EFFICIENT AND ROBUST PREDICTION ALGORITHMS FOR PROTEIN COMPLEXES USING GOMORY-HU TREES

A. MITROFANOVA*, M. FARACH-COLTON**, AND B. MISHRA*

**New York University, Department of Computer Science, New York, NY 10003*
***Rutgers University, Department of Computer Science, Piscataway, NJ 08855*
E-mail: antonina@cs.nyu.edu, farach@cs.rutgers.edu, mishra@nyu.edu

Two-Hybrid (Y2H) Protein-Protein interaction (PPI) data suffer from high False Positive and False Negative rates, thus making searching for protein complexes in PPI networks a challenge. To overcome these limitations, we propose an efficient approach which measures connectivity between proteins not by edges, but by edge-disjoint paths. We model the number of edge-disjoint paths as a network flow and efficiently represent it in a Gomory-Hu tree. By manipulating the tree, we are able to isolate groups of nodes sharing more edge-disjoint paths with each other than with the rest of the network, which are our putative protein complexes. We examine the performance of our algorithm with Variation of Information and Separation measures and show that it belongs to a group of techniques which are robust against increased false positive and false negative rates. We apply our approach to yeast , mouse, worm, and human Y2H PPI networks, where it shows promising results. On yeast network, we identify 38 statistically significant protein clusters, 20 of which correspond to protein complexes and 16 to functional modules.

1. Introduction

We wish to propose a new efficient and robust algorithm to infer protein complexes correctly from Y2H experiments. If the protein-protein interaction data were flaw-less and error free, then a fairly direct graph-theoretic algorithm working on graphs whose edges represent pair-wise interactions would have sufficed. The intuitively direct algorithms (e.g., clique detection, clustering or density-based methods) tend to be efficient, and work reasonably well with small number of errors that mislabel the edges falsely (both false positive and negative, errors). Our challenge is to devise more sophisticated algorithms that enjoy a comparable computational efficiency, and yet work robustly as the quality of the experimental data degrade substantially, as is common with practically all currently available PPI data. The fundamental conceptual innovation in our algorithm is to analyze structure of the graphs through their collections of edge-disjoint paths that remain relatively immune to the corrupting noises in the experiment, and yet lead to an efficient implementation through Gomory-Hu tree representations. Below, we further elaborate on these points.

Complexes of proteins are at the heart of many fundamental biological processes, including e.g. RNA metabolism, signal transduction, energy metabolism, and translation initiation. As noted, the process of efficiently purifying [5,4] protein complexes and identifying their structure and function has remained a challenge. The most common experimental techniques result in the Yeast two-hybrid protein-protein interaction networks, which encode pair-wise interactions between proteins, and thus hold the promise to yield information about large-scale phenomena such as participation in protein complexes, as examined in [3,12,7,9,13]. It has been a well-known problem that Y2H experiments suffer from high false positive (FP) and false negative (FN) rates. To overcome these limitations, one needs algorithmic approaches robust against high FP and FN rates. Thus, even when the details of protein complexes become "disguised" by FN or become intertwined with each other by FP, these algorithms could exploit the fact that proteins within complexes still remain connected by adequately many paths in the network. However, this qualitative statement requires a quantitative justification, namely, as the number of false edges (positive or negative) increases, how and when do these algorithms break down? What is the nature of the algorithmic degradation: slow and graceful, or sudden and catastrophic? What is the best algorithmic framework, in which they could be studied? Our main results are as follows:

Algorithmic Results: We devise and implement a novel algorithm based on max-flow and their representations through the classical Gomory-Hu tree data structures. We perform both theoretical and practical complexity analysis. We describe and conduct its performance and robustness analysis with respect to practical data using Meila's variational information [21] and Separation [22].

Experimental results: We consider *Saccharomyces cerevisiae* as a *model* organism for our study, since its Y2H network as well as its protein complex data are most complete. Data for protein Y2H pairwise interactions and protein complexes were taken from the BioGRID [2] and MIPS [1] databases (3930 proteins and 6219 Y2H interactions). On yeast network, we identify 38 statistically significant protein clusters, among which there are 20 protein complexes and 16 functional modules. Identified protein complexes cover 61% of all existing BioGRID/MIPS complexes, which have sufficient data coverage (or 72% of non-broken complexes, as described in Section 6).

