A Bayesian networks approach for predicting protein-protein interactions from genomic data

Ronald Jansen[1, *], Haiyuan Yu[1], Dov Greenbaum[1], Yuval Kluger[1], Nevan J Krogan[2], Sambath Chung[1, 3], Andrew Emili[2], Michael Snyder[3], Jack F Greenblatt[2] & Mark Gerstein[1, 4, †]

Department of Molecular Biophysics & Biochemistry, Yale[1]

Banting and Best Department of Medical Research[2], Department of Molecular and Medical Research, University of Toronto, Toronto, M5G 1L6, Ontario, Canada.

Department of Molecular, Cellular and Developmental Biology, Yale[3]

Department of Computer Science, Yale[4]

266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520, USA

†To whom correspondence should be addressed. Email: mark.gerstein@yale.edu

*Present address: Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 307 West 63rd Street, New York, NY 10021, USA

Accepted for publication in *Science* August 29, 2003.

Abstract

We developed an approach using Bayesian networks to predict protein-protein interactions genome-wide in yeast. Our method naturally weights and combines into reliable predictions genomic features only weakly associated with interaction (e.g., mRNA co-expression, co-essentiality and co-localization). In addition to de novo predictions, it can integrate often noisy, experimental interaction datasets. We observe that at given levels of sensitivity our predictions are more accurate than the existing high-throughput experimental datasets. We validate our predictions with new TAP-tagging experiments. Our analysis, which gives a comprehensive view of yeast interactions, is available at genecensus.org/intint.

Many fundamental cellular processes involve protein-protein interactions, and comprehensively identifying them is important to systematically defining the biological role of proteins. New experimental and computational methods have vastly increased the number of known or putative interactions, catalogued in databases (1-7). Much genomic information also relates to interactions indirectly: Interacting proteins are often significantly co-expressed (as shown by microarrays) and co-localized (to the same subcellular compartment) (8, 9).

Unfortunately, interaction datasets are often incomplete and contradictory (10-12). In the context of genome-wide analyses these inaccuracies are greatly magnified because the protein pairs that do not interact (negatives) far outnumber those that do (positives). For instance, in yeast the ~6000 proteins allow for ~18 million potential interactions, but the estimated number of actual interactions is below 100,000 (10, 13, 14). Thus, even reliable techniques can generate many false positives when applied genome-wide. This is similar to a diagnostic with a 1% false-positive rate for a rare disease occurring in 0.1% of the population, which would roughly produce one true positive for every 10 false ones. Further information is necessary.

Consequently, when evaluating protein-protein interactions, one needs to integrate evidence from many different sources (15-17). Here, we propose a Bayesian approach for integrating interaction information that allows for the probabilistic combination of multiple datasets and demonstrate its application to yeast (18). Our approach can be used for combining noisy interaction datasets and for predicting interactions de novo, from other genomic information. The basic idea is to assess each source of evidence for interactions by comparing it against samples of known positives and negatives ('gold-

standards'), yielding a statistical reliability. Then, extrapolating genome-wide, we predict the chance of possible interactions for every protein pair by combining each independent evidence source according to its reliability. We verified our predictions by comparing them against existing experimental interaction data (not in the gold-standard) as well as new TAP (tandem-affinity-purification) tagging experiments.

Among the many possible machine-learning approaches that could be applied to predicting interactions (ranging from simple unions and intersections of datasets to neural networks, decision trees, and support-vector machines), Bayesian networks have clear advantages (19): They allow for combining highly dissimilar types of data (i.e. numerical and categorical), converting them to a common probabilistic framework, without unnecessary simplification. They readily accommodate missing data. And they naturally weight each information source according to its reliability. In contrast to 'black-box' predictors, Bayesian networks are readily interpretable as they represent conditional probability relationships among information sources.

