how many protein complexes have been predicted by the software accurately. However, the relative performances of the software are the same when they are measured by average accuracy.

Note that MCL, MCODE, and CFinder predict different numbers of clusters without ranking. Since we cannot control the number of predictions from these tools, to provide a reasonable comparison, if a tool reports x clusters, our approach uses the top x clusters, if possible or all clusters from our software for the comparison.

3.3. Results on predicting protein complexes

The tables below show the performance of our algorithm (Core) compared to the other three tools on three different datasets and three different thresholds for the accuracy measure ($t = 0.6, 0.7, 0.8$). A known complex is regarded as being found by the software if there is a predicted cluster reported by the software which matches this known complex with accuracy at least t. Table 2 shows our comparison results with MCL core predicted more complexes in all cases and on average about 30% more complexes than MCL. We can also achieve a higher specificity in all cases, that is, among our predicted clusters, we have a high percentage of them that match known complexes.

Table 3 shows our comparison results with MCODE. Core is a clear winner in all cases. In many cases, the number of predicted complexes is about double or even triple that of MCODE and with a higher specificity in all cases. Tables 4 and 5 show our comparison results with CFinder for clique size (k) of 3 and 4, respectively. Core produces much better results.

TABLE 2. MCL VERSUS CORE

	No. of known complexes correctly predicted		
MCL/Core	$t \geq 0.6$	$t \geq 0.7$	$t \geq 0.8$
Krogan	30/36	20/26	10/15
DIP	28/37	16/21	7/11
Gavin	32/35	26/29	11/17

The numbers of predicted complexes by MCL are 633, 1250, and 232 for Krogan, DIP, and Gavin, respectively.

TABLE 3. MCODE VERSUS CORE

	No. of known complexes correctly predicted		
MCODE/Core	$t \geq 0.6$	$t \geq 0.7$	$t \geq 0.8$
Krogan	17/29	13/22	7/13
DIP	6/24	5/16	2/8
Gavin	23/32	19/27	8/17

The numbers of predicted complexes by MCODE are 72, 77, and 105 for Krogan, DIP, and Gavin, respectively.

TABLE 4. CFINDER ($k = 3$) VERSUS CORE

	No. of known complexes correctly predicted		
CFinder/Core	$t \geq 0.6$	$t \geq 0.7$	$t \geq 0.8$
Krogan	19/32	16/24	11/14
DIP	17/35	11/21	6/11
Gavin	22/31	16/27	12/17

The numbers of predicted complexes by CFinder ($k = 3$) are 115, 245, and 98 for Krogan, DIP, and Gavin, respectively.

TABLE 5. CFINDER ($k = 4$) VERSUS CORE

	No. of known complexes correctly predicted		
CFinder/Core	$t \geq 0.6$	$t \geq 0.7$	$t \geq 0.8$
Krogan	19/28	14/22	10/13
DIP	28/28	13/19	5/11
Gavin	25/29	20/26	13/16

The numbers of predicted complexes by CFinder ($k = 4$) are 66, 100, and 71 for Krogan, DIP, and Gavin, respectively.

Overall, Core performs much better than all existing tools in identifying more known complexes in all cases. In fact, when the required accuracy is increased from 0.6 to 0.8, Core still produces much better results than all the other tools.

3.4. Evaluation of predicted cores

We have also compared the quality of our predicted cores with the mediators predicted by Andreopoulos et al. (2007) on the dataset Gavin (Gavin et al., 2006). There are 491 cores found in Gavin et al. (2006), while there are 274 mediators predicted by Andreopoulos et al. (2007). We assume that a core can be found by a software tool if the accuracy between this core and a predicted core is λ. In our experiments, we used $\lambda = 0.4$, 0.5, 0.6, 0.7, and 0.8. We used the same accuracy measure for the predicted cores as for the predicted complexes. To compare with the 274 mediators, we used the cores from the top 274 complexes.

