BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 7 2008, pages 979–986
doi:10.1093/bioinformatics/btn036

Systems biology

From pull-down data to protein interaction networks and complexes with biological relevance

Bing Zhang[1,†,‡], Byung-Hoon Park[1,†], Tatiana Karpinets[1] and Nagiza F. Samatova[1,2,*]

[1]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 and [2]Computer Science Department, North Carolina State University, Raleigh, NC 27695, USA

Received on October 25, 2007; revised on January 2, 2008; accepted on January 22, 2008

Advance Access publication February 26, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Recent improvements in high-throughput Mass Spectrometry (MS) technology have expedited genome-wide discovery of protein–protein interactions by providing a capability of detecting protein complexes in a physiological setting. Computational inference of protein interaction networks and protein complexes from MS data are challenging. Advances are required in developing robust and seamlessly integrated procedures for assessment of protein–protein interaction affinities, mathematical representation of protein interaction networks, discovery of protein complexes and evaluation of their biological relevance.

Results: A multi-step but easy-to-follow framework for identifying protein complexes from MS pull-down data is introduced. It assesses interaction affinity between two proteins based on similarity of their co-purification patterns derived from MS data. It constructs a protein interaction network by adopting a knowledge-guided threshold selection method. Based on the network, it identifies protein complexes and infers their core components using a graph-theoretical approach. It deploys a statistical evaluation procedure to assess biological relevance of each found complex. On *Saccharomyces cerevisiae* pull-down data, the framework outperformed other more complicated schemes by at least 10% in F_1-measure and identified 610 protein complexes with high-functional homogeneity based on the enrichment in Gene Ontology (GO) annotation. Manual examination of the complexes brought forward the hypotheses on cause of false identifications. Namely, co-purification of different protein complexes as mediated by a common non-protein molecule, such as DNA, might be a source of false positives. Protein identification bias in pull-down technology, such as the hydrophilic bias could result in false negatives.

Contact: samatovan@ornl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Cellular functions are carried out by intricate network of interacting proteins (Alberts, 1998). In particular, groups of proteins working together to achieve relatively distinct biological functions through a series of protein interactions constitute the fundamental units in the network of interacting proteins: protein functional modules (Hartwell *et al.*, 1999), or protein complexes. From the network topological perspective, modular structure of protein interaction networks are indicated by the graph–theoretical property of high-clustering coefficient, i.e. cliquishness (Barabasi and Oltvai, 2004). Accordingly, protein complexes are typically interpreted as regions in the network, where vertices are more densely connected to each other than to the rest of the network (Bader and Hogue, 2003).

Protein complex identification inevitably depends on the quality of the protein–protein interaction networks that are in turn dependent on the experimental technologies for detecting the interactions. Due to the ability to detect physiological complexes in natural settings (Drewes and Bouwmeester, 2003; Gavin *et al.*, 2002), biochemical purification of proteins in combination with high-throughput tandem Mass Spectrometry (MS/MS) has become an important strategy for the genome-wide identification of protein interactions (Butland *et al.*, 2005; Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006). This strategy is commonly referred to as the pull-down technology, in which a bait protein is used to pull-down associated prey proteins, followed by the identification of the proteins through the MS/MS analysis. Although promising, protein interaction network construction and complex identification from pull-down data still requires careful attention in handling errors and noise.

Although alternative approaches have been proposed (Scholtens *et al.*, 2005), identification of protein complexes from pull-down data usually involves the following steps: (i) assessment of pair-wise protein interaction affinities, (ii) construction of a protein interaction network, (iii) identification of candidate protein complexes, and (iv) post-evaluation and finalization of the complexes. Every step is crucial and needs to be performed with great care. Also, all steps are closely coupled and thus should be performed in a coherent manner, as each subsequent step depends heavily on previous steps. For example, different sets of complexes can be obtained with different measurements of protein interaction affinity.

Assessment of protein interaction affinity is a process of addressing biological variability and technical limitations that often result in finding spurious interactions (false positives) and missing some true interactions (false negatives). Assessment of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Present address: Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232.

protein interaction affinities from pull-down studies requires special attention. Unlike technologies that produce a set of binary (pair-wise) interactions such as yeast two-hybrid, the pull-down approach purifies and analyzes whole protein complexes, thus requiring specific topological assumptions about bait and prey proteins to draw pair-wise protein interactions. Commonly used topologies include the 'spoke' and 'matrix' models that define interactions between bait and preys, and between all constituents in a complex, respectively (Bader and Hogue, 2002). It is suspected that the spoke and the matrix models tend to have high false negative rates and high false positive rates, respectively. Our empirical evaluation using known complexes from the Munich Information Center for Protein Sequences (MIPS) database and the pull-down data from (Gavin *et al*., 2006) indeed verifies this suspicion. Among the 13 384 pair-wise protein interactions derived from the MIPS complexes, the spoke model identified 2656 (19.8%) true interactions, and introduced 5214 (39.0%) false interactions. In contrast, although the matrix model was able to cover 9202 (68.8%) true interactions, it introduced 41 320 (308.7%) false interactions.