Supplementary material: Supplementary and output data are available from http://research.rutgers.edu/∼ amitrofa/predictions.html

We begin in section 2 with motivation and intuition behind our algorithm and its formal presentation in section 3. We present the robustness agains FP and FN by measuring the Variation of Information and Separation for several clustering techniques in section 4. In section 5 we discuss the criteria for statistical significance of computed clusters, and show results on yeast Y2H PPI data in section 6. Finally, we compare our results to previously published work in section 7 and conclude the paper.

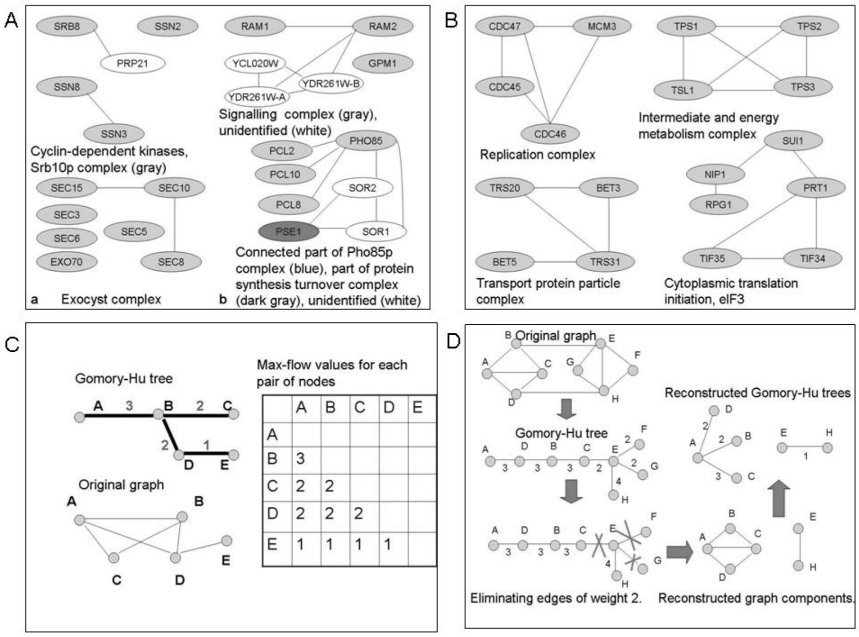
2. Background

The Y2H experiments are known for high false positive and false negative rates: two adjacent proteins might not belong to the same protein complex (FP; Figure 1 A: **b**) as well as proteins from the same complex might not share an edge (FN; Figure 1 A: **a**). These phenomena raise questions about the validity of the direct statistical examination of pure Y2H networks.

With current data coverage and high FN rates, protein complexes of the Y2H PPI networks suffer from low connectivity within. Among all existing Y2H edges, only 6.14% connect protein pairs which participate in the same protein complex. In fact, there are 788 protein complexes (from BioGRID and MIPS) with at least 3 nodes. Of those, 463 do not have a single Y2H edge in the complex, 129 have only one Y2H edge, and 71 have two edges. There are only 125 complexes which contain at least three Y2H edges in the complex and can *potentially* have a minimum level of connectivity necessary to be identified by a connectivity-based computational method.

The majority of graph-based methods for extracting protein complexes look for densely connected, clique-like regions of the PPI network [9,13,12,3]. However, the problem of noise in Y2H experiments required these methods to supplement pair-wise interactions with other biological markers, as co-expression [7], functional annotation [12], small-scale immunoprecipitation [13], microarrays [19], or inter-specie data for conserved protein complexes [3,20].

If a protein complex corresponds to a clique-like subgraph in the Y2H PPI graph, then increased FP and FN rates might at least interfere with and at most preclude the search for such structures. For example, as shown in Figure 1 A: **b** high FP rate can produce areas of "false" density or increase the connectivity between complexes making them impossible to be identified in the network. Likewise, high FN rate can disguise clique-like protein complexes. However, even if proteins from a complex lose a few


