A MAX-FLOW BASED APPROACH TO THE IDENTIFICATION OF PROTEIN COMPLEXES USING PROTEIN INTERACTION AND MICROARRAY DATA (EXTENDED ABSTRACT)

JIANXING FENG*
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
**Email: fengjx06@mails.tsinghua.edu.cn*

RUI JIANG
MOE Key Laboratory of Bioinformatics, Bioinformatics Division TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.
Email: ruijiang@tsinghua.edu.cn

TAO JIANG
Department of Computer Science, University of California, Riverside, CA 92521.
Email: jiang@cs.ucr.edu

The emergence of high-throughput technologies leads to abundant protein-protein interaction (PPI) data and microarray gene expression profiles, and provides a great opportunity for the identification of novel protein complexes using computational methods. Although it has been demonstrated in the literature that methods using protein-protein interaction data alone can successfully predict a large number of protein complexes, the incorporation of gene expression profiles could help refine the putative complexes and hence improve the accuracy of the computational methods.

By combining protein-protein interaction data and microarray gene expression profiles, we propose a novel *Graph Fragmentation Algorithm* (GFA) for protein complex identification. Adapted from a classical max-flow algorithm for finding the (weighted) densest subgraphs, GFA first finds large (weighted) dense subgraphs in a protein-protein interaction network and then breaks each such subgraph into fragments iteratively by weighting its nodes appropriately in terms of their corresponding log fold changes in the microarray data, until the fragment subgraphs are sufficiently small. Our extensive tests on three widely used protein-protein interaction datasets and comparisons with the latest methods for protein complex identification demonstrate the superior performance of our method in terms of accuracy, efficiency, and capability in predicting novel protein complexes. Given the high specificity (or precision) that our method has achieved, we conjecture that our prediction results imply more than 200 novel protein complexes.

1. INTRODUCTION

With the advances in modern biophysics and biochemistry, it has been widely accepted that the rise of complicated biological functions is largely due to the cooperative effects of multiple genes and/or gene products. This understanding leads to the emergence of high-throughput technologies for identifying interactions between biological molecules and results in the prosperity of interactomics in the post genomics and proteomics era. For example, with the use of yeast two-hybrid assays [1–3] and pull-down mass spectrometry experiments [4, 5], genome-wide protein-protein interactions (PPIs) have been identified and encoded into global PPI networks for the model species *Saccharomyces cerevisiae* (*i.e.* baker's yeast) [6–8]. With the improvement of instruments and increase in the throughput, these technologies have also been applied to identify interactions of human proteins, providing an increasing understanding of the global human PPI network [9]. Parallel to the boom of high-throughput identification of PPIs, genome-wide microarray experiments regarding the expression of genes and their products across a number of different conditions have also been conducted and resulted in publicly available databases such as the *gene expression omnibus* [10].

*To whom correspondence should be addressed.

As a major form of the cooperative effects of two or more proteins, protein complexes play important roles in the formation of complicated biological functions such as the transcription of DNA, the translation of mRNA, and many others. Traditionally, protein complexes are identified using experimental techniques such as the X-ray crystallography and the nuclear magnetic resonance (NMR), or computational methods such as protein-protein docking. These methods, though successful, can hardly meet the requirement of identifying all protein complexes in known organisms, due to the large number of proteins, the cost of biological experiments, and the limited availability of protein structure information. On the other hand, since a protein complex is composed of a group of two or more proteins that are associated by stable protein-protein interactions, computational methods that can make use of abundant data given by the above high-throughput technologies have been demonstrating increasing successes [11–15].

Many studies use PPI data alone for the purpose of identifying protein complexes or biologically functional modules. These methods assume that densely connected components in PPI networks are likely to form functional modules and hence are likely to be protein complexes [16]. With this assumption, the methods generally use the density of interactions as a main criterion and identify protein complexes by finding dense regions in PPI networks. To mention a few, Bader and Hoque proposed a clustering algorithm called MCODE that isolates dense regions in a PPI network by weighting each vertex according to the topological properties of its neighborhood [11]. Andreopoulos *et al.* presented a layered clustering algorithm that groups proteins by the similarity of their direct neighborhoods [17]. Spirin and Mirny applied three methods (*i.e.* clique enumeration, super paramagnetic clustering, and Monte Carlo simulation) to the MIPS PPI network for yeast [7] and produced about 100 dense subgraphs that were predicted to be protein complexes [12]. Their results were found to be superior to many others in terms of accuracy. Pei and Zhang introduced the use of a subgraph quality measure as well as a "seed-refine" algorithm to search for possible subgraphs in a PPI network [13]. King *et al* gave a clustering algorithm based on restricted neighborhood search to partition a PPI network into clusters using some cost function [18]. Bu *et al.* introduced a spectral method derived from graph theory to uncover hidden topological structures that consist of biologically relevant functional groups [19]. Li *et al.* found maximal dense regions by merging local cliques according to their affinity [14]. In a subsequent work, Li *et al.* devised an algorithm, called DECAFF, to address two major issues in current high-throughout PPI data, namely, incompleteness and high data noise [15].

Another group of methods combine PPI data and microarray gene expression profiles for the purpose of identifying protein complexes. These methods regard PPIs as static descriptions of the potential collaborative effects between proteins and treat gene expression profiles as dynamic information of genes under various conditions. Since proteins of a complex usually work together to complete certain biological functions, and there exists a simple mapping between genes and their products, the combination of PPI and microarray gene expression data can clearly help the discovery of protein complexes or functional modules [20, 21]. Moreover, such a combination is also often used in the search for regulatory modules and signalling circuits [22]. As an example, Guo *et al.* identified condition-responsive subnetworks in a PPI network by weighting its edges based on gene expression profiles [23].