For this reason, we are witnessing hybrid approaches that combine the two models. 'Socio-affinity' index (Gavin *et al*., 2006) is one such effort that computes affinity between proteins by measuring the log-odds of the number of times two proteins are observed together as bait and a prey, or a prey and a prey, relative to what would be expected from their frequency in the data set. However, such a method is most suitable for data sets where all proteins appear as both baits and preys. Since we do not always expect such reciprocal appearances, a modified version based on the probabilistic framework of a naïve Bayes classifier has been developed and demonstrated to outperform the socio-affinity index (Collins *et al*., 2007).

Protein complex identification usually relies on protein interaction networks that are constructed from statistically assessed pair-wise protein interaction affinities (Gilchrist *et al*., 2004). Typically, a network is constructed by setting a threshold for the pair-wise affinity. Such a network is usually represented as an un-weighted graph, where vertices and edges correspond to proteins and interactions between them, respectively. To be more specific, an edge is drawn between two vertices (proteins) if their affinity level is above a pre-selected threshold. In many cases, an arbitrary and empirical threshold is used. However, it should be noted that a choice of threshold can significantly affect the integrity of the network and the protein complexes derived from it. Once the initial candidate set of complexes is found from the network, a rigorous post analysis should be performed to derive the final results. This last step is critical in revealing the dynamic organization property of the complexes (Gavin *et al*., 2006), and associating complexes to underlying biological roles.

In this article, we introduce a multi-step framework that coherently integrates the aforementioned four steps for protein complex identification. Not only do we advance each step, but also provide an easy-to-follow scheme to integrate steps to maximize the informativeness of the available data. For the assessment of protein interaction affinity (Step 1), we suggest comparing the co-purification patterns of two proteins across different pull-down experiments. Our underlying assumption is that similarity of co-purification patterns indicates the likelihood of protein–protein affinity. We adopt the Dice coefficient to quantify such an affinity. Affinity scores thus found are used to construct a protein interaction network through a knowledge-guided information theoretic thresholding method (Step 2). Specifically, the affinity threshold that maximizes a balanced sensitivity and specificity of the resulting interaction network with respect to the known evidences is sought. From the network graph constructed with a chosen threshold, we seek to find all maximal cliques as candidate protein complexes (Step 3). A maximal clique is a clique that is not contained in any other clique, while a clique is a complete graph in which all vertices are pairwisely connected. Since we assume that all interaction partners have similar co-purification pattern, maximal cliques could serve as good candidate complexes. However, due to high false negative rates (undetected true interactions) of protein interactions found in pull-down experiments (Yu *et al*., 2006) and the dynamic property of protein complexes (Gavin *et al*., 2006), we note that a protein complex can be represented by many nearly identical cliques, similar to the complex isoforms described in Gavin *et al*. (2006). For this, in Step 4, we propose an algorithm that merges the isoforms (highly overlapped cliques) into single protein complexes, and divide the proteins in a complex into core components (common to most of the isoforms) and attachment components (unique to specific isoforms).

We evaluate our proposed framework by comparing the performance with those of the original iterative hierarchical clustering-based method (IHC) (Gavin *et al*., 2006) and two popular protein complex identification algorithms, the Mollecular Complex Detection (MCODE) algorithm (Bader and Hogue, 2003) and the Markov Clustering (MCL) algorithm (Enright *et al*., 2002). In particular, we compare the recalls and precisions for different methods based on the manually curated complex data from MIPS.

Finally, we analyze the biological relevance of the identified complexes. Based on the GO category enrichment analysis, the complexes are associated with the most enriched GO categories in order to assign each complex a predominant biological function.

2 METHODS

2.1 Generation of protein by pull-down matrix

The pull-down data is transformed into a binary protein by pull-down matrix, where each row corresponds to a protein and each column corresponds to a pull-down experiment. A cell (i,j) in the matrix is one if protein i is pulled down as a prey in the experiment j, and a zero, otherwise.

2.2 Pair-wise affinity score calculation

The Dice coefficient is used to score the interaction affinity between two proteins. With the matrix representation (Section 2.1), a protein is a binary vector. For two protein vectors (i and j), the Dice coefficient is computed by counting the number of elements (experiments) on which both proteins take the same or different values. Specifically, let q denote the number of elements that both protein i and j have ones. Let r denote the number of elements that protein i has ones, but zero for protein j.

Likewise, let s denote the number of elements that protein j has ones, but zeros for protein i. Then the Dice coefficient is formally defined as,

$$D(i,j) = \frac{2q}{2q + r + s}.$$

For comparison, the socio-affinity (SA) score was calculated according to Gavin *et al*. (2006), the Gilchrist score (GS) was calculated according to Gilchrist *et al*. (2004) and the Purification Enrichment (PE) score for the same data set was downloaded from http://interactome-cmp.ucsf.edu.