A
SRB8
SSN2
PRP21
SSN8
SSN3
Cyclin-dependent kinases, Srb10p complex (gray)
SEC15
SEC10
SEC3
SEC6
SEC5
EXO70
SEC8
a Exocyst complex
RAM1
RAM2
YCL020W
GPM1
YDR261W-B
YDR261W-A
Signalling complex (gray), unidentified (white)
PHO85
PCL2
PCL10
SOR2
PCL8
PSE1
SOR1
b Connected part of Pho85p complex (blue), part of protein synthesis turnover complex (dark gray), unidentified (white)
B
CDC47
MCM3
CDC45
CDC46
Replication complex
TPS1
TPS2
TSL1
TPS3
Intermediate and energy metabolism complex
TRS20
BET3
BET5
TRS31
Transport protein particle complex
SUI1
NIP1
RPG1
PRT1
TIF35
TIF34
Cytoplasmic translation initiation, eIF3
C
Gomory-Hu tree
A 3 B 2 C
2 1
D E
Original graph
A B C D E
Max-flow values for each pair of nodes
D
Original graph
B E A C G F D H
Gomory-Hu tree
A D B C E F G H
3 3 3 2 2 2 4
Reconstructed Gomory-Hu trees
D 2 A 2 B 3 C
E 1 H
A D B C E F G H
3 3 3 4
Eliminating edges of weight 2.
A B C D E H
Reconstructed graph components.

	A	B	C	D	E
A					
B	3				
C	2	2			
D	2	2	2		
E	1	1	1	1	

Figure 1. **A**: **(a)**Protein complexes that have low Y2H connectivity. **(b)** Protein complexes with "fused" out-of-complex proteins. **B**: Examples of protein complexes that contain 2-edge connected subgraphs. **C.** Gomory-Hu tree and its matrix representation. **D.** Cutting the small-weight edges of a Gomory-Hu tree to induce a partition on the nodes.

edges, they should still be connected by enough paths in the network.

If we take a *path* between two proteins as evidence that they are in the same complex, then the number of *edge-disjoint paths* is related to the degree of confidence we have of complex co-membership (an edge is also a path, but a path is not limited to an edge).

We examined the yeast Y2H PPI network for the number of edge-disjoint paths between protein pairs that belong to the same protein complex (in-complex) vs those that do not belong to the same protein complex (non-complex), thus covering all possible protein pairs. In Figure 2,A, we show an example of a distribution of edge-disjoint paths in each group: it is more common for non-complex group to share just one path, and in-complex group shows a clear evidence of sharing two and more paths compared to non-complex group. Overall, the fraction of protein pairs sharing one path over those sharing more than one path for the in-complex group is 1.059 and

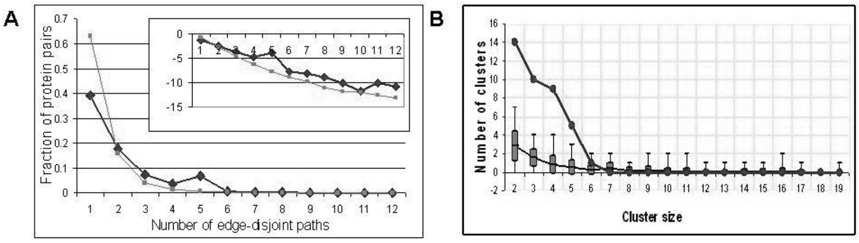


Figure 2. **A.**Distribution of edge-disjoint paths for protein pairs that belong to the same protein complex (thick dark color) vs pairs that do not belong to the same protein complex (light color). The embedded chart shows the same values in log_2 scale. **B**. Number of clusters as a function of cluster size in the whole yeast Y2H PPI network (thick dark upper line) and in the random graphs (light lower line). On the lower line: rectangles represent standard deviation, with max and min as up/down bars.

for non-complex group is 2.868, emphasizing the importance of the greater number of edge-disjoint paths for proteins from the same complex. For pairs of proteins that *do not share* an edge, the same dynamics is observed: the above is 1.081 for in-complex and 2.873 for non-complex group.

The number of edge-disjoint paths between a pair of nodes in the network (in our case unweighted and undirected) corresponds to the value of the *maximum flow* between that pair. However, there is no need to consider all $\binom{n}{2}$ node pairs in the network, since the number of edge-disjoint paths (or maximum flow) for all pairs of nodes can be calculated in only $n - 1$ steps and succinctly represented in a *Gomory-Hu tree* [6], as detailed below.