Besides these methods, there exist some other methods that aim at identifying protein complexes by using comparative interactomics. For example, Sharan *et al.* identified protein complexes by a comparative analysis of the PPI networks from yeast and bacteria [24]. Hirsh and Sharan developed a probabilistic model for protein complexes that are conserved across two species and applied it to yeast and fly [25]. These methods based on comparative analysis require the availability of quality PPI networks from multiple species and can only identify protein complexes conserved in multiple species.

Despite differences in the approach and use of data, most of the computational methods mentioned above follow a bottom-up local search strategy. For example, Li *et al.* first finds small dense subgraphs (or components) in a PPI network and then merges these components gradually to form protein complex-like subgraphs [15]. Pei and Zhang greedily

expands some carefully selected seed subgraphs until a given criterion is met [13]. Because a local search strategy does not return the optimal solutions in general, the above bottom-up methods are not guaranteed to find the densest subgraphs in the input PPI network and therefore may miss many important protein complexes that are truely dense.

To overcome this drawback, we present a top-down method to identify protein complexes that explicitly utilizes the density information in PPI networks as well as microarray gene expression profiles. This work combines the classic maximum network-flow based *Densest Subgraph Algorithm* (DSA) [26] to find the densest subgraphs with a novel application of microarray data. Our algorithm, named the *Graph Fragmentation Algorithm* (GFA), first finds dense subgraphs in a PPI network (many of which could potentially be large), and breaks each of them into fragments iteratively by weighting its nodes appropriately in terms their corresponding log fold changes in the microarray data, until the fragment subgraphs are sufficiently small. In order to test the performance of our method, we apply GFA to three widely used yeast PPI networks (*i.e.* the MIPS, DIP and BioGRID PPI networks) and compare our predictions with the known protein complexes in the MIPS database as well as with those of the latest methods for protein complex identification (that are not based on comparative analysis) [12, 15]. The test results clearly demonstrate the superior performance of our method in terms of accuracy, efficiency, and capability in predicting novel protein complexes. For example, GFA could be tuned to achieve sensitivity 73% and specificity 85% simultaneously on the DIP PPI network. Our method also provides a ranking of the predicted complexes, taking advantage of the multiple conditions (or samples) in the microarray expression data. Putative complexes with higher ranks are believed to have a larger likelihood to be true protein complexes. Moreover, our predictions result in more than 200 highly ranked dense subgraphs that share no common proteins with the known complexes in MIPS and are thus likely to be novel protein complexes.

For the convenience of presentation, some of the figures and tables are omitted in the main text and given in the appendix.

2. MATERIALS AND METHODS

2.1. Data sources

Three PPI datasets concerning *Saccharomyces cerevisiae* are used. The first one is the MIPS protein-protein interaction network dataset [7], which is believed to contain the most credible PPI data and will simply be denoted as MIPS-PPI. The second one is the DIP protein-protein interaction network dataset [6], denoted as DIP-PPI. The third one is BioGRID protein-protein interaction dataset [8], which is the most comprehensive one and will be denoted as BioGRID-PPI. Because a PPI network is treated as an undirected simple graph, at most one edge will be kept between any pair of proteins. The numbers of nodes (or edges) in the MIPS, DIP and BioGRID PPI networks are 4,554 (or 12,319), 4,932 (or 17,201) and 5,201 (or 71,044), respectively.

We retrieved 58 sets of microarray gene expression data concerning yeast from the GEO database [10]. The expression levels have been log transformed, and the microarray data contain a total of 716 samples (or conditions). Since the genes expressed in each sample are different, and they could also be different from the genes contained in a PPI network, we will use a sample of the microarray data on a PPI network if it covers at least 90% of the genes in the network. This criterion results in 477, 571 and 623 samples that can be applied to the MIPS, DIP and BioGRID PPI networks, respectively.

As in the previous studies [11, 12, 14, 15], the MIPS complex database [7] is used as the benchmark (*i.e.* the truth) to evaluate the protein complexes predicted by our method. This database contains protein complexes manually verified and those identified by high throughput experiments. We denote the set of complexes verified manually as MIPS-MAN and the set of all protein complexes in the database as MIPS-ALL. Furthermore, our GFA algorithm only outputs connected subgraphs, but many complexes in MIPS-ALL are not connected in the above PPI networks. To evaluate our results in a more reasonable way, we decompose each MIPS complex into connected components according to the PPI network under study. We will use MIPS-MAN-COMP and MIPS-ALL-COMP to denote the sets of connected complex components obtained from MIPS-

MAN and MIPS-ALL, respectively. Finally, since GFA does not output subgraphs consisting of a single node or edge (because they are trivial), all complexes or complex components with sizes 1 or 2 are removed from MIPS-MAN-COMP and MIPS-ALL-COMP. Note that the contents of MIPS-MAN-COMP and MIPS-ALL-COMP depend on the underlying PPI network used. In Table 1, we summarize sizes of the benchmark sets with respect to each PPI network.

Table 1. Sizes of the benchmark sets of protein complexes with respect to each PPI network.

Benchmark	MIPS-PPI	DIP-PPI	BioGRID-PPI
MIPS-MAN-COMP	100	114	134
MIPS-ALL-COMP	272	759	804

2.2. An outline of GFA

A PPI network is considered as an undirected simple graph, where nodes represent proteins and edges denote interactions between two nodes. To find dense subgraphs, various computational methods have been proposed (see the introduction section). Nevertheless, these methods are mostly based on local search strategies and can hardly find the densest subgraphs in a given PPI network.

