Predicting interactions in protein networks by completing defective cliques

Haiyuan Yu*, Alberto Paccanaro*, Valery Trifonov*, and Mark Gerstein¶

Department of Molecular Biophysics and Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520, USA

*These authors contributed equally to this work.
¶ To whom correspondence should be addressed.
Tel: +1 203 432 6105; Fax: +1 360 838 7861;
Email: Mark.Gerstein@yale.edu

Submitted to *Bioinformatics*

ABSTRACT

Datasets obtained by large-scale, high-throughput methods for detecting protein-protein interactions typically suffer from a relatively high level of noise. We describe a novel method for improving the quality of these datasets by predicting missed protein-protein interactions, using only the topology of the protein interaction network observed by the large-scale experiment. The central idea of the method is to search the protein interaction network for defective cliques (nearly complete complexes of pair-wise interacting proteins), and predict the interactions that complete them. We formulate an algorithm for applying this method to large-scale networks, and show that in practice it is efficient and has good predictive performance. More information can be found on our website: http://topnet.gersteinlab.org/clique/.

1 INTRODUCTION

A fundamental problem in modern biology is the identification of the complete set of interactions among the proteins in a cell (Jansen, et al., 2003; Marcotte, et al., 1999; 1999). Different experimental methods are available to identify such interactions, and they can be roughly divided into two main categories: small-scale (low throughput) and large-scale (high throughput) techniques. Given a set of proteins, small-scale techniques such as co-IP determine the interaction between one pair of proteins at a time (Bader, et al., 2003; Mewes, et al., 2002; Xenarios, et al., 2002; Xia, et al., 2004). On the other hand, large-scale techniques, e.g. yeast two-hybrid and TAP-tagging, allow identifying a large number of interacting pairs in a single experiment (Gavin, et al., 2002; Ho, et al., 2002; Ito, et al., 2000; Uetz, et al., 2000).

With the advent of genome-wide analysis, we are interested in the identification of the interaction among a great number of proteins (even of all the proteins in a genome). When the number of proteins is in the thousands, the number of possible interacting pairs is in the millions (Kumar and Snyder, 2002). To discover all these interactions using small-scale experiments becomes very labor-intensive and time-consuming, and in this situation large-scale experiments are preferred.

However, low throughput experiments allow much more precise identification of the interacting pairs than high throughput experiments – the latter are known to be more error-prone (Jansen, et al., 2002; von Mering, et al., 2002).

Two types of errors are possible: the large-scale experiment can wrongly indicate that an interaction exists, i.e. yield a false positive (FP); or it can fail to detect an interaction that actually exists, thus producing a false negative (FN). However, experimentalists would agree that these two types of errors occur with different frequency in large-scale experiments. While false positives have "higher visibility" due to the relatively small number of true interactions, it is generally observed that experiments allow a higher absolute degree of confidence when an interaction is observed, but a much lower degree when no interaction is detected. In other words, most of the errors (as an absolute count, not relative to the numbers of actual interacting or non-interacting protein pairs) are false negatives: it is believed that when no interaction is detected, it is not unlikely that the interaction actually exists, but the experiment has failed to detect it. In support of this observation, Figure 1 shows the differences between the low-throughput and high-throughput experimental data on protein-protein interactions in a subset of 56 proteins of S. cerevisiae, for which we were able to obtain complete matrices of experimental results.

The results of the two types of experiments were the same for 1033 of the 1596 pairs of proteins (including possible self-interactions); of the 563 cases when the results were different, 521 (92.5%) were false negatives and 42 (7.5%) were false positives. Ideally, we would like to have a computational method which would be able to correct many of the errors made by large-scale interaction experiments. In this paper we propose a new method, based purely on topological properties of graphs representing protein interaction networks, that attempts to detect those interactions that have been missed by large-scale experiments. Our algorithm searches for defective cliques in these graphs and predicts the interactions which complete them to full cliques.

The basic idea of the algorithm derives from the way in which large-scale experiments are carried out, and particularly from the matrix model interpretation of their results (Bader and Hogue, 2002; Gavin, et al., 2002; Ho, et al., 2002; Rigaut, et al., 1999). In these experiments one protein, the bait, is used to pull out the set of proteins interacting with it, i.e. its protein complex, in the form of a list. When such lists differ only in a few elements, it is reasonable to assume that this is due to experimental errors, and the missing elements should therefore be added. Each list can be represented as a fully connected graph in which proteins occupy the nodes. Then the problem of identifying lists that differ in only a few elements is equivalent to finding a clique with a few missing edges, which we shall call defective clique.

The rest of this paper is organized as follows. In section 2 we shall introduce some basic notions and give an overview of our method. In section 2.2 we present a more efficient and practically useful algorithm

implementing the method, and in section 3 we present the results of applying the method to several datasets obtained from experimental observations of the protein interaction network of yeast.