3. Methods

We begin by computing a Gomory-Hu tree for each connected component of the PPI graph. A Gomory-Hu tree is a weighted tree that spans nodes of a graph such that the max-flow between any two nodes in the graph is the same as the max-flow between them in the tree. That is, the max-flow from p_α to p_β in the network has value equal to the minimum edge on the path between these nodes in the Gomory-Hu tree, as shown in Figure 1 **C**. To compute max-flow value, we use a Ford-Fulkerson method: the best known deterministic max-flow algorithm for the undirected unweighted graph is one proposed by Matula [10] and Nagamochi and Ibaraki [11] that runs in $O(|P||E|)$ steps (where $|P|$ is the number of nodes/proteins and $|E|$ is the number of edges). Thus, the time complexity of our algorithm is $O(|P|^2|E|)$.

First we remove minimum-weighted edges from the Gomory-Hu tree. Removing an edge induces a bipartition between the nodes of the tree.

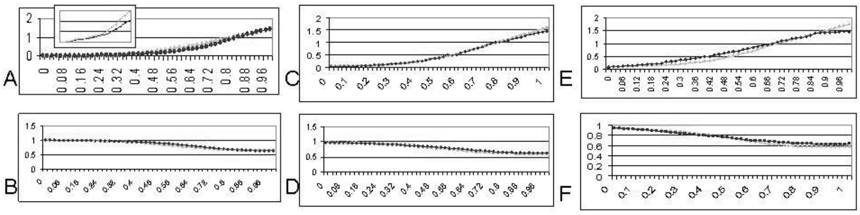
Thus an edge in the Gomory-Hu tree corresponds to an edge-cut in the PPI graph. After such elimination we *recompute* a Gomory-Hu tree for each induced connected component, since the forest obtained by removing edges (of weight > 1) from the Gomory-Hu tree is no longer the Gomory-Hu trees of the partitions, as for example shown in the Figure 1 **D**. We proceed recursively, by eliminating least weighted edges and recomputing Gomory-Hu tree for each induced connected component until there are no more edges to eliminate (see full formal description in *Supplementary material*).

We call the set of nodes in each connected component of the Gomory-Hu forest a **cluster**. We eliminate singleton nodes at each phase, saying that they *disappears*. With every elimination phase, each Gomory-Hu tree becomes smaller, splitting clusters or reducing their size until each cluster disappears. Clusters found this way are then subjected to further selection according to criteria of statistical significance, as described in 5.2.

4. Robustness via Statistical Analysis

We examine the robustness of our algorithm by computing Variation of Information (VI) [21] and Separation [22] as we vary the number of FP and FN in a randomly constructed network. Since we think of protein complexes as highly connected "clique-like" structures in the network [9,13,12,3], we build our random *test graph* in the following way: we introduce complete graphs of size from 10 to 2 and singletons (following the power-law distribution: 10 graphs of size 10, 20 graphs of size 9, etc, 300 singletons), similarly to the approach described in [22]. These groups of nodes are our initial complexes. Then we delete some % of random edges (FN) and/or add edges (FP).

Separation measure is relevant to the geometric mean of sensitivity and positive predictive value, as defined in [22]. High separation values indicate bidirectional correspondence between a cluster and a complex and thus are more favorable. *Variation of Information* is another useful metric based information-theoretic criterion that measures how much information is lost or gained in going from clustering C to C'. In our case, C corresponds to the initial complexes. If we let n be the number of nodes and K be the total number of clusters, with n_k being a size of cluster C_k, then the uncertainty (or entropy) of the clustering C is defined as $H(C) = -\sum_{k=1}^{K} P(k)\log(P(k))$, where $P(k) = n_k/n$. The joint distribution that a point belongs to cluster C_k in C and to cluster $C'_{k'}$ in C' is $P(k,k') = \frac{|C_k \bigcap C'_{k'}|}{n}$. Then the mutual information between clustering C and C' is $I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log(\frac{P(k,k')}{P(k)P'(k')})$ And finally, the



G j	2	4	6	8	10	12	14	16	18	20
Y2H PPI network	2221	807	400	238	160	114	85	67	56	49
Protein complexes	658	428	251	132	95	62	42	35	22	17
Within-complex degrees	209	28	7	1	0	0	0	0	0	0

Figure 3. **A-F** Each curve (black for our method and gray for MC) represents the value of VI (first row) or Separation (second row) as the % of edges removed increases, averaged over 10 random runs. **(A-B)** edge removal **(C-D)** edge removal with 5% of randomly added edges. **(E-F)** edge removal with 10% of randomly added edges. **G**. Results on the training set of the yeast Y2H PPI network: threshold $d = 6, \ldots, 17$. All nodes with degree $\geq d$ are eliminated.