A widely used definition of the density for a subgraph is $\delta = 2 \cdot |E|/(|V| \cdot (|V|-1))$ [11, 12], where E and V denote the sets of edges and nodes in the subgraph, respectively. An alternative definition is $\delta = |E|/|V|$. In general, the former definition favors small subgraphs (see Spirin and Mirny [12]), while the latter favors large subgraphs. However, both definitions are sensitive to the size of a subgraph. In fact, when the first definition is applied, we have to add a lower bound on $|V|$ to make the result interesting. Considering this, we use the latter definition of density in this work, since there exists an elegant algorithm to find the densest subgraph under this definition. Besides, our experimental results also demonstrate that this definition of density works well in finding protein complexes.

Theoretically, the problem of finding a subgraph with the greatest density in a graph under the first definition is much harder than that under the second one. The problem under the first definition is basically equivalent to finding the largest clique in a graph, a classical NP-hard problem in theoretical computer science [27]. However, there is an elegant and fast algorithm to solve the problem under the second density definition. This algorithm, simply denoted as DSA (*i.e.* the *Densest Subgraph Algorithm*), finds the densest subgraph in a graph by iteratively solving a series of maximum flow problems and has the time complexity of $O(|E| \cdot |V| \cdot \log(|V|^2/|E|))$ [26]. Although DSA can be iterated to find many dense subgraphs in a PPI network, this approach (alone) will likely not work well in terms of finding protein complex-like subgraphs, since it tends to find large dense subgraphs while protein complexes are usually small (*i.e.* containing no more than 20 proteins). Nevertheless, DSA will form the core ingredient of our GFA algorithm for finding protein complexes. GFA actually uses a generalized version of the second density definition: $\delta = |E|/w(V)$, where we assume that the nodes in the graph are weighted (*e.g.* using the log fold changes in some sample of microarray data) and $w(V)$ denotes the total weight of the nodes in the subgraph. The DSA algorithm mentioned above also works for this generalized definition.

GFA consists of two phases: (1) identify candidate subgraphs from the input PPI network using a single sample of gene expression data, and (2) combine candidate subgraphs from multiple samples to form a ranked list of predicted protein complexes. The basic idea behind the first phase is to iterate DSA to obtain (large) dense subgraphs and then break each large dense subgraph into fragment subgraphs by weighting its nodes appropriately using the log fold changes of the nodes in the sample. This phase is executed on each sample separately. In the second phase, we detect and remove redundant (or overlapping) subgraphs found using different samples and rank the subgraphs according to the times that they are found in all samples. The worst case time complexity of GFA is largely determined by the time complexity of phase 1, which is $O(|E| \cdot |V|^2 \cdot \log(|V|^2/|E|) \cdot \mathit{MaxIter} \cdot \mathit{SampleSize})$, where *MaxIter* is a parameter defined below and *SampleSize* is the number of samples of the microarray data used in the computation.

2.3. Identification of candidate subgraphs

Again, the idea is to break each large dense subgraph found by DSA into smaller ones by weighting its nodes appropriately using gene expression data. Recall that the gene expression data contains hundreds of samples. In this phase, we look at one sample at a time. The log fold change of the expression value of gene A in the sample is denoted as $expr(A)$. At the beginning, the nodes in the input PPI network with degree 1 are removed. This will reduce the size of the network and will not affect our final result much because a dense subgraph is not expected to contain nodes with degree 1. Then we weight every node uniformly as 1 and run DSA to find the densest subgraph. If the size of the subgraph identified is above a certain threshold (denoted as *MaxSize*), the weight of each node A in the subgraph is multiplied by a factor of $e^{-expr(A)}$ and DSA is applied again to the subgraph. The effect of this multiplication is that the weights of highly differentially expressed genes in the subgraph are reduced. The exponentially factor of $e^{-expr(A)}$ in this adjustment was chosen empirically. Note that, since DSA maximizes the ratio $|E|/w(V)$, it tends now to find a subgraph with nodes bearing small weights. In other words, the above weighting adjustment favors genes that are highly differentially expressed in the sample. As an effect, some nodes with large weights may be removed and the subgraph is fragmented. This step is executed iteratively, until either a given maximum iteration count (denoted as *MaxIter*) is reached or the size of the subgraph is below *MaxSize*.

Once a sufficiently small dense subgraph is found, all the nodes in the subgraph and all the edges adjacent to any one of the nodes in the subgraph are removed from the PPI network. Then, we remove all the nodes with degree 1 in the remaining network and reiterate the above process of using DSA to find the next sufficiently small dense subgraph. The whole process ends when the PPI network exhausts.

Now we discuss the two parameters *MaxSize* and *MaxIter*. *MaxSize* determines the maximum size of a subgraph found by GFA. In principle, it should be set as the largest possible size of an expected protein complex component (see Section 2.1 for the definition of protein complex components) for a given PPI network. For example, in our experiments, for MIPS-PPI, we select 20 as the bound because the maximum size of a protein complex component in MIPS-ALL-COMP does not exceed 20. However, our experiments show that GFA is quite robust with respect to this parameter and it is fine to use a slightly larger *MaxSize*, especially when the microarray data contains many samples, because only the "core" of a subgraph will be found in multiple samples. For example, we also tried to set *MaxSize* as 30 on MIPS-PPI and got almost the same result. The parameter *MaxIter* reflects how strictly we enforce the size bound. A small *MaxIter* may lead to subgraphs with sizes above *MaxSize*. This is useful when there are a few protein complexes that are very dense and much larger than the other protein complexes and we do not want to make *MaxSize* too large. So, the parameters *MaxSize* and *MaxIter* together control the sizes of the output subgraphs. Fortunately, our test results show that the final result of GFA is not very sensitive to either of these parameters.

2.4. Combining candidate subgraphs

The above phase 1 of GFA generates a set of candidate subgraphs for each sample of the microarray data. However, many of these subgraphs are duplicated or similar. We define the *overlap score* of two sets A and B as $overlap(A, B) = 2|A \bigcap B|/(|A| + |B|)$. This step aims to distill promising subgraphs from the candidate subgraphs. More specifically, duplicates and trivial subgraphs with sizes 1 or 2 are removed and similar subgraphs will be merged. However, because of the drastic difference in the densities of the three PPI networks considered in this paper, we have to use two different strategies in this phase. We use a simple strategy for MIPS-PPI and a more general, slightly more complicated strategy for DIP-PPI and BioGRID-PPI. The latter networks are much denser.

