BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3013–3020
doi:10.1093/bioinformatics/bth351

Protein complex prediction via cost-based clustering

A. D. King[1], N. Pržulj[1] and I. Jurisica[1,2,]*

[1]Department of Computer Science, University of Toronto, Toronto, M5S 3G4, Canada and [2]Ontario Cancer Institute, Division of Cancer Informatics, Toronto, M5G 2M9, Canada

Received on April 6, 2004; revised May 5, 2004; accepted May 14, 2004
Advance Access publication June 4, 2004

ABSTRACT

Motivation: Understanding principles of cellular organization and function can be enhanced if we detect known and predict still undiscovered protein complexes within the cell's protein–protein interaction (PPI) network. Such predictions may be used as an inexpensive tool to direct biological experiments. The increasing amount of available PPI data necessitates an accurate and scalable approach to protein complex identification.

Results: We have developed the Restricted Neighborhood Search Clustering Algorithm (RNSC) to efficiently partition networks into clusters using a cost function. We applied this cost-based clustering algorithm to PPI networks of *Saccharomyces cerevisiae, Drosophila melanogaster* and *Caenorhabditis elegans* to identify and predict protein complexes. We have determined functional and graph-theoretic properties of true protein complexes from the MIPS database. Based on these properties, we defined filters to distinguish between identified network clusters and true protein complexes.

Conclusions: Our application of the cost-based clustering algorithm provides an accurate and scalable method of detecting and predicting protein complexes within a PPI network.

Availability: The RNSC algorithm and data processing code are available upon request from the authors.

Contact: ij@uhnres.utoronto.ca

Supplementary Information: Supplementary data are available at http://www.cs.utoronto.ca/~juris/data/ppi04/

1 INTRODUCTION

Recent developments in the rapidly expanding fields of network biology and cell biology have resulted in a deluge of protein–protein interaction (PPI) data with accompanying data on protein complexes emerging from these PPI networks (Uetz *et al*., 2000; Ito *et al*., 2000, 2001; Giot *et al*., 2003; Li *et al*., 2004; Gavin *et al*., 2003; Ho *et al*., 2003). An inevitable consequence of this wealth of data is the need for efficient and accurate automated tools to identify and quantify significant portions of these data. Our method relies on modeling the PPI network with a graph, where nodes represent proteins and edges correspond to interactions, and applying principles of both graph theory and gene ontology to identify likely protein complexes with scalable accuracy.

Modeling PPI networks with simple graphs has been used for many applications, one of which is the prediction of protein complexes within the PPI networks (Bader and Hogue, 2003; Pržulj *et al*., 2004). Protein complexes generally correspond to dense subgraphs in the PPI network; thus, proteins in a given complex are highly interactive with each other (Bader and Hogue, 2003; Pržulj *et al*., 2004). Previous approaches to graph-theoretic cluster prediction include simple clustering methods such as identification of k-cores (Bader and Hogue, 2003) super-paramagnetic clustering (Spirin and Mirny, 2003) and the highly connected subgraph approach (Hartuv and Shamir, 2000; Pržulj *et al*., 2004). Although in this paper we focus on graph clustering only, many other important application of graph theory to cellular biology exist (e.g. Barabási and Oltvai, 2004; Newman, 2003; Albert and Barabási, 2002; Strogatz, 2001; Pržulj, 2004). The last one focuses specifically on PPI networks.

We have developed and applied the Restricted Neighborhood Search Clustering algorithm (RNSC), which partitions the network's node set into clusters based on a cost function that is assigned to each partitioning. We then filtered the RNSC output so that only clusters that share characteristics of known protein complexes are considered. This method was applied to four *Saccharomyces cerevisiae* PPI networks discussed in Pržulj *et al*. (2004), two *Drosophila melanogaster* PPI networks (Giot *et al*., 2003) and a *Caenorhabditis elegans* PPI network (Li *et al*., 2004). Our criteria for filtering the clusters included: cluster size, cluster density and functional homogeneity, all of which are discussed later in this paper. We compared the results of our method with known yeast protein complexes (Mewes *et al*., 2002) and found that optimizing filter cutoff values leads to high matching rates and large cluster sample sizes.

*To whom correspondence should be addressed.

2 SYSTEMS AND METHODS

Our protein complex prediction method relies on modeling PPI data as graphs (or networks). A graph $G = (V, E)$ is a set V of nodes, representing proteins, and a set E of edges, representing interactions between pairs of proteins. Each edge joins two nodes. We also use $G(V)$ to denote the set of nodes V of G (West, 2001).