2.3 Protein interaction network construction

A protein interaction network is represented as an undirected and unweighted graph. Generating one such graph from a pull-down data set requires a pre-selected threshold of the pair-wise affinity scores. Since an arbitrarily chosen threshold may result in poor complex identifications, we use manually annotated protein complexes in the MIPS database (ftp://ftpmips.gsf.de/yeast/catalogues/complexcat) to guide the threshold selection. In order to maximize both recall and precision, we use F_1-measure, which is the harmonic mean of recall and precision with an equal weight, as the metric. A series of thresholds are tested and the one that produces the best F_1-measure is chosen. Specifically, given a threshold, protein pairs are classified into four categories as shown Table 1. Then recall, precision and F_1-measure are calculated as shown in Table 1.

2.4 Protein complex identification

As in Gavin *et al*. (2006), we define protein complexes as sets of proteins where all pair-wise affinity scores are greater than a predefined threshold. From a graph theoretical point of view, a complex can be represented as a completely connected subgraph (or clique) of proteins, where an edge corresponds to a pair-wise affinity between proteins. Then protein complex identification is reduced to the maximal clique finding problem. A maximal clique is a clique that is not part of any other larger cliques, i.e. inclusion of any other vertex to a maximal clique will violate its completeness. We apply our maximal clique finding algorithm (Zhang *et al*., 2005) to enumerate cliques of size three or higher from the protein interaction network.

In practice, maximal cliques found from a protein interaction network need to be further processed. Due to technical errors as well as the dynamic organization of the complexes, a number of almost identical cliques can be produced. Merging these cliques into a single bigger complex will not only reduce false negatives, but also help revealing the dynamic organization of protein complexes.

The clique merging algorithm adopts the Meet/min coefficient (Goldberg and Roth, 2003) as the similarity measure between two cliques. For a given pair of cliques c_i and c_j, let q be the number of proteins that belong to both c_i and c_j, and let r and s be the numbers of proteins that only belong to c_i or c_j, respectively. The Meet/min coefficient is then calculated by,

$$M(i,j) = \frac{q}{\min(r,s)},$$

where min(r,s) is the minimum of r and s.

The clique merging algorithm iterates over a series of sessions, where in each session all candidate complexes are examined for potential merges. The algorithm stops when no changes in the candidate complex set between two consecutive sessions are observed. Initially, the candidate complex set consists of cliques identified from maximal clique algorithm, i.e. $C_0 = \{c_1, c_2, \ldots, c_l\}$, where each c_i is a clique. During each session, for each c_i, Meet/min coefficients between c_i and all other complexes are measured. Then the closest c_j (complex with the highest coefficient) is merged with c_i if the coefficient is above the given threshold, and new candidate complex $c_i \cup c_j$ is added into the candidate complex set. At the end of the session, both c_i and c_j are removed from the candidate set, and C_k becomes C_{k+1} if it was the k-th session. The algorithm takes a greedy approach to suppress the explosion of a candidate complex set. It may thus prevent some pairs of close (similar) complexes from being merged. However, if the pair is closest at least in one direction (c_i is the closest to c_j or c_j is the closest to c_i), it is guaranteed to be merged.

Proteins in the merged complexes are further separated into core proteins and attachment proteins. For each merged complex, we define the core proteins as those present in 2/3 or more of the original cliques used for the merging, and the other proteins are defined as attachment proteins.

For comparison, the MCODE software was downloaded from http://cbio.mskcc.org/~bader/software/mcode/, and the MCL software was downloaded from http://micans.org/mcl/.

Table 1. Contingency table for protein affinity scores

	Within a known protein complex	Not within a known protein complex
Above the threshold	True positive (*TP*)	False positive (*FP*)
Below the threshold	False negative (*FN*)	True negative (*TN*)

recall $= \frac{TP}{TP+FN}$, precision $= \frac{TP}{TP+FP}$, and F_1-measure $= \frac{2*\text{recall}*\text{precision}}{\text{recall}+\text{precision}}$.

2.5 Evaluation of identified protein complexes

We use the enrichment analysis to evaluate the protein complexes identified in this study based on the GO annotation downloaded from the Gene Ontology website, http://geneontology.org/GO.current.annotations.shtml. GO has a hierarchical structure, and a protein can be mapped into multiple categories in the same or different hierarchies. Since the aim of this study is to assess the biological relevance of each inferred protein complex in terms of a consensus in GO categories, we consider all categories in every level of the hierarchy for the evaluation (Zhang *et al*., 2004).

In order to identify GO categories that are enriched in a protein complex, we compare the statistical likelihood of proteins in the complex in each GO category to those in the reference set, i.e. C_{total}. Specifically, for a given protein complex and GO category X, let m and K be the numbers of proteins in the complex and category X, respectively. Let us further assume that k out of the m proteins in the complex are in category X. Then category X is said to be enriched if k exceeds its expected value, $m(K/C_{\text{total}})$. To assess the significance of enrichment (P) for a given category, we perform the hypergeometric test as described in (Zhang *et al*., 2005). Namely,

$$P = \sum_{i=k}^{m} \frac{\binom{C_{\text{total}} - K}{m - i}\binom{K}{i}}{\binom{C_{\text{total}}}{m}}$$

Finally, for a protein complex, we report the most significantly enriched categories under biological process, molecular function and cellular components, respectively.