2.4.1. *The simple strategy*

Here we simply count the frequency of each candidate subgraph in all samples and rank the subgraphs by their frequencies. A subgraph with a high frequency is expected to be a promising protein complex (or complex component), since it is dense and many of

its nodes are highly differentially expressed in multiple samples. After the frequency of each candidate subgraph is calculated, we check if two candidate subgraphs overlap. If the overlap score between two graphs (computed using their vertex sets) is above a certain cutoff (denoted as *MaxOverlap*), they are deemed duplicates and the one with a smaller frequency is simply removed.

Note that, the result of this removal step depends on the order that we process the candidate subgraphs. For example, consider subgraphs A, B, C with sizes a, b, c respectively, with $a > b > c$. A overlaps with B and B overlaps with C, but A and C do not overlap according to the given overlap criterion. If A and B are processed after B and C are processed, only A remains. But if we process A and B first, then both A and C will remain. So, for consistency, we consider the pairs of candidate subgraphs in decreasing order of their overlap. This simple strategy is also applied to the following more general strategy and the combined strategy.

As shown in our experimental results, this simple strategy works very well on MIPS-PPI, mostly due to its sparsity. It also works on the DIP-PPI and BioGRID-PPI, although it appears to be too conservative in dealing with similar candidate subgraphs.

2.4.2. *The more general strategy*

Dense subgraphs in dense PPI networks tend to be larger and we cannot expect that the subgraph corresponding to a real protein complex will be discovered by GFA from many samples exactly, since the samples generally have different expression levels. Thus, the simple strategy is too conservative for this situation. Moreover, when the input PPI network is large (such as BioGRID-PPI), DSA becomes quite slow and we may not want to spend the time to examine every sample of the microarray data. Hence, in this case, we need revise the definition of frequency and introduce a more general strategy to combine results from different samples. Our basic idea here is to combine similar candidate subgraphs. Due to the page limit, this general strategy and a combined method to integrate it with the simple stragegy is omitted in this extended abstract but will be given in the full paper.

3. RESULTS

3.1. Some useful definitions and notations

Before discussing the results, we need introduce several definitions and notations. First, since we will mainly validate our predictions against benchmark protein complexes in MIPS, we define the *effective size* of a predicted protein complex as the number of proteins shared by this predicted complex and the complexes in the relevant benchmark (*i.e.* MIPS-MAN-COMP or MIPS-ALL-COMP). Obviously, we could only hope to validate predicted protein complexes with large effective sizes. We say that a protein complex (component) A in a benchmark set is *identified* by a predicted complex B with some cutoff p if $|A \bigcap B|^2/(|A| \cdot |B|) \geq p$. Since a commonly used value for p in the literature is 0.2 [11, 15], we say that B *matches* A if A is identified by B with the cutoff $p = 0.2$. The following several (shorthand) notations will be convenient to use in tables and figures:

(1) *predicted* (or P for short): The number of predicted protein complexes.
(2) *matched* (or M for short): The number of predicted complexes that match some protein complex component in the relevant benchmark set.
(3) $P_{e \geq n}$: The number of predicted complex with effective sizes at least n.
(4) $P_{e=n}$: The number of predicted complexes with effective sizes exactly n.
(5) *identified*(p) (or $I(p)$ for short): The number of complex components in the relevant benchmark set that have been identified by any one of the predicted complexes with cutoff p. This parameter generally reflects the sensitivity of the prediction. Although the widely used p value is 0.2, we will also consider $p = 0.5$ since it could provide more insight into the prediction result.
(6) *effective specificity* : The number of predicted protein complexes that match complex components in the relevant benchmark set divided by the number of predicted complexes with effective sizes at least 2. In other words, it is equal to $M/P_{e \geq 2}$. Hereafter, the term *specificity* refers to *effective specificity* unless stated otherwise.

Note that, because of overlaps in the predicted results and the benchmark sets, the number of matched predicted complexes may not be the same as the number of the identified complex components in the relevant benchmark. In other words, M may be different from $I(0.2)$. For example, $M = 1$ and $I(0.2) = 2$ means that the result consists of one predicted complex that matches (and perhaps contains) two complex components in the benchmark. On the other hand, $M = 2$ and $I(0.2) = 1$ means that the result consists of two predicted complexes that match (and are perhaps contained in) a single benchmark complex component. In general, let us define the *efficiency* of a prediction as the ratio between $I(p)$ and M. Clearly, with the same $I(p)$ value (*i.e.* the same sensitivity), we would prefer prediction results with a small M since a smaller M would imply a higher efficiency. In our test results, an important property is that the number $P_{e\geq 2}$ is very close to M when the parameter *MinFrequency* is large. Hence, among the protein complexes predicted by GFA, a top ranked protein complex either has a match in the benchmark or has a very small effective size (*i.e.* it is largely disjoint from the benchmark).

3.2. Matching to the benchmark

For succinctness, we give a detailed report of the prediction result on MIPS-PPI and their matches in MIPS-MAN-COMP, and sketch the other results. As mentioned before, on MIPS-PPI, the simple strategy in phase 2 of GFA is applied. MIPS-MAN-COMP contains 100 complex components with respect to MIPS-PPI. The actual output of GFA depends on the parameters *MinFrequency* and *MaxOverlap* involved in phase 2. By choosing different values for these two parameters, we obtain prediction results with different combinations of sensitivity and specificity. In general, a big *MinFrequency* implies a high specificity and a low sensitivity.