We used four *S.cerevisiae* PPI networks originating from von Mering *et al*. (2002) comprising 2455, 11 000, 45 000 and 78 390 interactions. We call these networks Y_{2k}, Y_{11k}, Y_{45k} and Y_{78k} respectively, the smallest one containing high confidence interactions only, and the larger ones having an increasing number of lower confidence interactions. We used two *D.melanogaster* PPI networks, one derived from the entire fruitfly network of interactions given in Giot *et al*. (2003), and one derived from those interactions with confidence >0.5; these have 20 007 and 4637 interactions respectively, and we call these networks F_{20k} and F_{4k}. We also used a *C.elegans* PPI network, W_{5k}, consisting of 5222 interactions (Li *et al*., 2004) (also see supplementary information; www.cs.utoronto.ca/∼juris/data/ppi04/).

We have analyzed these networks using a two-step process. First, we clustered them using the RNSC algorithm. Second, we filtered the results based on cluster size, density and functional homogeneity. This approach preserves only those clusters that exhibit properties more frequently observed in true biological complexes.

To evaluate the effectiveness of our algorithm for detecting protein complexes, we compared the filtered clusters of the yeast PPI networks with known protein complexes in the MIPS yeast complex database (Mewes *et al*., 2002). Whether or not a given cluster is deemed to match a given MIPS complex depends on the matching criteria detailed below.

2.1 Clustering

The bulk of the computation time was spent clustering the PPI networks using RNSC algorithm, which is described in Section 2.6. The Results included very small clusters and clusters which were either insufficiently dense, or whose component proteins had only a weak functional homogeneity. To achieve a high prediction rate, we discarded these clusters. The appearance of these clusters is not a problem with the algorithm, where each protein must be assigned to a cluster (see Section 2.6); rather, it is a result of partitioning sparse networks.

2.2 Cluster size

The motivation to discard small clusters comes from two ideas. First, any overlap proportion between a small predicted complex and a known complex is more likely to be by chance than the same overlap proportion involving a larger predicted complex. Second, small known complexes frequently have low density in current PPI networks and are therefore difficult to detect using a clustering algorithm. We experimentally determined a lower bound for a cluster size and discarded all predicted complexes with size below this lower bound. The size bound is dependent on the PPI network in question.

2.3 Cluster density

Protein complexes usually exhibit high interaction rates with each other. Therefore, lower-density clusters are less likely corresponding to known protein complexes. By discarding clusters whose densities lie below a certain threshold, we can increase the prediction rate of our algorithm.

2.4 Functional homogeneity

Known protein complexes often exhibit high functional homogeneity (Bu *et al*., 2003; Pržulj *et al*., 2004), i.e. a large proportion of proteins within a known complex likely belongs to the same functional group. This property also holds for dense regions of PPI networks (Bu *et al*., 2003; Pržulj *et al*., 2004). The functional homogeneity P-value is the probability that a given set of proteins is enriched by a given functional group merely by chance, following the hypergeometric distribution. The P-value for a cluster C and a functional group F is:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|C|-i}}{\binom{|V|}{|C|}}, \qquad (1)$$

where C contains k proteins in F, and the entire PPI network contains $|V|$ proteins (also used in Bu *et al*., 2003; Pržulj *et al*., 2004). We consider the P-value of a cluster to be its smallest P-value over all functional groups. Functional group data are derived from von Mering *et al*. (2002) for the yeast networks.

We discarded all clusters with P-value above a given, experimentally derived threshold (see Section 3.1). Although our model of functional homogeneity is very simple, using it to evaluate PPI network clusters as potential protein complexes is effective, since known protein complexes have low P-values. Sensible cutoffs for the cluster P-values range from 10^{-2} to 10^{-8} for the networks. For our matching data, we chose a cutoff of 10^{-3} for each network, because it offers a compromise between complex-cluster matching rate and a cluster passing rate, i.e. we can get a large sample of clusters with high matching rates (see Section 3.2).

2.5 Matching criteria

We need to develop matching criteria to decide whether a given PPI network cluster matches a known biological complex. From the standpoint of considering the practicality of our results, it makes sense to consider a predicted cluster and a known protein complex to be matched if a large proportion of each protein (node) set overlaps, or if the set of cluster nodes is nearly entirely contained within a set of proteins in a complex. Having a large cluster containing a small complex is not as biologically revealing, and thus we do not consider this case.

For a very large protein complex and a matching PPI network cluster, a given overlap proportion is more significant than it would be in a small complex and a matching cluster. For example, an overlap of five proteins between a complex and a cluster each of size six is less significant (i.e. more likely to occur at random) than an overlap of 50 proteins between a complex and a cluster each of size 60. Bearing this in mind, we consider a cluster Cl to match a complex Co by overlap if both:

$$\frac{|V(Cl) \cup V(Co)|}{|V(Cl)|} \geq \frac{P_{\text{cluster}}}{\log_{10}[7 + |V(Cl)|]} \tag{2}$$

and

$$\frac{|V(Cl) \cup V(Co)|}{|V(Co)|} \geq \frac{P_{\text{complex}}}{\log_{10}[7 + |V(Co)|]} \tag{3}$$