The gold-standard dataset on which we train ('parameterize') the Bayesian network should ideally be: (i) independent from the data sources serving as evidence, (ii) sufficiently large for reliable statistics and (iii) free of systematic bias. We used the MIPS (Munich Information Center for Protein Sequences) complexes catalog as the gold-standard for positives (6). This hand-curated list of protein complexes is based on the literature (8,250 pairs). A negatives gold-standard is harder to define, but essential for successful training. Thus, we synthesized negatives from lists of proteins in separate subcellular compartments (9). These positive and negative gold-standards satisfy the first

two criteria and provide a good practical solution for the third. (Note, our goal was not to predict physical interactions, but whether two proteins exist in the same complex.)

As a measure of reliability, the overlap of information sources (i.e., 'interaction datasets', which could either be noisy experimental data or sets of genomic features) with the gold-standards can be expressed in terms of a 'likelihood ratio'. For example, consider a genomic feature f expressed in binary terms (i.e., 'present' or 'absent'). The likelihood ratio $L(f)$ is then defined as the fraction of gold-standard positives having feature f divided by the fraction of negatives having f. For two features f_1 and f_2 with uncorrelated evidence, the likelihood ratio of the combined evidence is simply the product $L(f_1, f_2) = L(f_1)L(f_2)$. For correlated evidence, $L(f_1, f_2)$ cannot be factorized this way. Bayesian networks are a formal representation of such relationships between features. The combined likelihood ratio is proportional to the estimated odds that two proteins are in the same complex given multiple sources of information.

We predict a protein pair as positive if its combined likelihood ratio exceeds a particular cutoff ($L > L_{cut}$) (negative otherwise). To get an overall assessment of how the prediction performs, we segmented the gold-standard into separate training and testing sets (using a seven-fold cross-validation protocol). Then we evaluated the number of true (*TP*) and false positive (*FP*) predictions in the testing set. Finally, we applied the Bayesian network beyond the testing set, computing likelihood ratios for all possible protein pairs in the genome.

Figure 1 schematically shows the information sources and results of our calculations. We term the results 'probabilistic interactomes' (PIs), in which each protein pair is associated

with a probability measure for being in the same complex (i.e. likelihood ratio *L*). Our procedure not only allows combining existing experimental interaction datasets (resulting in a PI-experimental or 'PIE'), but also the de novo prediction of protein complexes from genomic datasets (when the input data are not interaction datasets per se, resulting in a PI-predicted or 'PIP').

We combined four interaction datasets from high-throughput experiments into the PIE (1-4) (figure 1b). The PIE represents a transformation of the individual binary-valued interactions sets into a dataset where every protein pair is weighted according to the likelihood that it exists within a complex.

We computed the PIP from several genomic data sources: The correlation of mRNA amounts in two expression datasets (one with temporal profiles during the cell cycle, one of expression levels under 300 cellular conditions), two sets of information on biological function and information about whether proteins are essential for survival (6, 20-22). Although none of these information sources are interaction data per se, they contain information weakly associated with interaction: Two subunits of the same protein complex often have co-regulated mRNA expression and similar biological functions and are more likely to be both essential or non-essential (8).

For computing the PIE and the PIP we used two different types of Bayesian networks: a 'naïve' network for the PIP and a fully connected one for the PIE (19). The naïve network is simpler to compute, but requires information sources providing essentially uncorrelated evidence. In contrast, the fully connected Bayesian network accommodates correlated evidence, which is the case for the four experimental interaction datasets.

Finally, we combined the PIP, PIE and the gold-standard into a total PI (PIT), which represents our most comprehensive view of the known and putative protein complexes in yeast (note 1). Since the PIP and PIE data provide essentially uncorrelated evidence for protein-protein interactions, we chose a naïve network to construct the PIT.

Figure 1c gives an overview of how we compared the PIP, PIE, gold-standard and our new experiments. In particular, figure 2 shows the performance of the integration resulting in the PIP and PIE. When tested against the gold-standard, we observed that the ratio of true to false positives (*TP*/*FP*) increases monotonically with L_{cut}, confirming *L* as an appropriate measure of the odds of a real interaction. Conservatively estimated, protein pairs with $L > 600$ have a better than 50% chance of being in the same complex, suggesting $L_{cut} = 600$ as a useful threshold (19). Unless otherwise noted, we use this throughout our analysis. It gives 9,897 predicted interactions from the PIP and 163 from the PIE. In contrast, likelihood ratios derived from single genomic features (e.g., just mRNA co-expression) or from individual interaction experiments (e.g., just the Ho dataset) did not exceed the cutoff when used alone, with *TP*/*FP* values far below 1. This demonstrates that information sources that, taken alone, are only weak predictors of interactions can yield reliable predictions when combined.