2 METHOD

Before describing our approach, let us introduce some basic terms. A graph is a pair (V, E) of a set of *vertices* V and a set of *edges* $E \subseteq V \times V$, where each edge is a pair of the vertices it connects; if $\langle v_1, v_2 \rangle$ is in E , then the vertices v_1 and v_2 are *adjacent*. In a graph representing a protein interaction network, the vertices are proteins, and the edges are the pairs of interacting proteins.

A *clique* in a graph is a set K of vertices such that $K \times K \subseteq E$, i.e. each pair of vertices in K is connected by an edge in E. The *size* of this clique is the number of vertices in it.

As we discussed in section 1, under the matrix model interpretation of the results of large-scale experiments, two proteins interacting with the same protein clusters are likely to interact with each other. Thus in graph-theoretic terms our approach is based on the following observation about protein interaction networks:

> **(*) If vertices *P* and *Q* are both adjacent to each vertex in a clique *K*, then it is likely that *P* and *Q* are adjacent to each other, if they are not adjacent already.**

This observation can be depicted as shown in Figure 2A; in this example the size of the clique K is 5. The dashed edge between P and Q corresponds to an interaction which is missing from the experimental data, but which (according to observation (*)) is very likely to occur. We say that P, Q, and K form a *defective clique* KPQ with a missing edge PQ. (Note that a defective clique could in theory have more than one missing edge.)

Clearly the size of K plays an important role in determining how likely it is that P and Q interact. For example, if the size of K is 1 (i.e. P and Q both interact with one or more proteins, but those proteins do not interact among themselves), the likelihood of an interaction between P and Q is much smaller than in the case when the size of K is, say, 42. Thus a natural parameter of a prediction algorithm based on observation (*) is the minimal size k of K for which the interaction PQ is predicted.

Another parameter with which we can extend observation (*) is the number of edges missing from the clique when its size is sufficiently large. We will discuss the effects of this parameter in section 2.2, when we describe our algorithm in detail.

2.1 An algorithm for finding defective cliques

Our definition of a defective clique does not immediately suggest a method for finding such patterns in a protein interaction network. For this purpose it is useful to find an alternative characterization of a defective clique in standard graph-theoretic terms, which will allow us to use some off-the-shelf algorithms.

The main idea of our algorithm is based on the realization that a defective clique KPQ of size n with one missing edge is the union of two (complete) cliques of size n-1, namely $K \bigcup \{P\}$ and $K \bigcup \{Q\}$, as shown in Figure 2B. Thus we can reduce the algorithm for finding defective cliques to the following two main steps (which may be repeated until no new edges are added):

Step 1: Find all cliques in the network.
Step 2:

- Find pairs of cliques overlapping on all but one node each.
- In each of these pairs predict the edges between the non-overlapping nodes.
- Add the new edges to the network.

(Note that defective cliques with more than one missing edge could also be determined by applying this recipe.)

However, directly applying this naïve recipe to typical protein interaction networks is unrealistic, for the following reason: Since every subset of nodes in a given clique is itself a clique, the number of all cliques in a graph is at least 2^q, where q is the size of the largest clique in the graph. For example, the large-scale experimental data for the protein interaction network of S. cerevisiae we used to test our algorithm (see Section 3) contains four cliques of size 38; this yields more than 10^{12} cliques (even if we do not consider cliques of size less than 5, whose number is negligible), hence more than 10^{23} pairs of cliques to check in Step 2 of the algorithm. Since this number is prohibitively large, we need a more effective formulation of the algorithm. For this purpose in the next section we design an equivalent algorithm which only considers the maximal cliques in the graph.

2.2 Improving Efficiency Using Maximal Cliques

A *maximal* clique in a graph G is one which is not contained in any other clique in G. In the worst case the problem of finding all maximal cliques still takes time exponential in the size of the graph1[1]; however, if Step 1 is modified to only produce the maximal cliques in the graph, for the reasons discussed in the previous section the output of Step 1 would be reduced by a factor exponential in the size of the largest clique. This would lead to a corresponding reduction (by the square of that factor) of the running time of Step 2 of the algorithm.