Figure 1 shows the number of predicted complexes and their matching benchmark complexes with respect to various combinations of *MinFrequency* and *MaxOverlap*. An interesting observation is the high accordance among $P_{e\geq 2}$, M and $I(0.2)$. The accordance between the former two terms implies (as mentioned before) that a predicted protein complex has either a match in the benchmark or a very small (*i.e.* at most 1) effective size. While the accordance between the latter two terms indicates that GFA is very efficient and the accordance between the 1st and 3rd terms implies that GFA maintains a good (effective) specificity. The comparison between the prediction results for $MaxOverlap = 0.2$ and $MinOverlap = 0.5$ shows that the parameter *MaxOverlap* has little impact when *MinFrequency* is greater than 2. This means that the predicted protein complexes in general do not overlap too much with each other.

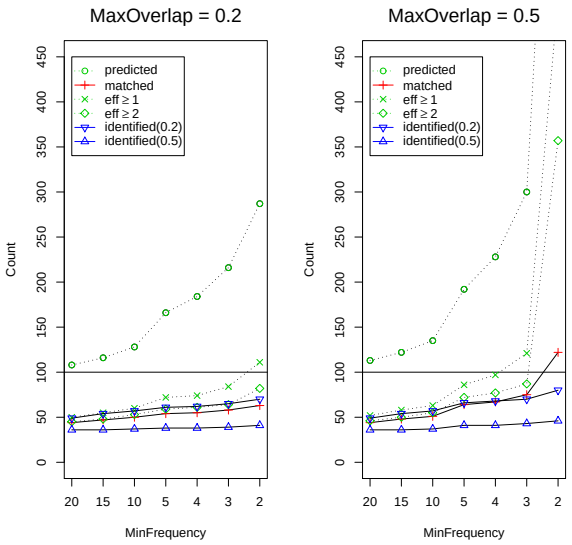

Fig. 1. Protein complexes predicted by GFA on MIPS-PPI and their matches in MIPS-MAN-COMP. Two *MaxOverlap* values, 0.2 (left) and 0.5 (right), are considered. The notation $eff \geq 1$ stands for $P_{e\geq 1}$.

Table 2 gives the detailed results when two extremal values of *MinFrequency* are considered, with *MaxOverlap* being fixed at 0.2. In the first group of results where MIPS-MAN-COMP is used as the benchmark (*i.e.* the more reliable benchmark), when $MinFrequency = 20$, 49 out of the 100 complex components in the benchmark are identified. Although the sensitivity is only 49%, 44 out of the 45 predicted complexes with large effective sizes (*i.e.* at least 2) have matches in the benchmark, which means that the (effective) specificity of this prediction is 97%. Moreover, among the 64 predicted protein complexes that have no matches in the benchmark, 58 of them have zero effective size. In other words, their proteins do not appear in the bench-

mark at all. We conjecture that these 58 predicted complexes represent novel protein complexes (or at least are involved in novel protein complexes).

On the other hand, if $MinFrequency = 2$, the predicted protein complexes identify 70% of the complex components in the benchmark, but the specificity drops. Among the 82 predicted complexes with large effective sizes, 63 of them match complex components in the benchmark, *i.e.* the specificity is 77%. Comparing the values of $I(0.2)$ and $I(0.5)$, we see that GFA could identify 21 additional complex components in the benchmark using $MinFrequency = 2$ than using $MinFrequency = 20$, as suggested by the values of $I(0.2)$, but only 5 of them have been identified with a high accuracy, as suggested by the values of $I(0.5)$. This means that generally speaking, complexes predicted by GFA with higher frequencies identify complex components in the benchmark more accurately. In other words, a predicted complex with a higher rank is more likely to be (or at least to be involved in) a true protein complex. Again, we conjecture that the 176 predicted complexes that share no proteins with the benchmark complexes represent novel complexes. Note that, the 176 novel complexes actually include the 58 novel complexes mentioned above.

By examining the sets of subgraphs output by GFA with $MinFrequency = 20$ and $MinFrequency = 2$ in detail, we find that the former set could already identify most of the large and dense complex components in the benchmark MIPS-MAN-COMP. 18 out of the 30 complex components in the benchmark missed by the latter (larger) set are trees with at most 6 nodes, and the remaining 12 missing complex components have densities at most 2 in MIPS-PPI. The details of these results are not shown here. It is also clear that GFA achieves very good efficiency in both cases, with the ratio $I(0.2)/M$ being about 1.11.

In the second part of Table 2 where MIPS-ALL-COMP is used as the benchmark, when $MinFrequency = 20$, 57 out of the 61 predicted complexes with large effective sizes have matches in the benchmark. Thus, we still have the same property that a protein complex predicted by GFA with a high frequency has either a match in the benchmark or a very small effective size. The sensitivity and specificity of the prediction are generally a bit worse than those using MIPS-MAN-COMP. The sensitivity is $135/272 = 50\%$ for $MinFrequency = 20$ and $179/272 = 66\%$ for $MinFrequency = 2$ and the specificity is $57/61 = 93\%$ for $MinFrequency = 20$ and $89/127 = 70\%$ for $MinFrequency = 2$. This is perhapss due to the noise in MIPS-ALL.

Table 2. Protein complexes predicted by GFA on MIPS-PPI and their matches in MIPS-MAN-COMP and MIPS-ALL–COMP. MAN and ALL stands for MIPS-MAN-COMP and MIPS-ALL-COMP, respectively. f stands for $MinFrequency$, and $MaxOverlap$ is set to 0.2.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
MAN, *f=20*	108	49	36	44	45	58
MAN, *f=2*	287	70	41	63	82	176
ALL, *f=20*	108	135	82	57	61	43
ALL, *f=2*	287	179	91	89	127	129

It is interesting to note that only a small fraction of the novel protein complexes conjectured above have matches in MIPS-ALL-COMP (*i.e.* at most $15 = 58 - 43$ for $MinFrequency = 20$ and at most $47 = 176 - 129$ for $MinFrequency = 2$).