The PIP had higher sensitivity than the PIE for comparable *TP*/*FP* ratios (figure 2c). ('Sensitivity' measures coverage and is defined as *TP*/*P*, where *P* is the number of gold-standard positives.) Specifically, the sensitivity of the PIP is ~27% at our cutoff. This may seem low, but compares favorably with the PIE, whose sensitivity was below 1%. This means that we can predict, at comparable error levels, more complex interactions de novo than are present in the high-throughput experimental interaction datasets.

One might ask whether simpler voting procedures can match the performance of more complicated machine-learning methods such as Bayesian networks. To test this, we compared the PIP with a voting procedure where each of the four genomic features contributes an additive vote towards positive classification. We found that the Bayesian network achieved greater sensitivity for comparable *TP*/*FP* ratios (figure 2c) (19).

Figure 3 shows parts of the PIP and PIE graphs and how these compare with the gold-standard and our new experiments. First, to test whether the thresholded PIP was biased towards certain complexes, we looked at the distribution of predictions amongst gold-standard positives (figure 3a); they were roughly equally apportioned amongst the different complexes, suggesting a lack of bias.

While we have thus far treated all interactions as independent, the joint distribution of interactions in the PIs can help identify large complexes: An ideal complex should be a 'clique' in an interaction graph (i.e., a subgraph with $N(N - 1)/2$ links between N proteins). Although this rarely happens in practice, because of incorrect or missing links, large complexes tend to have many interconnections within them, whereas false-positive links to outside proteins tend to occur randomly, without coherent pattern (figure 4).

Figure 3b shows parts of the thresholded PIP that are restricted to proteins with ≥20 links (figure 3b) (23), highlighting large complexes. Some predicted complexes overlap with the gold-standard positives (cytoplasmic ribosome) or the PIE (exosome, RNA polymerase I, 26S proteasome). Comparison with the gold-standard negatives showed where the PIP likely produced false complexes. Many protein associations only appear in the PIP and thus potentially represent new interactions and complexes. An interesting

example is the mitochondrial ribosome; it has appreciable overlap with both gold-standard positives and the PIE, and also contains plausible, newly predicted interactions with three proteins (19).

To further test the predictions in the PIP, we conducted TAP-tagging experiments, in which a protein expressed at its normal intracellular concentration ('bait') is tagged and used to 'pull down' endogenous protein complexes. We picked 98 proteins as TAP-tagging baits. These produced 424 experimental interactions overlapping with the PIP thresholded at L_{cut}=300. (185 of these, in turn, overlapped with gold-standard positives, and 16 with negatives, highlighting the reliability of our experiments.)

Figure 3c shows three examples of the overlap between the PIP and TAP-tagging. We predicted that the putative DEAD-box RNA helicase Dbp3 interacts with three other RNA helicases (Hca4, Mak5 and Dbp7), with proteins implicated in rRNA metabolism (e.g., Nop2, Rrp5, Mak5 and components of RNAPI), and with Nsr1, the yeast homolog of mammalian Nucleolin and a GAR-domain containing protein (24). When Dbp3 was TAP-tagged and purified, we found previously unknown interactions with Nsr1, Hca4 and Nop1 connecting Dbp3 with known rRNA processing proteins. Further purifications using TAP-tagged versions of Mak5, Rrp5, Dbp7, Dbp3, Nsr1, Hca4 and Nop2 doubly verified the physical association.