are satisfied, and we consider a cluster to match a complex by containment if:

$$\frac{|V(Cl) \cup V(Co)|}{|V(Cl)|} \geq P_{\text{contain}}. \tag{4}$$

For these three equations, P_{cluster}, P_{complex} and P_{contain} contain are all user-defined, experimentally derived proportions between 0 and 1. Note that in matching analysis, we do not consider proteins in a given protein complex if they do not appear in the applicable PPI network. Adding 7 to $|V(Cl)|$ in Equations (2) and (3) is done because it was empirically found to yield good thresholds. In fact, Equations (2)–(4) are entirely the result of empirical optimization: the equations generate sensible values such that a match is an overlap that represents high statistical significance without being too stringent a requirement (Fig. 1 and Section 3.2).

2.6 The RNSC algorithm

A clustering of a network $G(V, E)$ is a decomposition of the set of nodes V into subsets of nodes that are highly interconnected (i.e. these subsets of nodes induce dense subgraphs). Our clustering algorithm is the RNSC, which is a cost-based local search algorithm based loosely on the tabu search metaheuristic (Glover, 1989). In the context of this algorithm, a clustering of a network $G = (V, E)$ is equivalent to a partitioning of the node set V. The RNSC efficiently searches the space of partitions of V, each of which is assigned a cost, for a clustering with low cost. The algorithm searches using a simple integer-valued cost function (called the naive cost function) as a preprocessor before it searches using a more expressive (but less efficient) real-valued cost function (called the scaled cost function). The initial clustering is random or user-input.

The RNSC searches for a low-cost clustering by first composing an initial random clustering, then iteratively moving one node from one cluster to another in a randomized fashion to improve the clustering's cost. A general move is one that reduces the clustering cost by a near-optimal amount.

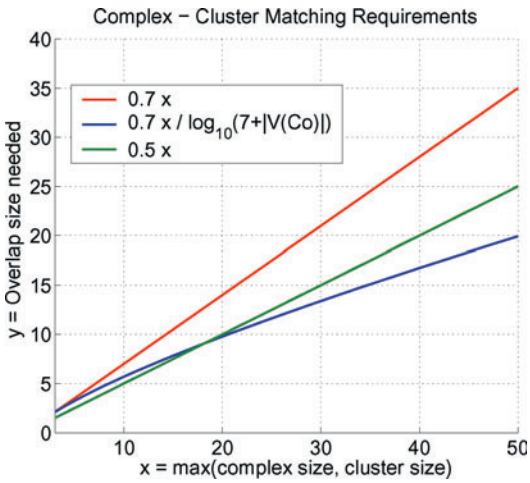

Fig. 1. The overlap requirements for a match between a cluster and a complex. The x-axis is the larger of the complex size and the cluster size, and the y-axis is the overlap size needed to consider the complex and the cluster to be matched. The lines $y = 0.5x$ and $y = 0.7x$ are given for reference only. This figure can be viewed in colour on *Bioinformatics* online.

The common problem among local search algorithms is their tendency to settle in poor local minima. This problem can be largely avoided by using diversification and multiple experiments. Thus, our algorithm makes diversification moves, which shuffle the clustering by occasionally dispersing the contents of a cluster at random. In addition, the RNSC maintains a list of tabu (forbidden) moves to prevent cycling back to the previously explored partitioning. Since the RNSC is randomized, different runs on the same input data will result in different clusterings.

The algorithm maintains many data structures and incurs a large memory cost for the sake of time-efficiency. Ordinarily, maintenance of the data structures for such a search algorithm would present a prohibitive cost in computation. However, there are many problem-specific properties related to both graph clustering and the chosen cost functions that allow the RNSC to perform very efficiently (a more detailed explanation of the RNSC algorithm can be found in the supplementary information (King, 2004). To achieve high accuracy in predicting true protein complexes, the RNSC output is filtered, using the following criteria: setting a maximum P-value for functional homogeneity, a minimum density and a minimum size. Only clusters that satisfy these criteria are presented as predicted protein complexes.

3 EXPERIMENTS AND RESULTS

Each network was clustered at least four times using the RNSC algorithm running under Linux. Each run took between 4 s and 67 min on a 2.8 GHz processor, with Y_{2k} being the fastest and F_{20k} being the most time consuming. We considered the lowest-cost clustering produced by these runs for each

Table 1. Cluster size lower bounds for *S.cerevisiae* PPI networks' clusters, needed to pass through the filter

Network	Minimum size	Total clusters	Passing clusters
Y_{2k}	4	393	48
Y_{11k}	5	974	84
Y_{45k}	7	181	86
Y_{78k}	8	1811	90

For example, for Y_{2k} network, out of 393 clusters in total, 48 were of size at least 4.

network. Resulting clusters are available in the Supplementary information.