In practice, the protein interaction networks are rather sparse [e.g. less than 15,000 interactions are observed with high confidence in the network of *S. cerevisiae*, out of over 18 million possible pairs of about 6,000 proteins (von Mering, et al., 2002)]. Our results show that existing algorithms for finding maximal cliques (Tsukiyama, et al., 1977) are very efficient on graphs with this structure. However, if we only compare maximal cliques for overlap on all but one node each, as we did with all cliques in the naïve version, the output of this algorithm will not be the same as that of the naïve version. The reason is that if a defective clique consists of a core clique K and two nodes P and Q, Step 2 of the algorithm will not, in general, attempt to match the cliques $K \bigcup \{P\}$ and $K \bigcup \{Q\}$, but two maximal cliques they are contained in, say $K \bigcup K_P$ and $K \bigcup K_Q$ (note that K_P and K_Q always exist, but are not necessarily unique). However, K_P will in general contain other nodes in addition to P, and these nodes might not all appear in K_Q. As a result, the non-overlapping parts K_P and K_Q of the maximal cliques will consist of more than one node each, and Step 2 of the naïve algorithm will fail to predict the edge PQ.

Hence, to obtain the same results as with our original algorithm, we have to modify Step 2 of the algorithm to look for partial overlaps of maximal cliques which differ in more than one node. This leads us to a generalization of the notion of a defective clique, shown in Figure 2C. To obtain the same result as in the original approach, any pair of nodes P_i and Q_i, belonging to the two non-overlapping components K_P and K_Q respectively, must be predicted as interacting, because the original algorithm would have predicted it (since it completes the defective non-maximal clique KP_iQ_i). The maximal size l of non-overlapping sub-cliques K_P and K_Q is a parameter of the algorithm.

Thus one round of the algorithm we use in our experiments becomes:

Step 1: Find all maximal cliques in the network.
Step 2:

- For each pair of maximal cliques, overlapping on at least k nodes and with non-overlapping components of at most l nodes each, predict the edges between all pairs of nodes between the two non-overlapping components.
- Add the new edges to the network.

Since even the number of maximal cliques can be significant (in the hundreds of thousands for some of our experimental datasets), and their sizes can be in the hundreds of nodes, the number of comparisons between nodes in pairs of cliques in Step 2 can still be formidable in practice. We further reduce the time complexity

[1] More precisely, the problem is NP-complete, *i.e.* only exponential-time algorithms for solving it are known.

of Step 2 by organizing the cliques (represented as strings sorted by node index) in a suffix tree. This structure allows us to reuse some comparison results among cliques sharing a common prefix of nodes.

Step 1 of the algorithm has an upper bound of $O(nm\mu)$ for its time complexity (Tsukiyama, et al., 1977), where n is the number of nodes, m – the number of edges, and μ – the number of maximal cliques in the graph. This implies an upper bound of $O(nm\mu+\mu^2)$ for the time complexity of one round of the algorithm (Step 1 followed by Step 2). In our tests the running time was indeed dominated by the time spent in Step 2.

3 RESULTS

We tested our method on two datasets of protein-protein interactions in *S. cerevisiae*, obtained from large-scale experiments (Bader and Hogue, 2002; Yu, et al., 2004). In both cases we compared its predictions with a "gold standard" set of protein pairs, known with high degree of confidence to be "positive" [interacting– protein pairs in the same protein complex determined by the MIPS complex catalog (Mewes, et al., 2002)] or "negative" [non-interacting – protein pairs with different sub-cellular localizations (Kumar, et al., 2002)], as published in Jansen, et al. (2003). Here, the gold standard positives are a collection of small-scale experiment results (Mewes, et al., 2002).

Since neither the large-scale experimental datasets nor the gold standard set are complete (i.e. there are protein pairs for which no experiment has been performed), the question of how to treat missing information arises. We took a conservative approach, assuming that no information indicates no interaction. Since the set of new interactions, predicted by the algorithm, does not decrease with the addition of edges to the input data, our results represent a lower bound on the predictions that would be made with less missing information.

3.1 Performance on a complete dataset

To illustrate the method on a small example, in which we can accurately assess its performance, we considered a sub-graph of the protein interaction network of *S. cerevisiae* for a set of 43 proteins, for which the gold standard is complete (for each pair of proteins it is known if they interact or not, i.e. there is no missing information). Here, we used the large-scale protein interaction network obtained by Yu, et al. (2004). The graph of the gold standard on this subset of proteins consists of four components, all of them cliques; we will refer to them as to G_1 through G_4, as defined in Figure 3A. The graph of the large-scale experimental dataset consists of 6 maximal cliques named E_1 through E_6, of size at least 2, plus 15 singleton nodes (cliques of size 1), shown in Figures 3B and 3C, where the data is presented in the form obtained after running Step 1 of the algorithm. Note that all elements of clique E_2 except for MRPL38 appear also in clique E_1 (protein names in bold; see Figure 3B), and that the pairs of cliques E_3-E_4 and $E_4 – E_5$ each share a node.