The nucleosome, a fundamental unit within chromatin, furnishes a second example of overlap. It is composed of 8 histones (2 H2A, 2 H2B, 2 H3, and 2 H4), which can block RNA-polymerase-II progression. This blockage is relieved upon interaction with the FACT complex (also known as SPN or yFACT), which consists of Spt16 and Pob3 in

yeast. Mammalian Pob3 has an HMG domain for interaction with histones; however, yeast Pob3 lacks this. Instead, the HMG protein NHP6 (with two virtually identical isoforms, NHP6A and NHP6B) binds histones (25-27). (It also is known that NHP6 also binds DNA in competition with the nucleosome (28).) Our thresholded PIP and experimental data document a specific interaction between NHP6A and HHF1 (H4), pinpointing the contact between the nucleosome and NHP6 to the H3-H4 heterodimer (HHF1 and HHT1). This is plausible, as NHP6 has been shown not to influence nucleosome reassembly (29), it is unlikely that it binds with the H2A-H2B dimer, which needs to reassociate with the nucleosome after binding FACT.

The replication complex, a third experimental validation of the PIP, assembles and dissembles from transiently interacting sub-complexes (e.g. MCM proteins, ORC and polymerases) throughout the cell-cycle (8, 30). Our predicted and experimentally verified interactions connect it, probably transiently, to another sub-complex, Replication Factor A (RFA, composed of Rfa1, Rfa2 and Rfa3). Specifically, we predicted and verified interactions between RFA and two proteins associated with other replication sub-complexes: Rfa2 with Top2 (a component of the nuclear synaptonemal complex) and Rfa1 with Pri2 (DNA polymerase alpha-primase subunit).

Finally, we predicted and verified by TAP-tagging that two proteins involved in translation elongation (Tef2 and Eft2) interact. This is plausible given that protein elongation is mediated by three factors in yeast: EF-1 alpha (Tef1, Tef2), EF-2 (Eft1, Eft2), and EF-3 (Hef3, Yef3); most other eukaryotes lack EF-3. Previous experimental data suggest an interaction between yeast EF-1 alpha and EF-3 (31). An interaction between EF-1 alpha and EF-2 had not been demonstrated, although this is reasonable

given their similar roles in elongation and their overlapping binding sites on the ribosome (32).

In summary, we have developed a Bayesian approach for integrating weakly predictive genomic features into reliable predictions of protein-protein interactions. Our de novo prediction of complexes replicated interactions found in the gold-standard positives and PIE. In addition, we were able to confirm several of our predictions with new experiments. The accuracy of the PIP was comparable to that of the PIE whilst simultaneously achieving greater coverage.

Our procedure lends itself naturally to the addition of more features, possibly further improving results. We anticipate that protein-protein interactions in organisms other than yeast can be explored in similar ways.

Figure captions

Figure 1: The information sources we integrated and how we compared them with each other. **Part A**: The three different types of data we used: (i) Interaction data from high-throughput experiments. These comprise large-scale two-hybrid screens (Y2H) (1, 2) and in-vivo pull-down experiments (3, 4). (ii) Other genomic features. We considered expression data, biological function of proteins (from Gene Ontology biological process and the MIPS functional catalog) and data about whether proteins are essential (6, 19-22). (iii) Gold-standards of known interactions and non-interacting protein pairs. (Note, the MIPS functional catalog is different than the MIPS complexes catalog used for the gold standard.) **Part B**: Combination of datasets into probabilistic interactomes. **Part C** shows how we compared the probabilistic interactomes with the gold standards and our new experimental data. Numbers next to the arrows indicate which subsequent figures refer to these various comparisons.

Figure 2: Comparison of PIP and PIE with each other and with the individual information sources. **Part A**: the *TP*/*FP* ratio as a function of L_{cut} for the PIP and the individual data from which it was computed. The ratio is computed as follows:

$$TP(L_{cut})/FP(L_{cut}) = \sum_{L>L_{cut}} pos(L) \Big/ \sum_{L>L_{cut}} neg(L)$$

where *pos*(*L*) and *neg*(*L*) are the number of positives and negatives in the gold-standard with a given likelihood ratio *L*. The vertical line indicates our standard threshold L_{cut} = 600. **Part B** shows the same plot as part A, but this time for the PIE. **Part C**: Comparison of *TP*/*FP* ratios between the PIP and PIE. The abscissa represents the sensitivity of the probabilistic interactomes. The gray area indicates the gain of sensitivity of the PIP over the PIE for equal *TP*/*FP* ratios. The arrow shows the difference in sensitivity at *TP*/*FP* = 0.3. At this level, the PIP contains 183,295 protein pairs, of which 6,179 are gold-standard positives (75% sensitivity), whereas the PIE contains 31,511 protein pairs and 1,758 gold standard positives among these (21% sensitivity). The white circles show the performance of a voting procedure in which each of the four genomic features (from which we computed the PIP) contributed an additive vote. There are four possible outcomes in the additive voting procedure, depending on how many datasets contribute a positive vote (19).