The thresholds for cluster size, density and functional homogeneity are a matter of compromise; although increasing the strictness of the thresholds generally increases the prediction rate, it reduces the number of passing predictions (see Section 3.1). In the case where few protein complexes are known for the PPI network (e.g. fruitfly and worm), this scalability is extremely useful, since we can make the thresholds strict at the beginning, and relax them as we analyze the growing set of predicted protein complexes (clusters). We have chosen the following matching thresholds: $P_{\text{cluster}} = P_{\text{complex}} = 0.7$ and $P_{\text{cluster}} = 0.9$.

3.1 Filter cutoffs

All of the three filter cutoffs (for cluster size, density and functional homogeneity) were chosen to yield reasonably high sample sizes while ensuring that clusters passing through the filter had a good chance of matching known complexes. In the case of the yeast networks, the minimum cluster size cutoff increased with the size of the network accordingly. Table 1 shows the chosen size cutoffs for the yeast PPI networks, along with the sizes of the cluster sets that pass the size cutoff.

We imposed a lower bound on the density of predicted complexes. As seen in Figure 2, a significant decrease in the passing rate of the RNSC clusters occurs when the cluster density cutoff is between 0.65 and 0.75. In general, known complexes tend to have high density in the PPI network, but very few large complexes have density 1 (See Supplementary Information). A density cutoff in the range of 0.65 and 0.75 allows a good compromise between passing sample size and prediction rate, but a cutoff closer to 0.9 may give a very high passing rate in a small sample size. For experimental results in the yeast networks, we used a cutoff of 0.7.

As with cluster size and density, for functional homogeneity (P-value) filtering we wish to maintain both a reasonable sample size and a high matching rate among passing clusters. Figure 3 shows the effect of changing thresholds for both density and P-value (after filtering for size) in Y_{78k}. Figure 4 shows the effect of these thresholds on the sample size in Y_{78k}.

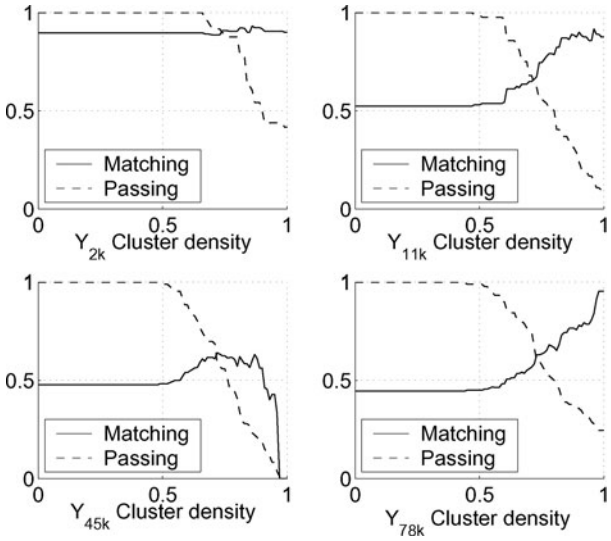

Fig. 2. The proportion of RNSC clusters which pass the cluster density filter (i.e. cluster passing rate) and the proportion of these passing clusters that match known complexes (cluster matching rate) for yeast networks Y_{2k}, Y_{11k}, Y_{45k} and Y_{78k}. These rates are for clusters that have already been filtered for size, but not for functional homogeneity.

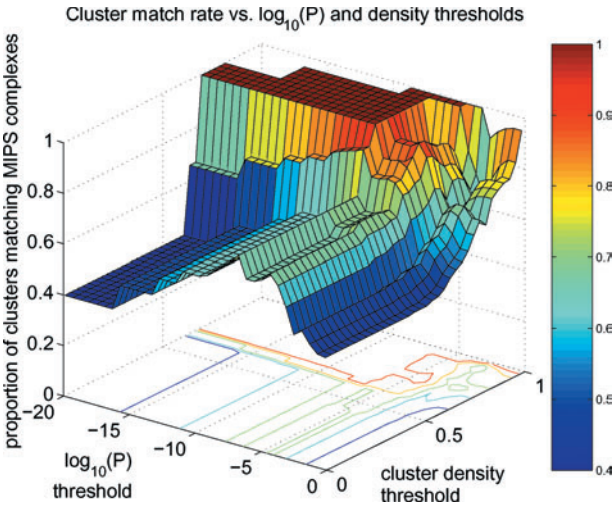

Fig. 3. Proportion of passing clusters in Y_{78k} which match a known complex from MIPS. The sample is the set of clusters passing first the size restriction, then the P-value restriction and density restriction. The P-value and density restrictions are given on the x- and y-axes. We chose 0.7 and 10^{-3} as our density and P-value cutoffs, respectively. This figure can be viewed in colour on *Bioinformatics* online.

For our experimental cluster passing rates, we chose a P-value cutoff of 10^{-3}.