Applying Step 2 of the algorithm with parameters $k = 6$ and $l = 17$ to the experimental data finds the partial overlap between cliques E_1 and E_2, and predicts the interactions between MRPL38 and all nodes in E_1 which are not in E_2. After these edges are added, the cliques E_1 and E_2 are merged into one clique of size 24, which is a subset of the gold standard clique G_1 , missing only the protein MRPL49 (in the experimental dataset this node is a singleton, so the clique completion is unable to recover its interactions). These are the only new interactions the algorithm predicts. All of them are positive in the gold standard for this set of proteins, therefore all of the predictions in this case are correct. G_1, E_1, and E_2 consist of Mitochondrial ribosomal proteins.

3.2 Performance on the available S. cerevisiae dataset

We applied the clique completion method to a large-scale experimental dataset of the protein interaction network of *S. cerevisiae* obtained by Bader and Hogue (2002). Unlike the smaller set analyzed in the previous section, the gold standard for the proteins in this dataset is incomplete (as is the dataset itself).

The initial graph contains 6645 edges between 2283 nodes. In this graph Step 1 of the algorithm found 4934 maximal cliques. Step 2 of the algorithm, configured to search for partial overlaps of size at least $k = 4$ and non-overlapping parts of size at most $l = 3$, predicted 437 new interactions. Adding these interactions reduces the number of maximal cliques by 276, showing consolidation of smaller complexes into larger ones, as expected.

As a criterion of the effectiveness of the algorithm we used the likelihood ratio of the predicted interactions, defined in Jansen, et al. (2003) as

$$L = \frac{\frac{P_+}{G_+}}{\frac{P_-}{G_-}}$$

where
P_+ is the number of true positives – predicted interactions which are positive in the gold standard;
P_- is the number of false positives – predicted interactions which are negative in the gold standard;
G_+ is the total number of positive pairs in the gold standard; and
G_- is the total number of negative pairs in the gold standard.

Higher values of L correspond to sets of predictions having higher overlap with the positive and/or lower overlap with the negative gold standard, and generally indicate better predictors. One of the advantages of calculating L is that it naturally takes into account the biased sampling between positives and negative, which is often the case for biological data (see supplementary materials).

The gold standard set contained $G_+ = 8250$ positive and $G_- = 2{,}708{,}622$ negative pairs when restricted to the proteins in this experimental dataset. Of the 437 interactions predicted by the method in this test, 94 were in the gold standard set; of them 73 were positive (P values $< 10^{-10}$; see supplementary materials) and 21 negative, which yields a likelihood ratio of 1141.3, significantly higher than the likelihood ratios of other single features reported in Jansen, et al. (2003) (essentiality, expression correlation, MIPS function, and GO biological process), which are below 400.

The values of the parameters chosen in our test are in a "plateau" of relative stability of the results. In a wider spectrum of parameter values, the likelihood ratio of the predicted set was between 59.13 and 3720.94 when varying the parameters of the algorithm as follows: k (the minimal overlap size) between 4 and 7, and l (the maximal size of the non-overlapping parts) between 1 and 20; the number of predicted interactions was between 12 and 8993. The average running time was below 4 seconds on a desktop machine. We also calculated the ROC curve, and compared our method to the four available large-scale yeast interaction experiments. Our method out-performs all of them (see Figure 4).

Another parameter indicating the quality of the prediction is the functional enrichment in the predicted interactions, i.e. the ratio of the frequency of functionally similar pairs among the predictions to the expected frequency in the yeast genome (see supplementary materials). (Note that functional similarity is not a feature taken into account when constructing the input set or predicting the new interactions.)

The distribution of the likelihood ratio and functional enrichment of the predicted edges as a function of the maximal size of a defective clique they complete is shown in Figure 5A. They show that even for small sizes of the overlap the predicted edges are much more likely to be in the positive than in the negative gold standard, and are significantly more likely to be functionally similar than the average interacting pair.

Taking into account the size of the predicted set and the fact that the predictions were made only on the basis of the topology of the input set, we believe the high value of these measures is a strong argument for the usefulness of this method as a predictor of new interactions.

3.3 Biological examples

With the addition of the 437 predicted interactions, we were able to discover many protein complexes that are not present in the initial network. For example, Casein kinase II complex is composed of two catalytically active subunits (CKA1 and CKA2) and two regulatory subunits (CKB1 and CKB2). It is involved in regulating cell growth and proliferation (Ackermann, et al., 2001). However, based on the original large-scale interaction experiments, the interaction between CKA2 and CKB1 is missing. Therefore, the whole complex could not be determined as a fully-connected clique. We were only able to discover two three-cliques: {CKA1, CKA2, CKB2} and {CKA1, CKB1, CKB2}. Only after our defective clique procedure, CKA2 and CKB1 was predicted to be connected and the whole complex became a four-clique (see Figure 6A).