Figure 3 shows representations of the thresholded PIP (de novo prediction) compared with different datasets. **Part A** shows the complete set of gold-standard positives and their overlap with the PIP. Here, the PIP (green) covers 27% of the gold standard positives (yellow). **Part B** shows a graph of the largest complexes in the PIP, i.e., only those proteins in the thresholded PIP having ≥20 links. On the left, overlapping gold-standard positives are shown in green, PIE links in blue and overlaps with both the PIE and gold-standard positives in black. On the right, overlapping gold-standard negatives are shown in red. Regions with many red links indicate potential false-positive predictions. **Part C** shows three PIP complexes that we partially verified by TAP-tagging. Each complex contains the proteins linked to a central protein (gray) after thresholding the PIP at $L_{cut} = 300$. Interactions verified by our TAP-tagging are shown in dark blue and PIE links in light blue; black links indicate where TAP-tagging overlapped with PIE links.

Figure 4: *TP*/*FP* for subsets of the thresholded PIP that only include proteins with a minimum number of links. Requiring a minimum number of links isolates large complexes in the thresholded PIP graph (figure 3b). Increasing the minimum number of links raises *TP*/*FP* by preserving the interactions among proteins in large complexes, while filtering out false positive interactions with heterogeneous groups of proteins outside the complexes.

References

1. P. Uetz *et al.*, *Nature* **403**, 623-7. (2000).
2. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **98**, 4569-74. (2001).
3. A. C. Gavin *et al.*, *Nature* **415**, 141-7. (2002).
4. Y. Ho *et al.*, *Nature* **415**, 180-3. (2002).
5. I. Xenarios *et al.*, *Nucleic Acids Res* **30**, 303-5. (2002).
6. H. W. Mewes *et al.*, *Nucleic Acids Res* **30**, 31-4. (2002).
7. G. D. Bader *et al.*, *Nucleic Acids Res* **29**, 242-5. (2001).
8. R. Jansen, D. Greenbaum, M. Gerstein, *Genome Res* **12**, 37-46. (2002).
9. A. Kumar *et al.*, *Genes Dev* **16**, 707-19. (2002).
10. C. von Mering *et al.*, *Nature* **417**, 399-403. (2002).
11. A. M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, *Mol Cell Proteomics* **1**, 349-56. (2002).
12. A. M. Edwards *et al.*, *Trends Genet* **18**, 529-36. (2002).
13. G. D. Bader, C. W. Hogue, *Nat Biotechnol* **20**, 991-7. (2002).
14. A. Kumar, M. Snyder, *Nature* **415**, 123-4. (2002).
15. A. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, D. Eisenberg, *Nature* **402**, 83-6. (1999).
16. M. Gerstein, N. Lan, R. Jansen, *Science* **295**, 284-7. (2002).
17. R. Jansen, N. Lan, J. Qian, M. Gerstein, *J Struct Funct Genomics* **2**, 71-81 (2002).
18. A. Drawid, M. Gerstein, *J Mol Biol* **301**, 1059-75. (2000).
19. Materials and methods are available as supporting material.
20. T. R. Hughes et al., *Cell* **102**, 109-26. (2000).
21. R. J. Cho et al., *Mol Cell* **2**, 65-73. (1998).
22. M. Ashburner et al., *Nat Genet* **25**, 25-9. (2000).
23. http://genecensus.org/intint
24. I. P. Girard et al., *EMBO J* **11**, 673-82. (1992).
25. N. K. Brewster, G. C. Johnston, R. A. Singer, *Mol Cell Biol* **21**, 3491-502. (2001).
26. A. A. Travers, *EMBO Rep* **4**, 131-6. (2003).
27. T. Formosa et al., *Genetics* **162**, 1557-71. (2002).
28. Y. Yu, P. Eriksson, L. T. Bhoite, D. J. Stillman, *Mol Cell Biol* **23**, 1910-21. (2003).
29. R. C. Bash, J. M. Vargason, S. Cornejo, P. S. Ho, D. Lohr, *J Biol Chem* **276**, 861-6. (2001).