Variation of Information is $VI(C, C') = H(C) - I(C, C') + H(C') - I(C, C')$. Higher VI corresponds to bigger deviation from the original clustering C.

We compare our method with Markov Clustering (MC) [23], which is reported as the most robust clustering on PPI networks in [22]. Since the protein complexes in current PPI networks suffer from low connectivity, it is more important to examine the robustness of the algorithms against increasing FN rates. We present some results in Figure 3, which show that our approach is equally or more robust compared to MC when examined against increased FP (by 5% and by 10%, which are most likely to exist in Y2H PPI networks) and varying FN rates. Both methods show smooth curves toward increased FN rates.

5. Experiments

5.1. *High degree nodes*

To minimize the number of non-selective (possibly FP) interactions that would give statistically insignificant clusters, the common practice is to eliminate "excessive-degree" nodes from the Y2H PPI graph, as for example exercised in [13]. To learn the degree threshold, we select a so-called "training set" , which corresponds to about $\frac{1}{4}$ of the network (the remaining part is called a "testing set"). To choose a training set, we start with a random protein in the graph and accumulate the desired number of nodes by breadth-first search. In the learning, we eliminate nodes and the outgoing edges according to various degree thresholds d and evaluate the node/edge elimination effects by various performance measures. First, we calculate the percent coverage, P, of how many final clusters fully correspond to MIPS/BioGRID protein complexes. Additionally, we introduce a new measure of protein complex coverage, i.e., 2-edge connectedness. The graph is 2-edge connected if there are at least two edge-disjoint paths between every pair of nodes in the graph, consider some examples shown in Figure 1,B.

We found that from 125 MIPS protein complexes with at least three Y2H edges, 74 (59.2 %) are fully or partially 2-edge connected. However, these protein complexes often overlap with each other or are the subsets of each other, producing data redundancy that can negatively influence the analysis. There are 33 not overlapping non-redundant 2-edge connected protein complexes, which we use in our further statistical analysis.

We measure Q, the recall rate for 2-edge connected complexes, that are in the training set. For example, in our training set initially there exist 6 2-edge connected protein complexes and Q is the % of these 6 that we identify in each run. We show P and Q for each run in Figure 3, G : the highest Q values are observed with $d = 13-16$. Among those, $d = 13$ executes highest P, which we consider our threshold and eliminate all nodes of degree higher than 13 from our dataset (85 nodes or 2.16 % of total network nodes).

5.2. *Statistical Significance of Clusters*

As the algorithm proceeds, many clusters of different sizes are generated. The final part of our algorithm is to estimate statistical significance of computed clusters and decide which correspond to protein complexes and functional modules. To measure the statistical significance of the cluster, we need to account for the probability of finding such cluster in a random graph. To generate random graphs, we use the Maslov-Sneppen procedure [14], which shuffles the edges of the original Y2H PPI network so that the number of interactions for each protein in the network is preserved.

Size-based p-value: First, we calculate p-value for clusters of different sizes. The Figure 2,B shows enrichment in the number of clusters of sizes 2 to 6 in the original Y2H PPI graph, as compared to results on 100 random graphs. Clusters of size 7 and higher, in contrast, appear more often at random. For each cluster of size s, we calculate p-value as a probability of finding a cluster of size s at random, fit to a normal distribution. Clusters of size $2, 3, 4$ and 5 showed p-value $p < 1 \times 10^{-4}$, which we consider statistically significant. Clusters of size 6 showed $p = 0.20$ and therefore can appear at random with reasonably high probability.

Density-based p-value: We define a "cluster-network density", *CND*, as the difference between the average number of edge-disjoint paths per pair of proteins in the cluster, ED_c, and the average number of edge disjoint paths from the proteins of this cluster to the proteins in the rest of the network (ignoring proteins from different connected components), ED_r. Thus $CND = ED_c - ED_r$ reflects the difference between the connectivity inside the cluster and connectivity of this cluster with the rest of the network. Of course, all original clusters from the yeast Y2H PPI network show CND greater than 0. Here we again consider 100 random graphs generated by the Maslov-Sneppen procedure [14] and calculate p-value (fit to a normal distribution) *per cluster* produced. For each cluster of the original Y2H PPI network, p-value reflects the probability that CND at random would be greater or equal to CND in the original network. To correct for multiple hypothesis tested, we apply Bonferroni Correction (the number of hypothesis tested is equal to the number of observed original clusters). We consider those clusters with corrected p-values less than 1×10^{-4} as statistically significant. It appeared that clusters with p-values $< 10^{-4}$ do not violate the statistically significant sizes shown in Figure 2,B. We report 38 out of 56 clusters as being statistically significant according to the criteria described above. Among clusters with p-value $> 10^{-4}$, two represent protein complexes and two correspond to functional modules.