Another good example is the exosome complex, consisting of 7 proteins (see Figure 6B). It is involved in RNA processing (Mewes, et al., 2002; Mitchell, et al., 1997). In the original large-scale interaction network, RRP43, RRP4, and RRP42 are disconnected. Therefore, the whole complex is divided into three five-cliques: {RRP42, RRP46, SKI6, DIS3, RRP45}, {RRP43, RRP46, SKI6, DIS3, RRP45}, and {RRP4, RRP46, SKI6, DIS3, RRP45}. Our defective clique procedure successfully predicted the interactions among RRP43, RRP4, and RRP42. The complex, thus, became a seven-clique as described in the MIPS complex catalog (Mewes, et al., 2002).

3.4 Comparison with related work

King *et al.* have designed the Restricted Neighborhood Search Clustering (RNSC) algorithm, which partitions proteins into clusters depending on their interactions (King, et al., 2004). The RNSC algorithm can also be viewed as a method for predicting new interactions, if we consider all pairs of proteins in a predicted cluster to be interacting. We compared the predictions of RNSC with those of our algorithm on the datasets published in King, et al. (2004), which are based on the data of von Mering, et al. (2002); the results are shown in Figure 5B. Since the two algorithms represent very different approaches to discovering clusters, the overlap of their predictions is noteworthy.

Bader and Hogue also proposed the Molecular Complex Detection (MCODE) algorithm to discover protein complexes, which can be views as another way to predict new interactions (Bader and Hogue, 2003). The method essentially looks for k-cores in the network. A k-core is a sub-graph G of n ($n \geq k$) vertices with minimal degree k (in G, degree (v) $\geq k$ for every $v \subseteq G$). By definition, all defective cliques determined with the parameters k and l are at least (k+1)-cores, i.e. results from our method will be a subset of the MCODE method. Therefore, our predictions are much more stringent than theirs, whereas the MCODE method could potentially discover more interactions.

4 CONCLUSION

We presented a method for predicting new protein-protein interactions, based purely on topological properties of networks of observed interactions. Comparing the results with the gold standard set and functional annotations confirmed that it is a very good predictor. While computationally expensive, we believe the method has the advantage of being more robust by virtue of its independence of non-topological features such as functional classification.

Acknowledgment

MG acknowledges support from the NIH grant 5P50GM062413-03.

Figure captions

Figure 1. A graphical representation of the symmetric matrix of the differences between complete protein-protein interaction data obtained in small-scale and large-scale experiments on 56 proteins of *S. cerevisiae* (only the upper triangle is shown). There is a colored box for each cell in the matrix indicating the type of the interaction between proteins *i* and *j*. White boxes indicate interactions observed in small-scale but not in large-scale experiments (false negatives); black boxes stand for interactions observed in large-scale but not in small-scale experiments (false positives); gray boxes show protein pairs for which both the small- and the large-scale experiments produced the same result. The number of false negatives exceeds the number of false positives by an order of magnitude.

Figure 2. Schematic illustrations of a defective clique and how the concept evolved. (A). A defective clique in a protein interaction network. *KP* and *KQ* are both (*k+1*)-cliques, with *k* overlapping vertices (i.e. clique *K*). The dashed edge between proteins *P* and *Q* corresponds to a predicted interaction. *KPQ* is a defective clique with a missing edge *PQ*. (B). The decomposition of the defective clique (*KPQ*) into the union of two overlapping cliques (*KP* and *KQ*). (C). Generalized defective cliques. In general, a defective clique consists of two cliques: $K \bigcup K_P$ and $K \bigcup K_Q$. There are two parameters to determine a defective clique: *k*, the size of the overlapping subclique (i.e. *K*); *l*, the size of the nonoverlapping subcliques (i.e. $K_P \bigcup K_Q$). In the defective clique $K \bigcup K_P \bigcup K_Q$, the dashed edges between subcliques K_P and K_Q correspond to predicted interactions.

Figure 3. Subset of the gold standard set without missing information and the corresponding large-scale interaction sub-network. (A). There are 4 maximal cliques in the gold standard set. Please see Supplementary Figure 1 for the network view of these 4 cliques. (B). There are 6 maximal cliques and 15 singleton nodes in the large-scale experimental data. (C). Network view of the experimental data in (B), excluding the singletons.

Figure 4. The trade-off between detection rate and error rate for different values of *k* and *l* to evaluate the performance of our defective clique method. The curve is also known as the Receiver Operating Characteristic curve (i.e. ROC curve) (Egan, 1975). The inset highlights the lower left corner of the ROC curve to show the comparison between our method and the four large-scale experimental datasets.

Figure 5. (A). Distribution of the likelihood ratio *L* of predicted edges as a function of the maximal size of a defective clique they complete. (B). Comparison with the predictions of the RNSC algorithm on three of the datasets published in King, et al. (2004).