3.2 Results

Matching rates for the yeast networks are shown in Table 2 for density $\geq$0.7 and $P \leq 10^{-3}$, using the size cutoffs found

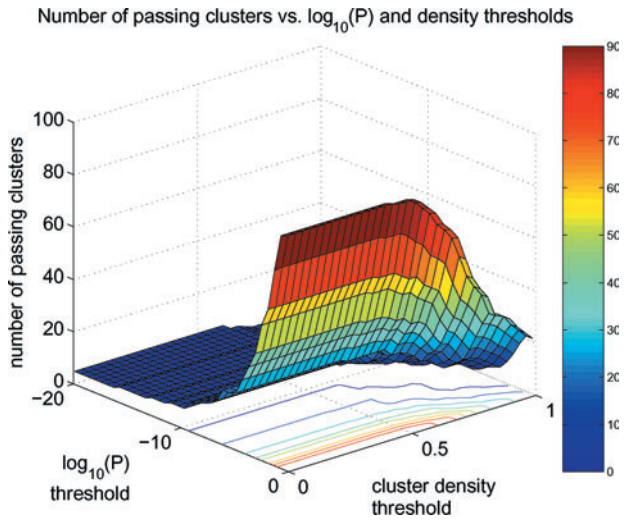

Fig. 4. The effect of changing P-value and density cutoffs on the sample size, i.e. the number of clusters that pass the filter criteria for Y_{78k}. Clusters are first filtered by size, then by P-value and density. We chose 0.7 and 10^{-3} as our density and P-value cutoffs, respectively. This figure can be viewed in colour on *Bioinformatics* online.

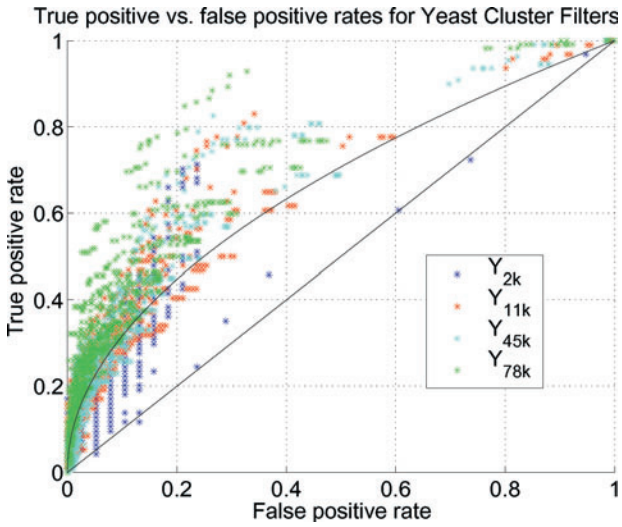

Fig. 5. True positive rate versus false positive rate for filtered yeast clusters: the proportion of matched clusters accepted by the filter versus the proportion of unmatched clusters accepted by the filter. The points represent all tested filter cutoffs for size, P-value and density. Clusters of size less than two are ignored in these data. The curve $y = \sqrt{x}$ is given as a reference. This figure can be viewed in colour on *Bioinformatics* online.

Table 2. Matching data for density $\geq$0.7 and $P \leq 10^{-3}$; using both RNSC and MCL algorithms (van Dongen, 2002); MCL values shown in parentheses

Graph	Minimum size	Passing clusters	Matched clusters	Prediction rate (%)
Y_{2k}	4	28 (8)	23 (9)	82.1 (88.9)
Y_{11k}	5	45 (8)	30 (12)	66.7 (75)
Y_{45k}	7	32 (2)	21 (4)	65.6 (50)
Y_{78k}	8	31 (3)	22 (6)	73.3 (50)

Passing clusters are those that pass all filtering criteria, and matching clusters are those passing clusters that satisfy the matching criteria with at least one complex from MIPS.

to provide good passing sample sizes (described in Table 1). The table also contains results reached by replacing the output of the RNSC algorithm with the output of the MCL (Markov Cluster) algorithm (van Dongen, 2002). The fact that all of the filter cutoffs can be adjusted means that there are countless samples of varying size and matching rate. An example is presented in Figure 5, where each choice of filter cutoffs is represented as a point. In spite of the noise, the results for Y_{78k} are the best: for a given false positive rate, the true positive rate for Y_{78k} is the highest of the four. This may be because the larger dataset carries much more statistical significance, in spite of it containing more noise.