3 RESULTS

We tested our method on the pull-down data in *S.cerevisiae* (Gavin *et al.*, 2006), where 3260 pull-down experiments are stored among 1993 bait and 2760 prey proteins.

3.1 Evaluation of protein interaction affinity scores

We assessed our Dice coefficient-based scoring (DC) scheme for protein interaction affinity by comparing its performance with those of the socio-affinity (SA), Gilchrist score (GS), and Purification Enrichment (PE). All four schemes are designed to assign pair-wise affinity scores for all proteins identified in an experiment. We investigated how each scheme successfully distinguishes true interactions from false interactions. To be more specific, with a selected threshold for each scheme, a pair of proteins is classified to be of a true interaction if their affinity score is above the threshold, or of a false interaction, otherwise. The 267 manually annotated protein complexes downloaded from the MIPS database were used to validate classifications. The performances of the three approaches were compared by the F_1-measure. For a fair comparison, we chose a large variety of thresholds within the score ranges for each scoring system. DC and GS generated scores ranging from 0 to 1, while SA and PE ranged from -3 to 22 and 0 to 30, respectively. Figure 1 shows the F_1-measures for different scoring systems based on different thresholds. DC and PE performed the best, with the best F_1-measure of 0.53 achieved at the thresholds of 0.15 and 1.5, respectively. SA achieved the best F_1-measures of 0.43 at the threshold of 4. GS seems very insensitive, the best F_1-measure of 0.25 was observed for thresholds varying from 0.05 to 0.5. We also calculated the F_1-measure for the pair-wise interactions derived from the spoke and matrix models. The F_1-measures for the spoke model and the matrix model were 0.25 and 0.29, respectively. These results indicate that DC and PE best represent the affinity of protein interaction. Therefore, we used these two scoring schemes for the network construction in the following step.

3.2 Identification of candidate protein complexes via clique enumeration

Based on the above result, we constructed two versions of yeast protein interaction networks. Network D is based on the DC threshold of 0.15, while Network P is based on the PE threshold of 1.5. Network D has 2109 vertices and 16 169 edges, while Network P has 4543 vertices and 37 000 edges.

Using the maximal clique finding algorithm, we identified 4123 and 19 242 cliques of size three or more from Network D and Network P, respectively. There is a large amount of overlap among the cliques. For example, the maximum clique (the largest maximal clique) in Network D overlapped with 43 other maximal cliques, among which 25 overlaps covered more than half of the proteins in the corresponding maximal cliques. More severely, the maximum clique in Network P overlapped with 10 954 other maximal cliques, among which 768 overlaps covered more than half of the proteins in the corresponding maximal cliques. We used the Meet/min coefficient to quantify the overlap between two cliques, and found that 308 812 and 31 821 406 clique pairs showed a Meet/min coefficient of greater than 0.5 in Network D and P, respectively. Many cliques differed only in a few proteins. The high degree of overlap can be partially attributed to the undetected true interactions (Yu *et al.*, 2006), which reflects the technical imperfection. However, it may also reflect the dynamic organization of the protein complexes (Gavin *et al.*, 2006). Merging these overlapping cliques into a single bigger complex may not only help reduce false negatives, but also reveal interesting biological dynamics.

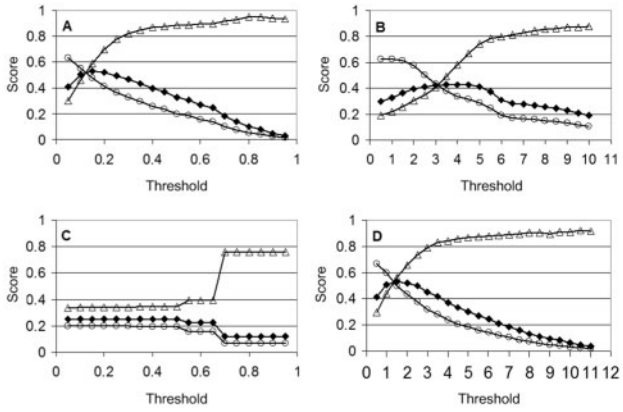

Fig. 1. Performance evaluation of the protein interaction affinity scoring schemes based on the MIPS complex catalogue. (**A**) Dice coefficient (DC), (**B**) Socio-affinity index (SA), (**C**) Gilchrist *et al.*'s score (GS), (**D**) Purification enrichment score (PE). Solid diamond, open circle lines, and open triangle lines represent F_1-measure, recall and precision, respectively.