Figure 6. Two biological examples of protein complexes that can only be discovered by our defective clique method. (A). Casein Kinase II complex consisting of 4 proteins. (B). Exosome complex consisting of 7 proteins.

References

Ackermann, K., Waxmann, A., Glover, C.V. and Pyerin, W. (2001) Genes targeted by protein kinase CK2: a genome-wide expression array analysis in yeast, *Mol Cell Biochem*, **227**, 59-66.

Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Res*, **31**, 248-250.

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

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

Egan, J.P. (1975) *Signal detection theory and ROC-analysis*. Academic Press, New York.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141-147.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, **415**, 180-183.

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 U S A*, **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.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, **302**, 449-453.

King, A.D., Przulj, N. and Jurisica, I. (2004) Protein complex prediction via cost-based clustering, *Bioinformatics*, **20**, 3013-3020.

Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.H., Miller, P., Gerstein, M., Roeder, G.S. and Snyder, M. (2002) Subcellular localization of the yeast proteome, *Genes Dev*, **16**, 707-719.

Kumar, A. and Snyder, M. (2002) Protein complexes take the bait., *Nature*, **415**, 123-124.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences, *Science*, **285**, 751-753.
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.
Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. and Tollervey, D. (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases, *Cell*, **91**, 457-466.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci U S A*, **96**, 4285-4288.
Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration, *Nat Biotechnol*, **17**, 1030-1032.
Tsukiyama, S., Ide, M., Ariyoshi, H. and Shirakawa, I. (1977) A new algorithm for generating all the maximal independent sets, *SIAM J. Comput.*, **6**, 505-517.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae., *Nature*, **403**, 623-627.
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.
Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res*, **30**, 303-305.
Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks, *Annu Rev Biochem*, **73**, 1051-1087.
Yu, H., Zhu, X., Greenbaum, D., Karro, J. and Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics, *Nucleic Acids Res*, **32**, 328-337.

Figure 1

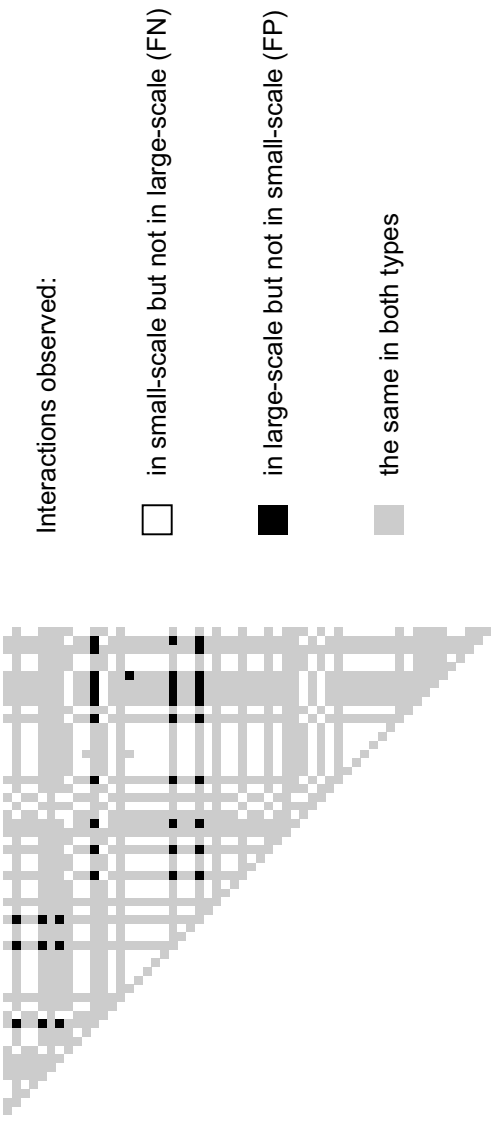
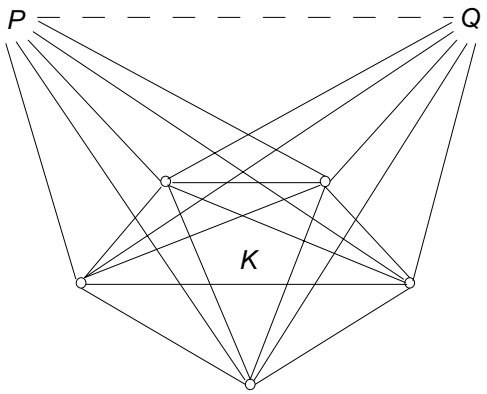
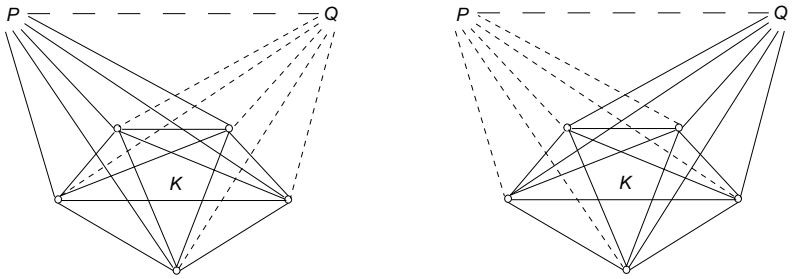