Figure 6A shows an example of a predicted complex (i.e. a RNSC cluster) and the true protein complex from MIPS that it matches in the yeast network Y_{11k}. The RNSC cluster has size 8, density 0.964, and P-value 3.98×10^{-8}. The known cluster, COPI, has size 8, density 0.786, and P-value 3.29×10^{-10}. COPI is an intracellular transport complex that contributes to the coating of membrane vesicles within the cell. Although the ADPribosylation factor protein (Arf1), which is in COPI but not the predicted complex, has the same gene ontological function as the rest of the proteins in COPI, it is incident with only one edge in the complex. Gpt2, which is contained in the predicted complex but not the known complex, is incident with six edges in the cluster. Gpt2 is glutamic pyruvate transaminase 2 with a strong similarity to Sct1p (GAT1), and is responsible for transaminase and transferase activity and lipid biosynthesis; Gpt2 is located in endoplasmic reticulum. Although Gpt2 is assigned a different functional group by von Mering *et al*. (2002) according to MIPS, each of these nine proteins is responsible for cellular transport, transport facilitation and transport routes (Mewes *et al*., 2002; von Mering *et al*., 2002), and most are listed as probable members of membrane biogenesis and traffic complex. This suggests that Gpt2 likely belongs to the COPI complex.

Figure 6B shows an unmatched RNSC cluster in Y_{11k}. However, it exhibits all of the properties that we are looking for: it has sufficient size, 7, and high density, 0.810; its functional homogeneity P-value is 9.31×10^{-6}, with six of its seven proteins contributing to metabolism. Moreover, it comprises five members involved in pyrimidine base biosynthesis (Ura1, 2, 3, 5, 10). This suggests that biological validation of this set of proteins forming a protein complex may be worthwhile.

Figure 6C shows an example of a containment match: an RNSC cluster in Y_{11k} is contained within a MIPS complex (note that the cluster contains most of the edges within the