3.3 Production of protein complexes through merging highly overlapping cliques

The final sets of protein complexes for Networks D and P were produced by merging highly overlapping cliques. As per results, 851 and 622 complexes were inferred from Network D and Network P, respectively. To assess the qualities of the identified complexes, we evaluated the recall and precision using the manually annotated protein complexes in the MIPS database, and compared the clique merging (CM) results to those generated from MCODE and MCL. As clearly delineated in Figure 2C, DC outperformed PE for all the three complex identification methods applied, even though they showed similar results in the pair-wise based evaluation in 3.1. PE generated higher recalls, but sacrificed significantly in precision. Within the same scoring scheme, CM consistently beat MCODE and MCL. DC was less sensitive to the complex identification method, and both DC-CM and DC-MCL combinations achieved better F_1-measures than the SA-IHC combination used in the original study. Unfortunately, as the IHC procedure described in the article was hard to follow, we were not able to test this algorithm on DC and PE. It is possible that IHC may perform better with these scoring schemes as they both outperformed SA in 3.1.

In order to explore the dynamic organization nature of the protein complexes, we further separated proteins in the merged complexes into more static core components and more dynamic attachment components, as depicted in the four examples

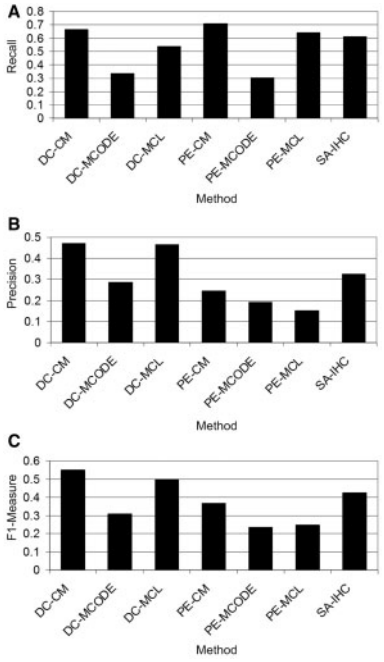

Fig. 2. Performance evaluation of the computationally predicted protein complexes based on the MIPS complex catalog. Recall (**A**), Precision (**B**), and F_1-measure (**C**) of six protein complexes sets found in this study through two different protein interaction affinity scoring schemes DC (Dice coefficient) and PE (Purification enrichment score), and three complex identification algorithms CM (Clique merging), MCODE (Molecular complex detection algorithm) and MCL (Markov clustering algorithm). The complexes predicted in the original paper based on the Socio-affinity index (SA) and iterative hierarchical clustering (IHC) is also plotted.

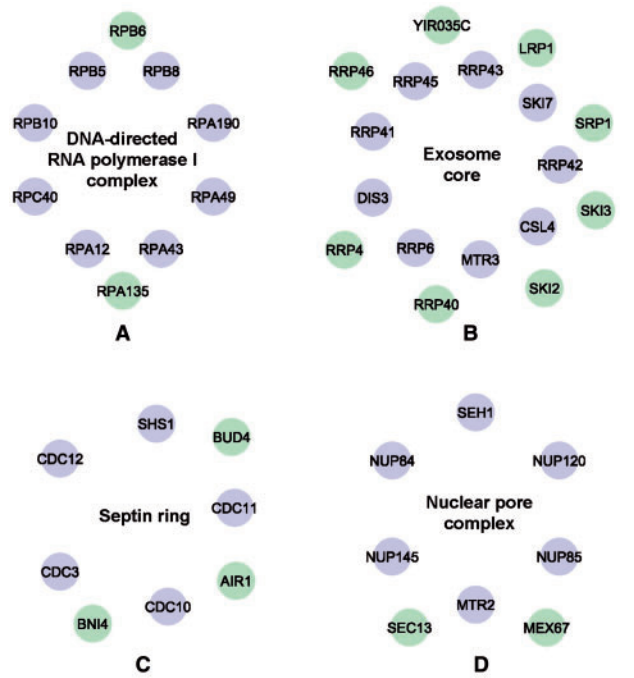

Fig. 3. Dynamic organization of protein complexes. Protein complexes are decomposed into more static core proteins (purple) and more dynamic attachment proteins (green).

in Figure 3. In complex A, all core proteins (purple) and attachment proteins (green) belong to the DNA-directed RNA polymerase I complex. Although we cannot exclude the possibility of the dynamic organization of this complex, the observation likely reflects technical errors that produced false negatives in individual complex isoforms (maximal cliques). In complex B, Ski complex proteins SKI2 and SKI3 were found as attachments to the Exosome core. This has been reported in the original analysis (Gavin *et al.*, 2006), and the association is critical to the cytoplasmic messenger RNA 3′–5′ decay (Orban and Izaurralde, 2005). In complex C, protein BNI4 was found as an attachment to the Septin ring complex. This attachment is required for normal septin function in yeast (DeMarini *et al.*, 1997). In complex D, protein MEX67 was found as an attachment to a core with the nuclear pore complex components, supporting previous report that nuclear mRNA export requires the interaction between MEX67 and the nuclear pore complex protein MTR2 (Santos-Rosa *et al.*, 1998). Cores and attachments for all 851 complexes are listed in the Supplementary Information.