The GFA prediction results on DIP-PPI and BioGRID-PPI and their matches in MIPS-MAN-COMP are sketched below. The details are given in the appendix (see Tables 5 and 6, and Figures 2 and 3). On DIP-PPI and BioGRID-PPI, the combined strategy in phase 2 of GFA is used. In both cases, $MinFrequency = 3$ is selected as the smallest frequency threshold instead of $MinFrequency = 2$. This is because the combined strategy in phase 2 introduces noise (*i.e.* spurious subgraphs) as it relaxes the definition of frequency. Such spurious subgraphs typically have very low frequencies and could potentially be eliminated by using a moderate *MinFrequency* threshold.

On DIP-PPI, the parameter *MaxOverlap* is set as 0.2 as before. With $MinFrequency = 20$, GFA predicts 116 protein complexes. 51 of them are conjectured to be novel based on their (zero) effective sizes, using MIP-MAN-COMP as the benchmark. The sensitivity and specificity are 50% and 91%, respectively. With $MinFrequency = 3$, GFA predicts 318 protein complexes, and 204 of them are conjectured to be novel. The sensitivity and specificity

are 73% and 85%, respectively. Unlike on the MIPS or DIP PPI networks, the parameter *MaxOverlap* has a significant impact on the prediction results for BioGRID-PPI, since the network is much denser. We will take $MaxOverlap = 0.5$ as an example to show the results in this paper. With $MinFrequency = 20$, GFA predicts 166 protein complexes and 111 of them are conjectured to be novel. The sensitivity and specificity are 31% and 83%, respectively, still using MIPS-MAN-COMP as the benchmark. With $MinFrequency = 3$, GFA predicts 870 protein complexes and 529 of them are conjectured to be novel. The sensitivity and specificity in this case are 48% and 63%, respectively.

We note that in all of the above tests, GFA achieves the best sensitivity (of 73%) with a decent specificity (of 85%) on DIP-PPI, whereas its accuracy deteriorates significantly on BioGRID-PPI. This does not surprise us because although MIPS-PPI is supposed to be the most reliable one among the three PPI networks, it may also miss many true edges (interactions). In other words, it may be too conservative. These missing edges, some of which may exist in DIP-PPI, could provide useful density information in the computation of GFA. On the other hand. BioGRID-PPI may contain many false interactions that could mislead GFA. The prediction efficiency of GFA remains good on both DIP-PPI and BioGRID-PPI.

3.3. Comparison with the previous methods

In this section, we compare the performance of GFA with those of two existing methods for identifying protein complexes from PPI networks that are proposed or surveyed in Spirin and Mirny [12] and Li *et al.* [15]. We will not consider methods based on comparative analysis of PPI networks in this comparison, since we are mostly interested in the interplays between PPI data and microarray gene expression data in the current study and the issue of how gene expression profiles could help analysis of PPI networks. Because the previous methods all predict complexes that are connected in the input PPI network and contain at least three proteins, MIPS-MAN-COMP will be used as the benchmark for a fair comparison.

Table 3. Comparison of GFA and Spirin and Mirny [12] on MIPS-PPI. The row $MinFrequency = 58$ shows the result of GFA when *MinFrequency* is set as 58.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
V. Spirin	76	39	28	46	51	21
$MinFrequency = 58$	77	39	30	35	35	40

The first comparison is with the result reported in Spirin and Mirny [12], which is a bit old but still among the most accurate protein complex predictions. After removing duplicates from the protein complexes predicted in this reference, we obtain 76 subgraphs. To match this number, we set $MinFrequency = 58$ so that the number of subgraphs output by GFA is close to 76. Table 3 summarizes the performance of both prediction results. Both results identify almost the same number of complex components in the benchmark (so the same sensitivity) with the cutoff $p = 0.2$ or $p = 0.5$. But the values of the parameter M show that our result is more efficient, since the 35 matched predicted complexes in our result achieve the same sensitivity as that achieved by the 46 matched predicted complexes in Spirin and Mirny [12]. More importantly, because of this efficiency, our result suggests 19 more novel complexes that are completely disjoint from the proteins in the benchmark complexes, as shown in the $P_{e=0}$ column.

The comparison should be taken with a grain of salt because Spirin and Mirny [12] used an older PPI data from MIPS, which is no longer available. Note that, a more recent MIPS-PPI may not give their method a better result. Nonetheless, our algorithm GFA is much simpler than theirs, especially when the simple strategy is used in phase 2, because their result is a combination of the outputs of three totally different algorithms.

The second comparison is with DECAFF, an algorithm proposed by Li *et al.* [15] recently. Since they gave a detailed comparison between DECAFF and many existing methods for protein complex identification in the literature, including MCODE [11], LCMA [14], and an algorithm proposed by Altaf-Ul-Amin *et al.* [28], and demonstrated the superiority of DECAFF over these methods, we will only compare GFA with DECAFF in this paper.

DECAFF uses the same MIPS PPI data as that

used by GFA and predicts 1220 complexes. The first group of results in Table 4 shows the matching of the 1,220 complexes to the benchmark complexes. For comparison, the matching of the 287 complexes predicted by GFA with $MinFrequency = 2$ is listed here too. As we can see, the GFA prediction result contains less than 1/4 of the complexes predicted by DECAFF while only losing 3% sensitivity. This comparison also suggests that the complexes produced by DECAFF overlap with each other a lot. For a more informative comparison, we remove overlapped putative complexes as described in Section 2.4.1. Since the removal depends on the cutoff *MaxOverlap*, we consider two cutoff values here: 0.5 and 0.2.