Interactions observed:
in small-scale but not in large-scale (FN)
in large-scale but not in small-scale (FP)
the same in both types

Figure 2

A. A defective clique in a protein interaction network



B. The decomposition of a defective clique into the union of two overlapping cliques



C. Generalized defective cliques

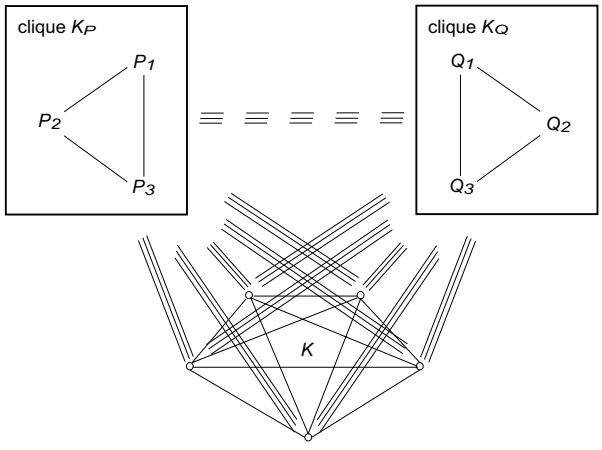


Figure 3

A. The gold standard

clique G_1:

MRPL16	MRPL36	MRPL27	MRPL7	MRPL35
MRP20	MRPL28	MRPL25	MRPL9	MRPL6
MRPL8	MRPL49	MRP49	MRPL38	MRPL13
MRPL20	MRPL4	MRPL3	MRPL24	MRPL44
MRP7	MRPL19	MRPL17	MRPL10	MRPL23

clique G_2:

RPL23A	RPL31A	RPP1A	RPP2B	RPL23B
RPL11B	RPL14B	RPL34B	RPL37A	RPL38
RPP0	RPL26A	RPL18B	RPL5	

clique G_3:

YTA7 RPN1

clique G_4:

TAF30 NUT2

B. The large-scale experimental data

clique E_1:

MRPL16	MRPL36	MRPL27	MRPL7	MRPL35
MRP20	MRPL28	MRPL25	MRPL9	MRPL6
MRPL8	MRP49	MRPL13	MRPL20	MRPL4
MRPL3	MRPL24	MRPL44	MRP7	MRPL19
MRPL17	MRPL10	MRPL23		

clique E_2:

MRPL35	MRPL28	MRPL8	MRPL38	MRPL4
MRPL3	MRP7			

clique E_3: RPP1A RPP0

clique E_4: RPP2B RPP0

clique E_5: RPP2B MRPL10

clique E_6: MRPL6 NUT2

singletons:

RPL23A	RPL31A	RPL23B	RPL11B	YTA7
RPL14B	RPN1	RPL34B	MRPL49	RPL37A
RPL38	RPL26A	RPL18B	TAF30	RPL5

C. Network view of the six cliques in the experimental data

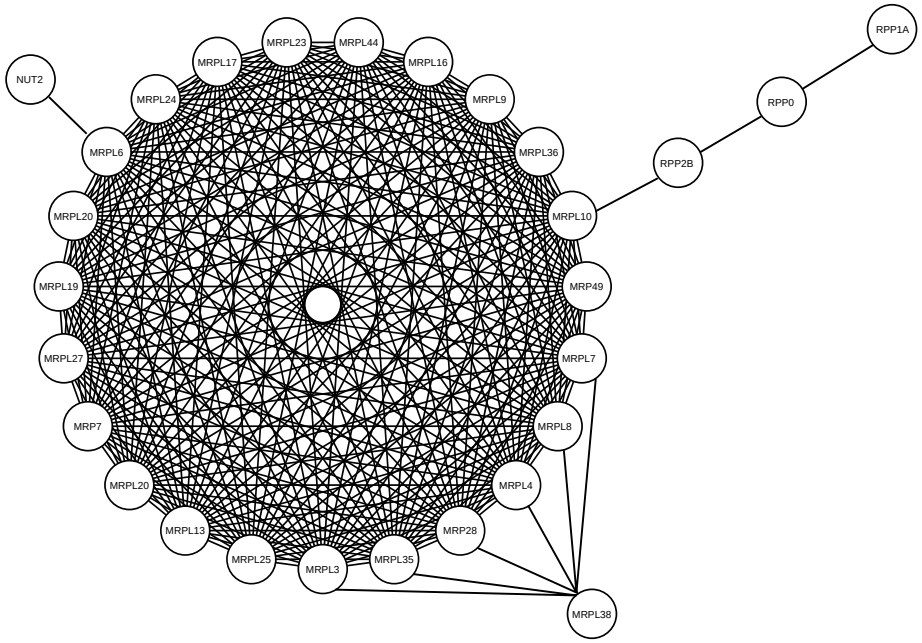


Figure 4

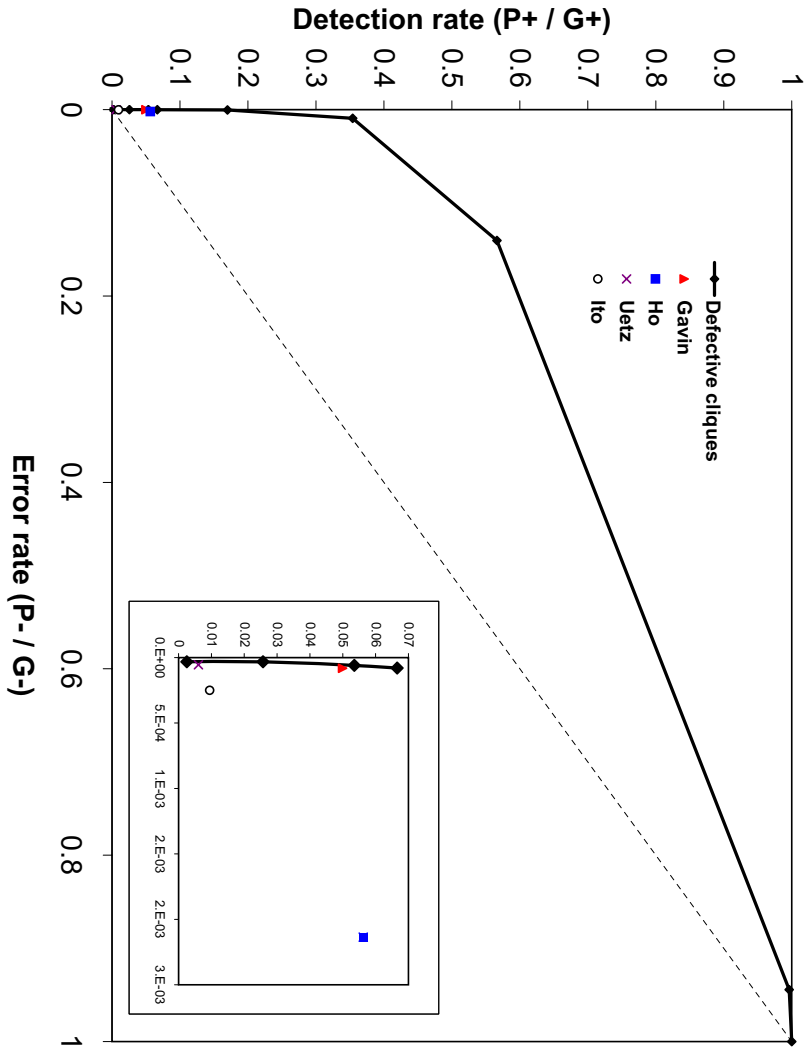

Detection rate (P+ / G+)
Error rate (P- / G-)
Defective cliques
Gavin
Ho
Uetz
Ito
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.E+00
5.E-04
1.E-03
2.E-03
2.E-03
3.E-03

Figure 5

A. Distribution of the likelihood ratio *L* of predicted edges as a function of the maximal size of a defective clique they complete

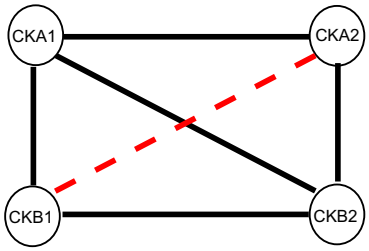
overlap size	total new edges	positive	negative	*L*	functional enrichment
4	259	31	5	2035.57	7.922
5	108	22	7	1031.86	9.784
6	52	18	7	844.25	11.173
7	14	2	2	328.32	8.69
8	4	0	0	N/A	11.173

B. Comparison with the predictions of the RNSC algorithm on three of the datasets published in King, et al. (2004)

number of proteins	number of observed interactions	new interactions predicted by RNSC	new interactions predicted by clique completion	overlap of predictions	probability of overlap of this size at random
988	2000	59	461	24	$<10^{-56}$
2401	11000	337	2710	101	$<10^{-120}$
5321	78000	1581	28180	112	$<10^{-120}$

Figure 6

A.Casein Kinase II (4)



B.Exosome complex (7)