3.5. Results

Table 6 shows the numbers of cores correctly predicted by Core and the software used in Andreopoulos et al. (2007) using different accuracy thresholds. It is clear that the performance of Core is significantly better than Andreopoulos et al. (2007). In all cases, Core can predict many more benchmark cores than the mediators with less false positives and false negatives.

TABLE 6. COMPARISON ON THE NUMBER OF CORRECTLY PREDICTED BENCHMARK CORES

	Accuracy threshold (λ)				
	$\geq$0.4	*$\geq$0.5*	*$\geq$0.6*	*$\geq$0.7*	*$\geq$0.8*
Core					
TP	267	244	169	150	103
FP	44	60	112	131	172
FN	224	247	322	341	388
Mediators					
TP	29	13	8	5	0
FP	238	258	265	268	274
FN	462	478	483	486	491

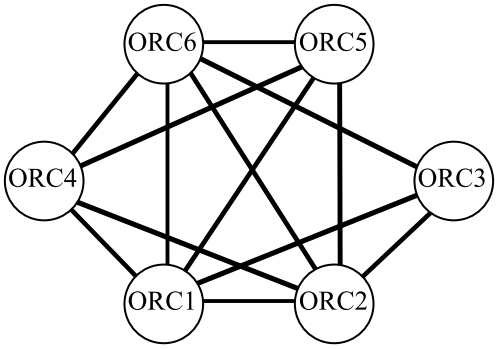

FIG. 2. Interactions between 6-protein core.

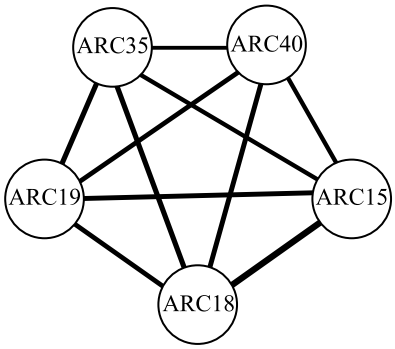

FIG. 3. Interactions between 5-protein core.

The quality of the cores produced by our software is much higher. In fact, there are quite a few cases where our predicted core matches completely with the known core found in Gavin et al. (2006). For example, Figure 2 shows a 6-protein core, which is exactly the same as the core given in Gavin et al. (2006). Based on the gene ontology (GO[5]) database, all the component proteins (ORC1, ORC2, ORC3, ORC4, ORC5, and ORC6) have the same function of directing DNA duplication by binding to replication origins and are involved in transcriptional silencing. The complex constructed from this core also matches very well with the known complex (post-replication complex). These results also support the core-attachment concept presented in Gavin et al. (2006).

Another example is the 5-protein core (Fig. 3) with proteins ARC15, ARC18, ARC19, ARC35, and ARC40. This core also matches exactly with a known core. The proteins are all required for the motility and integrity of cortical actin patches. The cluster predicted using this core matches very well with the known complex (Arp2p/Arp3p complex).

Table 7 shows the distribution of the sizes of the cores predicted by the tools when compared to the known cores found in Gavin et al. (2006). The cores predicted by our software seem to have a distribution more similar to that of the known cores than the mediators predicted by Andreopoulos et al. (2007).

4. DISCUSSION

We have performed a preliminary investigation on those predicted complexes reported by Core that are of high rank, but do not match any of the known complexes. We found that they may be potentially new

[5] *www.geneontology.org/.*

TABLE 7. DISTRIBUTION OF THE SIZES OF THE CORES

Size of cores	*Maximum*	*Minimum*	*Average*
Gavin et al. (2006)	23	1	2.98
Core	8	1	1.99
Mediators	152	1	12.31

complexes. For most of these complexes, their cores match quite well with the known cores given in Gavin et al. (2006), and the functions of the core proteins are very related. For example, the predicted complex of rank 1 consists of 39 proteins with two proteins (YOl077C and YHR052W) in its core. This core is found to be totally inside a known core given in Gavin et al. (2006). These two proteins have been predicted to be in the same complex also in Ho et al. (2002), Riffle et al. (2007), and Qiu et al. (2008). The functions of the core proteins are found to be closely related to ribosome biogenesis and assembly. Another example is the complex predicted to be of rank 4. This complex has 32 proteins with three proteins (YDR101C, YLR074C, and YGR245C) forming the core. This core is also totally inside a known core given in Gavin et al. (2006). The result is also supported by Riffle et al. (2007). Further investigation is needed to verify whether these predicted complexes are real complexes.