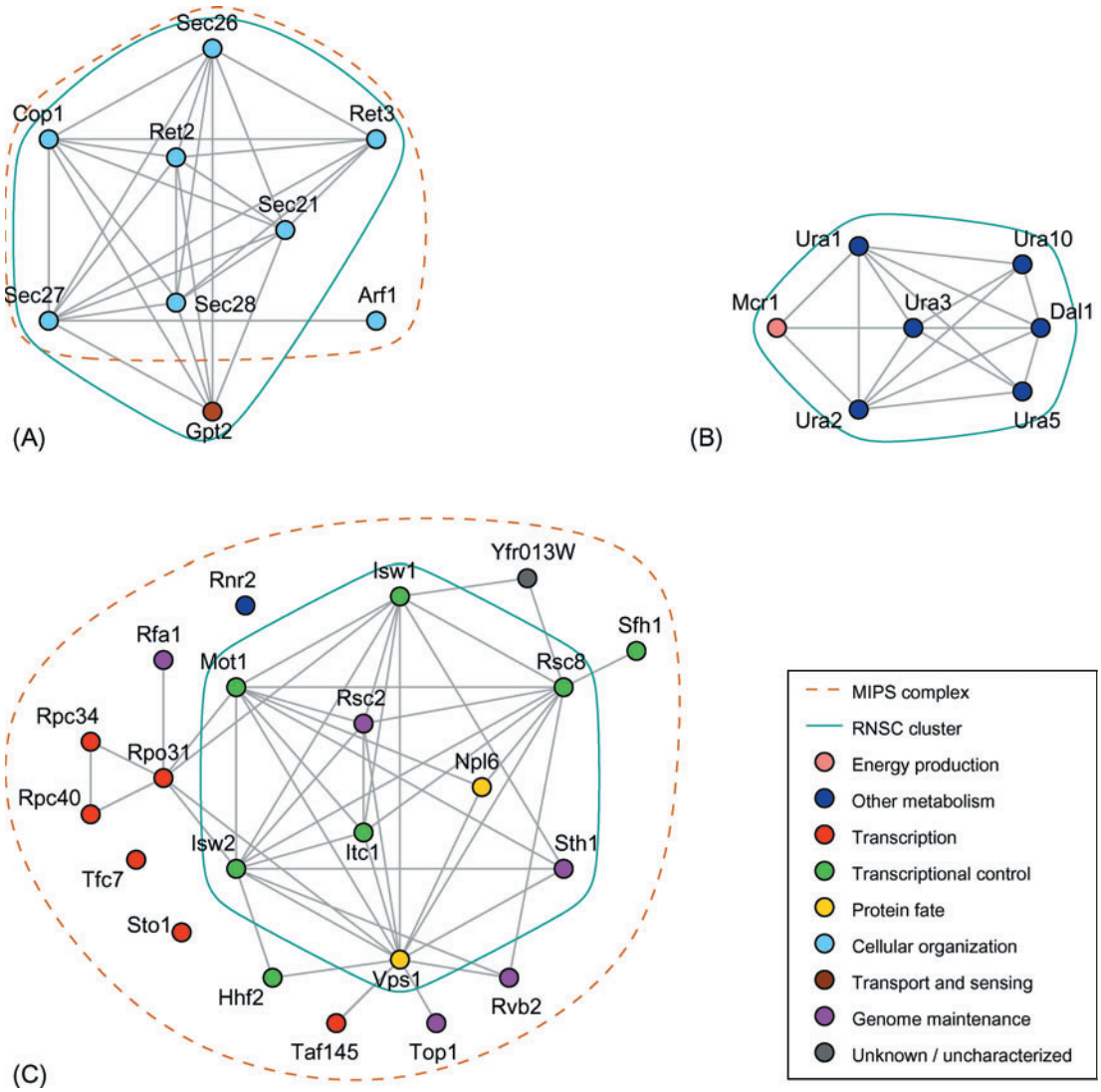

Fig. 6. Examples of matched and predicted protein complexes: (**A**) MIPS complex COPI in the yeast network Y_{11k} and the matching complex predicted by RNSC. Each has size 8, and their overlap is 7. (**B**) An unmatched cluster in the yeast network Y_{11k}. The cluster has no overlap greater than one protein with any known complex. It passed through the filter, and exhibits characteristics of a protein complex. (**C**) This RNSC cluster in Y_{11k} is contained within a larger MIPS complex. Note that the cluster contains most of the edges in the complex.

complex). Indeed, the nodes of the complex that are not included in the cluster do not exhibit the ideal graph-theoretic properties of protein complexes. They are sparsely connected and largely heterogeneous. This MIPS complex is responsible for transcription and transcriptional control, genome maintenance and chromatin structure remodeling. Clearly, decreasing the number of false negatives in current PPI databases should lead to better overlap between predicted and true protein complexes. Text analysis and manual curation of PubMed resources will help substantially (Shatkay and Feldman, 2003; Xenarios *et al*., 2000; Zanzoni *et al*., 2002; Peri *et al*., 2003).

Similarly as in Pržulj *et al*. (2004), RNSC identified Rib1-5, Rib7 as a functionally homogeneous cluster (riboflavin biosynthesis) with density 1.0 in Y_{11k}. In Y_{2k} RNSC identified only cluster comprising Rib1, Rib3 and Rib5 with a density 0.67. In Y_{45k} and Y_{78k}, Rib1–5 and Rib7 have density 1.0 among themselves, but the proteins are highly interactive with other proteins. Although SGD lists all six proteins directly annotated to the vitamin B2 biosynthesis (Cherry *et al*., 1998; Christie *et al*., 2004), Rib1–5, Rib7 do not form a cluster in either of these two PPI networks; rather, the Rib proteins are divided among several clusters. This is a case in which hierarchical cluster analysis may provide some insight; i.e., considering all four networks for yeast simultaneously.

The results for the fly and worm networks are less definitive. As there are no comprehensive sources for complexes and

functional classifications for these networks, we could neither construct P-values for the clusters nor compare them to a set of known complexes. In these networks, we filtered clusters for size and density. The predicted complexes are given in the Supplementary information. For F_{20k}, there are only five predicted complexes, the largest of which has size 5. This is due to the fact that the current fruitfly network is extremely sparse. For F_{5k}, the less noisy dataset for fruit-fly, there are 42 predicted complexes, all of size 3 and 4. For W_{5k}, there are 32 predicted complexes, including 3 of size $\geq$10. In the future, more complete PPI data will likely lead to a larger, more significant set of predicted complexes for fly and worm.

4 DISCUSSION

Our results suggest that true protein complexes exhibit certain graph-theoretic properties and functional homogeneity. Thus, using size, density and functional homogeneity as filtering criteria for network clusters is a reasonable approach to predict novel protein complexes. However, there are some problems with this approach. While protein complexes are usually expected to have high density in PPI networks, not all do. A related problem is the incompleteness of current PPI networks. The more complete and accurate our PPI and known protein complexes datasets are, the more accurately we can analyze the PPI networks. Further, the functional homogeneity, while accurate for the most part, seems to be an incomplete, oversimplified model. Many known complexes show low functional homogeneity. Also, many proteins belong to multiple functional groups. In addition, many proteins are of unknown function.

Even with such a simple filtering model and incomplete data, we managed to achieve very high matching rates between PPI network clusters and known protein complexes (Table 2). In comparison, Bader and Hogue generated a set of 209 predicted complexes, of which 54 match the MIPS database in at least 20% of their proteins in a yeast PPI network of some 15 000 interactions (Bader and Hogue, 2003). In Pržulj *et al*. (2004), a set of 31 predicted complexes is given for Y_{11k}, of which 27 were reported to have high overlaps with MIPS complexes. Jansen *et al*. (2002) predicted that pairs of nodes to be in the same cluster; they, like us, achieved low error rates (as low as 0% for five predicted pairs) that increased with the sample size. However, their findings cannot easily be applied to predicting entire complexes, but only interactions within them (Jansen *et al*., 2002). Our results complement these efforts to better understand protein complexes within networks or protein–protein interactions.

Many clustering algorithms are available. Most, including the single-linkage and UGPMA algorithms, are applicable to points in a geometric space rather than networks such as PPI networks. As shown in Table 2, using the MCL algorithm rather than the RNSC algorithm to cluster the graph results in less significant datasets. Further details of the comparison of the two algorithms is provided in King (2004).