3.4 Functional evaluation of inferred protein complexes

We evaluated the functional relevance of each of the 851 protein complexes identified from Network D using a hypergeometric enrichment test against all GO categories. 610 complexes showed high functional homogeneity, with a Bonferroni-adjusted *P*-value less than 0.05 in at least one of the biological process, molecular function, and cellular component categories. The most enriched GO categories for all the 851 complexes and associated *P*-values are listed in the Supplementary Information.

4 DISCUSSION

4.1 Computational framework of protein complex identification

With the wide-spreading interest of pull-down technology, a seamless integration of computational steps in identifying protein complexes from pull-down data is valuable. In this article, we introduced one such multi-stage framework. We particularly proposed a protein–protein interaction affinity-scoring scheme and demonstrated how graph theoretical approaches can be deployed to identify protein complexes.

We considered the affinity score between two proteins to be correlated with the strength of their co-purifications (co-appearances) over all pull-down experiments in which either (or both) of the proteins is purified, and used the Dice coefficient to quantify the affinity. Despite its simplicity, Dice coefficient performs better than existing scoring schemes in estimating the strength of protein interaction affinity, and provides an easy but effective alternative in analyzing pull-down data. Moreover, correlation based protein interaction affinity measurement can be easily adopted for the

semi-quantitative pull-down data. Recent studies have revealed the linear correlation between protein abundance and the spectrum count in MS/MS (Liu *et al*., 2004; Zhang *et al*., 2006). If such quantitative spectral counts for the preys from MS/MS are available, the information can be incorporated in the pair-wise affinity score calculation by simply replacing the Dice coefficient with the Pearson's correlation coefficient, for example. In this way, each bait-prey association is weighted by the prey abundance represented by corresponding spectral count, instead of a simple binary assignment, and thus better estimation of protein interaction affinity is expected.

Maximal clique enumeration-based algorithms have been applied in identifying modules from various biological networks (Baldwin *et al*., 2005; Palla *et al*., 2005; Spirin and Mirny, 2003; Zhang *et al*., 2006). Although the clique enumeration problem is NP-hard, recent progresses in exact, parallel and scalable computational solutions to this problem have made its genome-scale application feasible (Zhang *et al*., 2005, Park *et al*., 2007). As protein interaction networks are usually sparse with a scale-free distribution (Barabasi and Oltvai, 2004), even exact solutions can solve the problem quickly. For the networks used in this study, exact enumerations took less than 30 s on a regular computer with an Intel Xeon Processor 3.06 Ghz CPU and 1 GB Memory. As clique-based approaches require un-weighted graph as an input, how to select a threshold to convert the originally weighted graph into an un-weighted graph is a long-standing problem. Most often, a threshold is chosen empirically. In this article, we have employed the methods from the information retrieval field, and developed a knowledge-based, systematic approach for thresholding. One requirement of applying our method is the availability of a benchmark dataset to calculate recalls, precisions, and F_1-measures. Since yeast is the most extensively studied model organism, such benchmark can be found, and the MIPS database was used for our study. Even in a case when no such benchmark dataset is directly available in other organisms, alternative approaches could be considered. For example, one could use the protein complexes information in PDB (http://www.rcsb.org/pdb), and infer protein complexes in the organism under study through BLAST or PSI-BLAST. Moreover, one could use the GO-derived functional similarity scores described in Lord *et al*. (2003), or the functional association scores in the STRING database (http://string.embl.de/) to assist the threshold selection. Although these scores do not directly suggest physical interactions, correlation between functional association and protein interaction has been demonstrated in multiple studies.

Comparing to other complex identification algorithms, the clique merging approach has an obvious advantage of intuitiveness. For example, it might be possible to achieve better performance for MCL by changing the parameters used for the analysis, but setting the right parameters is non-trivial without a thorough understanding of the algorithm. Similarly, IHC might generate better results if DC or PE were used, however, the procedure described in the original paper is not easily reproducible. Although CM is much easier to follow than IHC, it shares the important feature of being able to reveal the dynamic organization of the protein complexes, a feature that is missing in other complex identification algorithms such as hierarchical clustering, MCL and MCODE.

Evaluating the biological relevance of identified protein complexes is a crucial step in complex identification. In this study, we evaluated the predicted complexes against the GO annotation. The assumption behind GO-based evaluation is that proteins belonging to a real biological complex should possess a related molecular function, be located in a similar cellular component, and be involved in an associated biological process. As GO aims at a structured, controlled vocabulary that describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner (Ashburner *et al*., 2000), these evaluation methods should be applicable in any organisms with a sufficiently good GO annotation. If the GO annotation is not available, other evaluation approaches could be used. For example, as proteins belong to a real complex should be generally conserved through the evolution process to act as an integrated functional unit, one could use the 'phylogenetic correlation' to evaluate the biological significance of the complexes (Poyatos and Hurst, 2004).