5. CONCLUSION

We have developed a new approach for predicting protein complexes from the PPI network of single species based on a recent study on the organization of complexes. We found that this approach is more effective than existing approaches and can identify more known complexes. It also has the potential of identifying new complexes.

In this work, emphasis is mainly on predicting cores, the step for adding attachment proteins was done based roughly on the idea of majority rule (that is, proteins have interactions with more than half of the cores proteins will be added as an attachment protein to the core). A better approach for selecting attachment proteins may further improve the prediction power of the software. Also, the current ranking function for the complexes (Step 3) may be slightly biased to large complexes. It may be interesting to derive a better ranking function. Also, we have analyzed all 81 known complexes in our study and have found that, in most of the cases that cannot be discovered by our software, there are only have a few interactions among the proteins. It is unlikely that these complexes can be recovered based only on one PPI network without additional information, e.g., GO annotation. It is known that there are missing interactions and also false positive interactions in the PPI network. We do not consider these issues in the current approach. Using multiple PPI networks on different species may help fill missing edges in the PPI networks. Extending our approach to consider the possibility of having false positive interactions in the PPI network may also produce more accurate results.

ACKNOWLEDGMENTS

This work was supported in part by Hong Kong RGC (grant HKU 711608E) and Seed Funding Programme for Basic Research (grant 200611159001) of the University of Hong Kong. The work of Q.X. was supported in part by Research Fund for the Doctoral Program of Higher Education of China (grant 4111279) and Natural Science Foundation of Guangdong Province, China (grant 4203176).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Adamcsek, B., et al. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023.

Andreopoulos, B., et al. 2007. Clustering by common friends finds locally significant proteins mediating modules. *Bioinformatics* 23, 1124–1131.

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

Blatt, M., et al. 1996. Superparamagnetic clustering of data. *Phys. Rev. Lett.* 76, 3251–3254.

Brohee, S., and Helden, J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinform.* 7, 488.

Dongen, S. 2000. Graph clustering by flow simulation [Ph.D. dissertation]. Centers for Mathematics and Computer Science, University of Utrecht.

Enright, A.J., et al. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584.

Gavin, A.C., et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636.

Ho, Y., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183.

Ito, T., et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 98, 4569–4574.

King, A.D., et al. 2004. Protein complex prediction via cost-based clustering. *Bioinformatics* 20, 3013–3020.

Krogan, N.J., et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 637–643.

McDonald, T.. et al. 2006. Expanding the subproteome of the inner mitochondria ising protein separation technologies. *Mol. Cell. Proteomics* 5, 2392–2411.

Mewes, H.W., et al. 2000. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40.

Qiu, J., and Noble, W.S. 2008. Predicting co-complexed protein pairs from heterogeneous data. *PLoS Comput. Biol.* 4, e1000054.

Riffle, M., and Davis, T.N. 2007. Protein complexes utilizing statistical cliques present in mass spectrometry datasets. Manuscript submitted.

Roh, M., and Margolis, B. 2003. Composition and function of PDZ protein complexes during cell polarization. *Am. J. Phys. Renal Physiol.* 285, F377–F387.

Sugiura, T., and Berditchevski, F. 1999. Function of $\alpha 3\beta 1$-tetraspanin protein complexes in tumor cell invasion. Evidence for the role of complexes in production of matrix metalloproteinase 2 (MMP-2). *J. Cell Biol.* 146, 1375–1389.

Xenarios, I., et al. 2000. DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* 28, 289–291.

Uetz, P., et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627.

Address reprint requests to:
Dr. Henry C.M. Leung
Department of Computer Science
University of Hong Kong
Pokfulam Road
Hong Kong

E-mail: cmleung2@cs.hku.hk