5 CONCLUSIONS AND FUTURE WORK

Using the RNSC algorithm to cluster PPI networks and filtering based on graph-theoretic resemblance to typical known protein complexes provides an effective method for predicting protein complexes. There is mounting evidence that employing graph-theoretic techniques can be useful in protein network analysis, as also demonstrated by recent research (Lappe and Holm, 2004; Pržulj *et al*., 2004; Pržulj, 2004; Yu *et al*., 2004). Our results suggest that we can predict protein complexes with high confidence using RNSC algorithm with filtering.

These predictions can be used to make biological experiments more focused, efficient and less expensive. Not only do the results warrant investigation where predicted complexes are unknown, but in some cases they warrant re-examination of current results. In order for this predictor (and other graph-theoretic tools) to work best, our knowledge of the networks needs to be improved. As more PPI data become available, automated tools for their analysis will need to become scalable and accurate.

There is a huge amount of further research to be done in the area of PPI network analysis. On the side of gene ontology, it will be helpful to investigate improved functional homogeneity models. Clearly, the mono-functional model of functional homogeneity that we use stands to be improved, most likely at a cost of simplicity. Just as protein function can be used to help predict protein complexes, knowledge of complexes can be used to predict previously unknown cellular function (Bu *et al*., 2003).

Clustering could likely be improved by applying hierarchical complex predictions. For example, in yeast we predicted protein complexes using four PPI networks of increasing size. Determining how the predicted complexes in one such network relate to those in another network will hopefully give us further insight to the nature of protein interactions.

We have developed an accurate and scalable method of predicting protein complexes from PPI data. Biological research will inevitably continue to be hypothesis driven, but computational analysis methods, such as ours, are likely to become indispensable for their ability to systematically identify areas of significance, and at a much lower cost.

ACKNOWLEDGEMENTS

A.K. would like to thank Rudi Mathon for his guidance and support. This work was supported by the University of Toronto (A.K.), OGS (N.P.), NIH P50 GM-62413 (N.P., I.J.), the National Science and Engineering Research Council RGPIN 203833–02 (I.J.).

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

Albert,R. and Barabási,A.-L. (2002) Statistical mechanics of complex networks. *Rev. Modern Phys.*, **74**, 47–97.

Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2. PMID: 12525261.

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

Bu,D., Zhao,Y., Cai,L., Xue,H., Zhu,X., Lu,H., Zhang,J., Sun,S., Ling,L., Zhang,N., Li,G. and Chen,R. (2003) Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.*, **31**, 2443–2450.

Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M. and Bauer,A. (2003) A functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al*. (2003) A protein interaction map of *Drosophila melanogaster. Science*, **302**, 1727–1736.

Glover,F. (1989) Tabu search, part I. *ORSA J. Comput.*, **1**, 190–206. ['ORSA' is called Informs today.].

Hartuv,E. and Shamir,R. (2000) A clustering algorithm based on graph connectivity. *Inform. Process. Lett.*, **76**, 175–181.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2003) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) 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*, **97**, 1143–1147.

Jansen,R., Lan,N., Qian,J. and Gerstein,M. (2002) Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genomics*, **2**, 71–81.

King,A., Pržulj,N. and Jurisica,I. (2004) *Protein complex prediction via cost-based clustering*. (Supplementary information) Bioinformatics http://www.cs.utoronto.ca/~juris/data/pp104/

King,A.D. (2004) Graph clustering with restricted neighbourhood search. Master's Thesis, University of Toronto, Toronto, Ontorio.

Lappe,M. and Holm,L. (2004) Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.*, **22**, 98–103.

Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.-O., Han,J.-D.J., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C.elegans. Science*, **303**, 540–543.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Newman,M.E.J. (2003) The structure and function of complex networks. *SIAM Rev.*, **45**, 167–256.

Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M., *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.

Pržulj,N. (2004) Graph theory approaches to protein interaction data analysis. In Jurisica,I. and Wigle,D. (eds), *Knowledge Discovery in proteomics. Throughput Biological Domains*. CRC Press. (in press).

Pržulj,N., Wigle,D. and Jurisica,I. (2004) Functional topology in a network of protein interactions. *Bioinformatics*, **20**, 340–348.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–55.

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

Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae. Nature*, **403**, 623–627.

van Dongen,S.M. (2002). Graph clustering by flow simulation. Ph.D. Thesis, University of Utrecht, Utrecht, The Netherlands

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

West,D.B. (2001) *Introduction to Graph Theory*, 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.

Yu,H., Zhu,X., Greenbaum,D., Karro,J. and Gerstein,M. (2004) TOPNET: a tool for comparing biological subnetworks, correlating protein properties with topological statistics. *Nucleic Acids Res.*, **32**, 328–337.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) Mint: a molecular interaction database. *FEBS Lett.*, **513**, 135–140.