4.2 Exploring possible sources of false identifications

As shown in Figure 3A, clique merging might help reduce false negative identifications by integrating information from highly overlapping cliques. Such kind of false negatives can be attributed to the errors (undetected true interactions) in individual pull-down experiments, and can be reduced by our computational framework. Here we further explore other possible sources of false positive and false negative identifications by manually going through the identified complexes. We have found that co-purification of different protein complexes as mediated by a common non-protein molecule, such as DNA, might lead to false positives, while the protein identification bias could generate false negatives.

The purification patterns of two proteins can be similar if there exists a common non-protein molecule, such as DNA. A protein may be a part of a complex that interacts with the DNA, and the DNA pulls down the whole complex. Such an interaction is likely maintained during the purification step. Consequently, our scheme of assigning protein interaction can detect both direct physical protein interactions and indirect associations mediated by a non-protein molecule. One salient example is complex 21 (see Supplementary Information), which consists of several components that are intra-associated, but inter-disassociated (Figure 4). In other words, proteins in the same component physically interact, but proteins from different components may not. However, different components in complex 21 are found to interact with a common non-protein molecule, DNA. These components include the RSC complex (red), the RNA polymerase III complex (blue), the Mot1 complex (purple) and the Nucleosomal protein complex (green).

An additional example is complex 39 (Supplementary table), which includes proteins from independent components such as the replication factor A complex (RFA1, RFA2 and RFA3), the RecQ helicase-Topo III complex (TOP3), and the DNA mismatch repair complex (MSH2, MSH3 and MSH6).

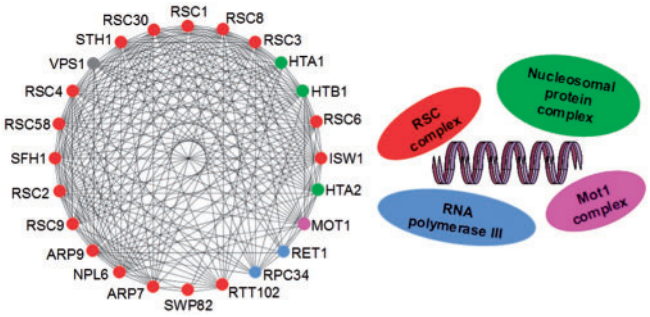

Fig. 4. Involvement of multiple protein complexes in a computationally predicted complex. The network among the proteins involved in the predicted complex is shown at the left, in which proteins are colored by their specific complex memberships, as indicated in the right cartoon. Each complex interacts with DNA independently.

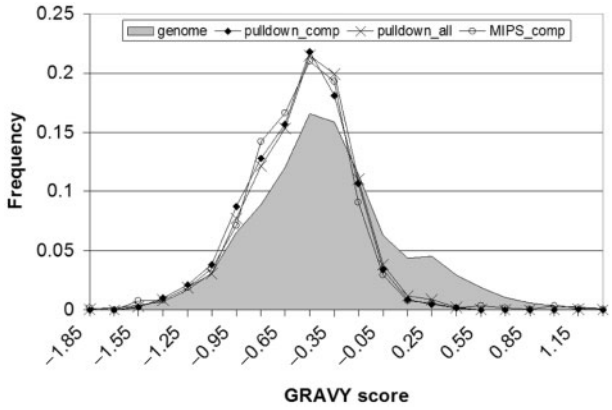

Fig. 5. Hydrophobicity score (GRAVY) distribution for proteins found in the entire genome (shaded area), identified protein complexes for the pull-down data (solid diamond line), the original pull-down data (cross line) and the MIPS complexes (circle line).

Despite differences in their functional roles, these components all bind to single strand DNAs. Other proteins in this complex also interact with single strand DNAs and participate in DNA repair process.

Although not archived in the current protein interaction database, we cannot rule out the possibility of direct interaction among the above complexes owing to the complexity of the cell. However, the above examples suggest that association through an additional molecule such as DNA or RNA is a possible explanation for the co-purification. Treatment of protein samples with nuclease to separate individual complexes from their common interacting nucleic acids may help distinguishing direct interactions from DNA or RNA-mediated co-purification, and eliminating the later to reduce false positives in complex identification.

While manually checking the identified complexes, we also found that protein complexes with hydrophobic subunits are usually incompletely identified, with the hydrophobic subunits missing, and therefore generating high false negatives. For example, complex 93 (see Supplementary Information) involves seven subunits of the yeast vacuolar (H)-ATPases (V-ATPases), which are ATP-dependent proton pumps that acidify intracellular vacuolar compartments (Compton *et al.*, 2006; Graham *et al.*, 2003). The V-ATPases consist of two separable domains: V1 and V0. The V1 domain associates with the vacuolar membrane and is composed of eight hydrophilic subunits catalyzing ATP hydrolysis. The V0 subcomplex is hydrophobic, composed of seven subunits and transports protons across the membrane (Flannery *et al.*, 2004). Complex 93 comprises six out of the eight (75%) hydrophilic subunits (VMA4, VMA5, VMA7, VMA8, VMA10 and VMA13), while only one out of the seven (14%) hydrophobic subunits (VPH1).