The second and third groups of results in Table 4 compare the predictions of GFA and DECAFF after the removal. In each case, the *MinFrequency* parameter in GFA is selected so that the number of predicted complexes by GFA is close to that by DECAFF. The comparison shows that GFA outperforms DECAFF in terms of sensitivity ($I/100$), specificity ($M/P_{e\geq 2}$) and efficiency (I/M). Moreover, GFA is able to find more novel protein complexes, as shown in the $P_{e=0}$ column.

Table 4. Comparison of GFA and DECAFF on MIPS-PPI. o and f stand for $MaxOverlap$ and $MinFrequency$, respectively.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
DECAFF	1,220	73	48	505	757	280
o=0.2, f=2	287	70	41	63	82	176
o=0.5, DECAFF	242	61	25	64	109	87
o=0.5, f=4	228	68	41	67	77	131
o=0.2, DECAFF	111	43	21	41	55	44
o=0.2, f=18	111	53	36	46	47	58

We also compare our results on BioGRID-PPI with that generated by DECAFF on the same PPI network as reported in Li *et al.* [15]. A comparison of the 2,840 predicted complexes predicted by DECAFF and the benchmark complexes is given in the first row of Table 7 in the appendix. Although this prediction has a high (perfect) sensitivity and decent specificity, it has a very low efficiency as the 118 complex components in the benchmark are identified by a large number (*i.e.* 1,141) of the predicted complexes. In other words, the predicted complexes overlap a lot with each other. For a more informative comparison, we again remove overlapped putative complexes using the method described in Section 2.4.1, with $MaxOverlap = 0.5$ or $MaxOverlap = 0.2$. The second and third groups of results in Table 7 compare the predictions of GFA and DECAFF after the removal. In each case, $MinFrequency$ is selected so that the number of predicted complexes by GFA is close to that by DECAFF. The table shows that GFA outperforms DECAFF significantly in terms of specificity ($M/P_{e\geq 2}$), efficiency (I/M), and the ability to predict novel protein complexes ($P_{e=0}$). It is only outperformed by DECAFF in sensitivity when $p = 0.2$. In fact, it achieves a better sensitivity than DECAFF when $p = 0.5$, although the sensitivities of both methods are all pretty low.

3.4. The effects of microarray data and parameters in phase 1

The experiments on the three PPI datasets show that the number of samples combined in GFA has a big impact on the final result, but the prediction results of GFA are not very sensitive to the parameters in phase 1. Due to the page limit, a detailed discussion on these effects or non-effects is omitted in this extended abstract but will be given in the full paper.

4. CONCLUSIONS AND DISCUSSION

We have presented a max-flow based algorithm, GFA, to identify complexes from PPI networks by incorporating microarray data. Compared to the previous methods, GFA is actually able to find the densest subgraphs in the input PPI network efficiently, rather than using some local search heuristic. Our experiments on the MIPS, DIP, and BioGRID PPI networks have demonstrated that GFA outperforms the previous methods in terms of specificity, efficiency and ability in predicting novel protein complexes, and it has a comparable sensitivity as those of the previous methods. One of the reasons that GFA was not able to identify some of the benchmark protein complexes is that it removes nodes of degree 1 from the network in every iteration. This step is necessary since it prevents GFA from producing many small spurious complexes. We may have to explore a different strategy in order to improve the

sensitivity.

In phase 1 of GFA, multiple rounds of DSA have to be executed in order to find a dense subgraph of a sufficiently small size. This is time consuming. To speed up this step, we can set a small *MaxIter*. We have demonstrated that the final result is not very sensitive to this parameter. An alternative is to assign larger weights to nodes based on expression data in each round.

Our discussion in the previous section shows that the performance of GFA generally improves when more samples are combined. However, the running time of GFA is proportional to the number of samples and could become a concern when the PPI network is large/dense.

Acknowledgements

This work was partly supported by the Natural Science Foundation of China grants 60621062, 60503001, 60528001, and 60575014, the Hi-Tech Research and Development Program of China (863 project) grants 2006AA01Z102 and 2006AA02Z325, the National Basic Research Program of China grant 2004CB518605, NSF grant IIS-0711129, NIH grant LM008991, a startup supporting plan at Tsinghua University, and a Changjiang Visiting Professorship at Tsinghua University.

References

1. Peter Uetz et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* 2000; 403:623–627.
2. Takashi Ito et al. Toward a protein-protein interaction map of the budding yeast: a. comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* 2000; 97(3):1143–1147.
3. Takashi Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* 2001; 98(8):4569–4574.
4. Yuen Ho et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature* 2002; 415:180–183.
5. Anne-Claude Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002; 415:141–147.
6. Lukasz Salwinski et al. The database of interacting proteins: 2004 update. *Nucleic Acids Research* 2004; 32:D449–D451.
7. U. Güldener et al. CYGD: the comprehensive yeast genome database. *Nucleic Acids Research* 2005; 33:D364–D368.
8. Chris Stark et al. BioGRID: a. general repository for interaction datasets. *Nucleic Acids Research* 2006; 34:D535–D539.
9. Ulrich Stelzl et al. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* 2005; 122(6):957–968.
10. Tanya Barrett et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research* 2007; 35:D760–D765.
11. Gary D. Bader and Christopher W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003; 4(2).
12. Victor Spirin and Leonid A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* 2003; 100(21):12123–12128.
13. Peng Jun Pei and Ai Dong Zhang. A 'seed-refine' algorithm for detecting protein complexes from protein interaction data. *IEEE Transactions on Nanobioscience* 2007; 6(1):43–50.
14. Xiao-Li Li et al. Interaction graph mining for protein complexes using local clique merging. *Genome Informatics* 2005; 16(2):260–269.
15. Xiao-Li Li et al. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. *Comput. Syst. Bioinformatics Conf.* 2007; 6:157–168.
16. Amy Hin Yan Tong et al. A combined experimental and computational strategy to dene protein interaction networks for peptide recognition modules. *Science* 2002; 295(5553):321–324.
17. Bill Andreopoulos et al. Clustering by common friends finds locally significant proteins mediating modules. *Bioinformatics* 2007; 23(9):1124–1131.
18. A. D. King et al. Protein complex prediction via cost-based clustering. *Bioinformatics* 2004; 20(17):3013–3020.
19. Dongbo Bu et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research* 2003; 31(9):2443–2450.
20. Sabine Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research* 2003; 31(21):6283–6289.
21. Yu Huang et al. Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics* 2007; 23(13):i222–i229.
22. Trey Ideker et al. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002; 18(90001):S233–S240.
23. Zheng Guo et al. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinfor-*