6. Results

We consider a cluster as a *match* if *all* of its proteins belong to the same protein MIPS complex (at the lowest hierarchical level). Also we use a measure of biological importance, similar to one defined in [13] – a *Functional Module*, as a group of proteins that participate in the same process in the same location, however not necessarily at the same time. In order for a cluster to be identified as a Functional module, its proteins should reside in the same cellular location and should share similar/relevant functions (Gene Ontology classification). Even stronger supporting evidence for Functional modules includes co-expression and literature co-citation (we used tables and criteria provided by [8] for pairwise log-odds scores). Since in many cases we cannot say with certainty whether proteins enter a process at different times or at the same time, clusters from this category are strong candidates for protein complex predictions.

Table 1. Final clusters of testing and training sets in the Yeast network.

Sets	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4} \bigcup \frac{3}{4}$
Total clusters with $p < 10^{-4}$	10	28	38
Clusters that cover MIPS complexes	5	15	20
FM with co-location and (co-expression or co-citation	1	4	5
FM with co-location	4	7	11
FM with limited information	0	2	2

We present results for both the $\frac{1}{4}$ and $\frac{3}{4}$ of the network in Table 1. Among 38 clusters with p-value $< 10^{-4}$, there are 20 MIPS complexes and 18 functional modules. Five of the functional modules are supported by co-expression of participating proteins or co-citation from the literature, thus making a strongly grounded predictions for new protein complexes, as shown in *Supplementary material.* Two clusters had weaker evidence of forming a functional module primarily due to lack of information about the functional annotation or cellular location of participating proteins.

For completeness, we also studied a recall rate, which corresponds to the proportion of 33 2-edge connected complexes covered. As we select our training and testing sets, 5 2-edge connected complexes become broken, resulting in 28 2-edge connected MIPS complexes in both sets, 20 of which we identify (yielding recall rate of 72%). Additionally, we characterize the performance of our method by a parameter M, the fraction of proteins in the *matched* cluster over proteins in the 2-edge connected part of the corresponding complex: 18 out of 20 clusters show $M = 1$.

In general, among the 2-edge connected complexes, there are 17 triangles, 13 4-node, two 5-node, and one 7-node graphs. Thus, the clusters that cover 2-edge connected complexes are partially bounded by the above sizes.

We have applied our method to other species: we list 31 clusters for human, 17 for mouse, and 29 for worm, whose sizes satisfy our stringent statistical-significance criteria, as shown in *Supplementary material.*

7. Discussion

In this section, we compare our method to those previously described in the literature (the comparison is made with the best available results for each method), as shown in Table 2 and discussed below.

King et al. [12] develop the restricted neighborhood search clustering algorithm using a cost function. After generating clusters, proteins are selectively chosen from clusters using a filtering model. They identified 23 clusters matching MIPS protein complexes by 90 % of proteins in a cluster (6 of which matched by 100% – identified by "→" in the Table 2). We identify 35 new clusters not covered by their method, among which there are 17 new protein complexes.

Table 2. Precision is the number of clusters covering complexes in 100% of cluster proteins over total number of clusters. Recall is the number of 2-edge connected complexes covered by clusters over a total number of 2-edge connected complexes. Intersection reflects the overlap between complexes identified by us and other methods.

	Precision	Recall	Intersection
Our method	20/38 = 53 %	20/33 =61% → 20/28=72%	20
King et al [12]	23/30 → 6/30=20%	4/33 = 12%	4
Bader et al [9]	54/209 = 26% → 0/209=0%	0/33=0%	0
Spirin et al [13]	30/67 = 45%	18/33= 55%	16

Bader and Hogue [9] present algorithm that detects densely connected regions in PPI networks. They generate 209 clusters, 52 of which matched MIPS complexes in at least 20% of their proteins (with the highest overlap being 43%, thus bringing the number of clusters that match protein complexes by 100% of their proteins to 0, as reflected in Table 2).