This observation suggests the insensitivity of the current pull-down technology in identifying hydrophobic proteins. We empirically tested this by computing hydrophobicity using the GRAVY score of proteins found in (1) the entire genome, (2) the identified protein complexes, (3) the original pull-down data and (4) the MIPS complexes. The GRAVY score is measured by averaging hydrophobicity indices of all amino acids in a protein (Kyte and Doolittle, 1982). Each amino acid is given a hydrophobicity index between 4.6 and −4.6, where 4.6 is the most hydrophobic and −4.6 is the most hydrophilic. Figure 5 shows a hydrophilic bias in proteins comprising the identified complexes. Hydrophobic proteins are clearly under-represented in the identified complexes (solid diamond line) as compared to the entire genome (shaded area). Moreover, the bias is not generated in the computational process for complex identification, as it exists in the original pull-down data (cross line) as well. Analyzing the GRAVY scores in known complexes taken from the MIPS database also showed similar hydrophilic bias in proteins comprising those complexes (circle line). This may suggest that the hydrophilic bias is not unique to the pull-down technology; instead, it is common to most of the analytical technologies used for generating the data in the MIPS database. However, it is also possible that hydrophobic proteins are less likely to take part in multi-protein complexes. If the hydrophilic bias is truly caused by the analytical technologies, it can't be easily fixed by computational analyses, and will require improvements from the experimental side.

ACKNOWLEDGEMENTS

We are thankful to the reviewers for the insightful suggestions that helped us improve the manuscript. This research has been supported by the 'Exploratory Data Intensive Computing for Complex Biological Systems' project from U.S. Department of Energy (Office of Advanced Scientific Computing Research, Office of Science). The work of N.F.S. was also sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DE-AC05-00OR22725.

Conflict of Interest: none declared.

REFERENCES

Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists,. *Cell*, **92**, 291–294.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,. *Nat. Genet.*, **25**, 25–29.

Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

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

Baldwin,N.E. *et al.* (2005) Computational, integrative and comparative methods for the elucidation of genetic co-expression networks. *J. Biomed. Biotechnol.*, **2**, 172–180.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Butland,G. *et al.* (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **433**, 531–537.

Collins,S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell Proteomics*, **6**, 439–450.

Compton,M.A. *et al.* (2006) Vma9p (subunit e) is an integral membrane V0 subunit of the yeast V-ATPase. *J. Biol. Chem.*, **281**, 15312–15319.

DeMarini,D.J. *et al.* (1997) A septin-based hierarchy of proteins required for localized deposition of chitin in the *Saccharomyces cerevisiae* cell wall. *J. Cell Biol.*, **139**, 75–93.

Drewes,G. and Bouwmeester,T. (2003) Global approaches to protein-protein interactions, *Curr. Opin. Cell Biol.*, **15**, 199–205.

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

Flannery,A.R. *et al.* (2004) Topological characterization of the c, c′, and c″ subunits of the Vacuolar ATPase from the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.*, M406767200.

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

Gavin,A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gilchrist,M.A. *et al.* (2004) A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, **20**, 689–700.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.

Graham,L.A. *et al.* (2003) Structure and assembly of the yeast V-ATPase. *J. Bioenerg. Biomembr.*, **35**, 301–312.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

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

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Liu,H. *et al.* (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193–4201.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Orban,T.I. and Izaurralde,E. (2005) Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. *RNA*, **11**, 459–469.

Palla,G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–818.

Park,B.H. *et al.* (2007) Data-driven, data-intensive computing for modelling and analysis of biological networks: application to bioethanol production. *J. Phys. Conf. Ser.*, **78**, 012061.

Poyatos,J. and Hurst,L. (2004) How biologically relevant are interaction-based modules in protein networks? *Genome Biol.*, **5**, R93.

Santos-Rosa,H. *et al.* (1998) Nuclear mRNA export requires complex formation between Mex67p and Mtr2p at the nuclear pores. *Mol. Cell Biol.*, **18**, 6826–6838.

Scholtens,D. *et al.* (2005) Local modeling of global interactome networks. *Bioinformatics*, **21**, 3548–3557.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Yu,H. *et al.* (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, **22**, 823–829.

Zhang,B. *et al.* (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucl. Acids Res.*, **33**, W741–W748.

Zhang,B. *et al.* (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics*, **5**, 16.

Zhang,B. *et al.* (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.*, **5**, 2909–2918.

Zhang,C. *et al.* (2006) Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *J. Proteome Res.*, **5**, 801–807.

Zhang,Y. *et al.* (2005) Genome-scale computational approaches to memory-intensive applications in systems biology. In *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, **12**.