matics 2007; 23(16):2121–2128.
24. Roded Sharan et al. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol* 2005; 12(6):835–846.
25. Eitan Hirsh and Roded Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics* 2007; 23(2):e170–e176.
26. Giorgio Gallo et al. A fast parametric maximum flow algorithm and applications. *SIAM J. Comput* 1989; 18(1):30–55.
27. Michael R. Garey and David S. Johnson. *Computers and intractability : a guide to the theory of NP-completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.
28. M. Altaf-Ul-Amin et al. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 2006; 7(207).

Appendix: additional figures and tables

Table 5. Protein complexes predicted by GFA on DIP-PPI and their matches in MIPS-MAN-COMP and MIPS-ALL-COMP. MAN and ALL stands for MIPS-MAN–COMP and MIPS-ALL-COMP, respectively. f stands for $MinFrequency$, and $MaxOverlap$ is set to 0.2.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
MAN, $f = 20$	116	57	35	49	54	51
MAN, $f = 3$	318	83	46	69	81	204
ALL, $f = 20$	116	171	75	77	97	6
ALL, $f = 3$	318	303	106	160	241	35

Table 6. Protein complexes predicted by GFA on BioGRID-PPI and their matches in MIPS-MAN-COMP. f and o stand for $MinFrequency$ and $MaxOverlap$, respectively.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
$o = 0.5, f = 20$	166	42	28	38	46	111
$o = 0.5, f = 3$	870	85	41	108	223	529
$o = 0.2, f = 20$	157	38	25	35	44	106
$o = 0.2, f = 3$	453	73	30	69	103	296

Table 7. Comparison of GFA and DECAFF on BioGRID-PPI. Again, o and f stand for $MaxOverlap$ and $MinFrequency$, respectively.

	P	$I(0.2)$	$I(0.5)$	M	$P_{e\geq 2}$	$P_{e=0}$
DECAFF	2,840	118	81	1141	1,871	533
$o = 0.5$, DECAFF	610	101	30	144	264	215
$o = 0.5$, $f = 4$	582	75	40	79	144	369
$o = 0.2$, DECAFF	226	53	15	48	78	113
$o = 0.2$, $f = 10$	221	51	25	46	56	150
$o = 0.2$, $f = 9$	234	52	25	46	58	160

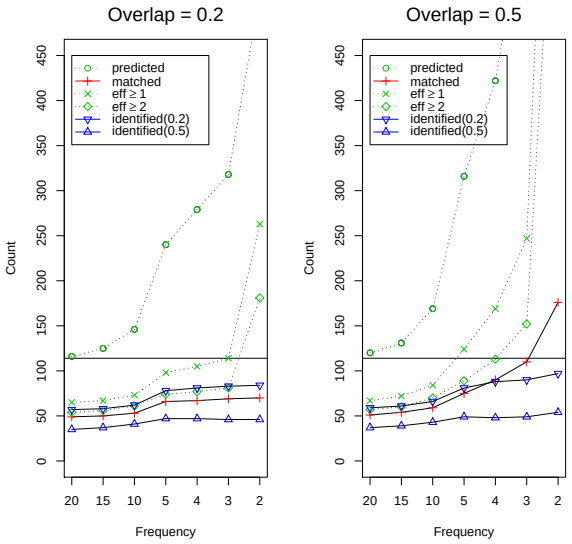

Fig. 2. Protein complexes predicted by GFA on DIP-PPI and their matches in MIPS-MAN-COMP. Two $MaxOverlap$ values, 0.2 (left) and 0.5 (right), are considered. Here, $eff \geq 1$ stands for $P_{e\geq 1}$.

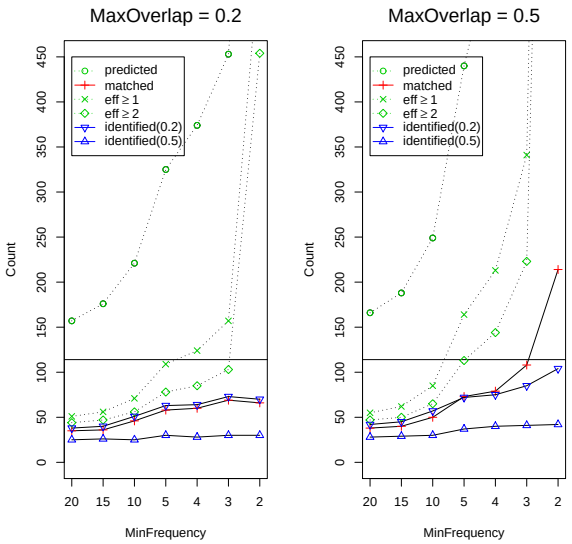

Fig. 3. Protein complexes predicted by GFA on BioGRID and their matches in MIPS-MAN-COMP. Two $MaxOverlap$ values are considered: 0.2 (left) and 0.5 (right). Again, $eff \geq 1$ stands for $P_{e\geq 1}$.