Spirin and Mirny [13] look for heavily connected, clique-like groups of nodes in the network (supplemented with hypothesis-driven studies such as coprecipitation, omitted by our method). The union of *three* different clustering methods identified 30 clusters corresponding to protein complexes. We identify 4 new protein complexes and 13 new functional modules. The MCL clustering [23] has not been applied to search for protein complexes yet, but only to cluster proteins based on their sequence similarity.

An interesting min-cut clustering approach of Tarjan et al [17], which was applied to find communities in web and citation networks, introduced an artificial sink node connected to all other nodes. We plan to expand our understanding of competitive bounds for communities' sizes addressed in [17]. Another interesting approach described by Newman in [15] (later extended in [24]) describes graph decomposition based on edge betweenness, defined as the number of shortest paths which go through an edge. Hartuv et al. in [16] present an algorithm based on min cut idea, which shows an improved time complexity and generates clusters with diameter 2 (two vertices are either adjacent or share one or more common neighbors). We do not require nodes in the cluster to be adjacent or to necessarily share a neighbor; however, they can be connected by much longer edge-disjoint paths.

One promising future direction for our method would be to assign a confidence score for each Y2H interaction (i.e. conservation of the interaction across species). It is possible to define a distance-based measure between proteins and use a Diffusion Map for spectral clustering, as in [18]. However, this method is very computationally expensive and hard to scale to large datasets. We plan to explore an efficient implementation of a continuous approach of diffusion maps with discrete approach of Gomory-Hu trees.

8. Conclusions

We have presented an efficient algorithm for identifying protein complexes through manipulation of the Gomory-Hu tree of the PPI Y2H network. Our method is shown to be robust against high FP and FN rates and capable of producing clusters of high quality when compared to other approaches. Identified functional modules are strong candidates for complex predictions and constitute reliable material for experimental research.

References

1. The MIPD database: http://mips.gsf.de/
2. The BioGRID database: http://www.thebiogrid.org/
3. R. Sharan, T. Ideker, B. Kelley, R. Shamir, B. Karp, *Recomb.* 282–289 (2004).
4. A. Gavin, M. Bosche and R. Krause, *Nature.* **415**, 141–147 (2002).
5. Y. Ho, A. Gruhler and A. Heilbut, *Nature.* **415**, 180-183 (2002).
6. R. Gomory and T. Hu, *J. of SIAM.* **4**, 551-570 (1961).
7. R. Jansen, H. Yu, et al *Science.* **302(5644)**, 449-53 (2003).
8. I. Lee, S. Date, A. Adai, E. Marcotte, *Science.* **306(5701)**, 1555-1558 (2004).
9. G. Bader and C. Hogue, *BMC Bioinformatics.* bf 4(1), 2 (2003).
10. D. Matula, *FOCS.* 249–251 (1987).
11. H. Nagamochi and T. Ibaraki, *SIAM J.Disc. Math.* **5**, 54–66 (1992).
12. A. King, N. Przulj, I. Jurisica, *Bioinformatics.* **20(17)**, 3013–3020 (2004).
13. V. Spirin and L.A. Mirny,L.A, *Proc Natl Acad Sci.* **100**, 12123–12128 (2003).
14. S Maslov and K. Sneppen, *Science.* **296**, 910-913 (2002).
15. M. Newman, *Phys. Rev. E.* **74**, 036104 (2006).
16. E. Hartuv and R. Shamir, *IPL.* **76(4-6)**, 175–181 (2000).
17. G. Flake, R. Tarjan, K. Tsioutsiouliklis, *Internet Math.* **1(4)**, 385-408 (2004).
18. G. Lerman and B.E. Shakhnovich, *PNAS.* **104(27)**, 11334-11339 (2007).
19. J. Chen and B. Yuan, *Bioinformatics.* **22(18)**, 2283–2290 (2006).
20. E. Hirsh and R. Sharan, *Bioinformatics.* **23(2)**, e170-e176 (2007).
21. M. Meila. *COLT.* (2003).
22. S. Brohee and J. Van Helden, *BMC Bioinformatics.* **7:488** (2006).
23. J. Enright, S. Dongen, A. Ouzounis, *Nucl. Acids Res.* **30(7)**, 1575-84 (2002).
24. F. Luo, Y. Yang, et al *Bioinformatics.* **23(2)** 207-14 (2007).