30. O. M. Aparicio, D. M. Weinstein, S. P. Bell, *Cell* **91**, 59-69. (1997).

31. M. Anand, K. Chakraburtty, M. J. Marton, A. G. Hinnebusch, T. G. Kinzy, *J Biol Chem* **278**, 6985-91. (2003).

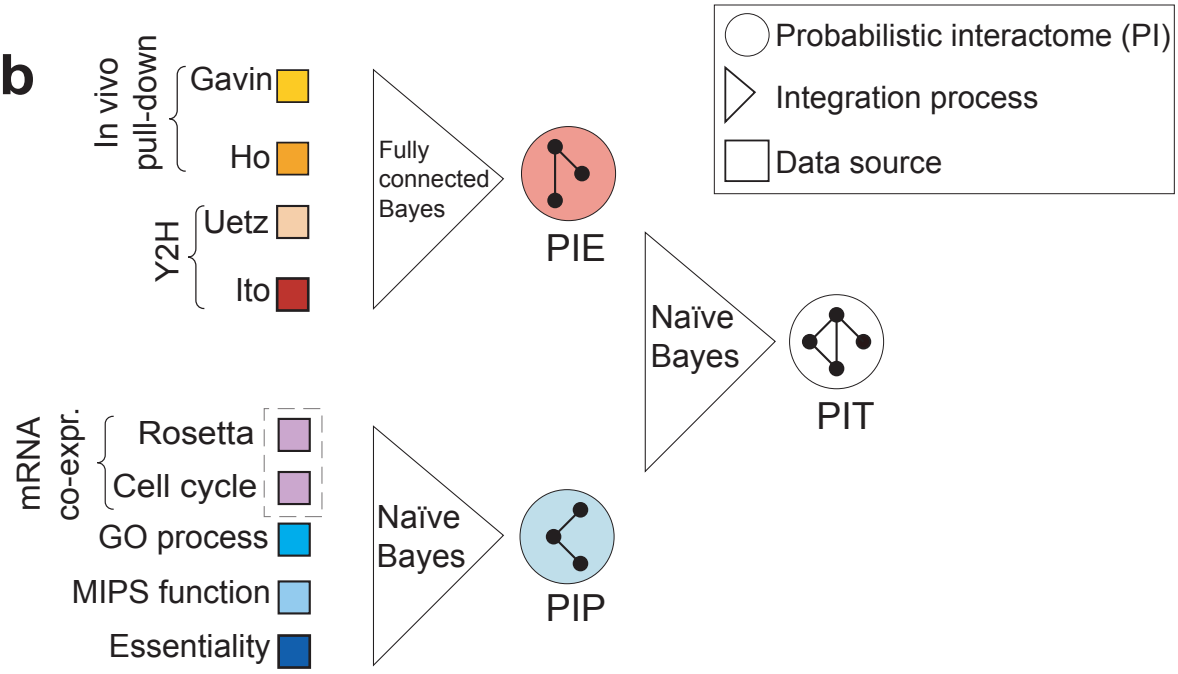
32. O. Kovalchuke, R. Kambampati, E. Pladies, K. Chakraburtty, *Eur J Biochem* **258**, 986-93. (1998).

Figure 1

a

Data type	Dataset		# protein pairs	Used for ...
Experimental interaction data	In-vivo pull-down	Gavin et al.	31,304	Integration of experimental interaction data (PIE)
		Ho et al.	25,333	
	Yeast two-hybrid	Uetz et al.	981	
		Ito et al.	4,393	
Other genomic features	Expression	Rosetta compendium	19,334,806	De novo prediction (PIP)
		Cell cycle	17,467,005	
	Biological function	GO biological process	3,146,286	
		MIPS function	6,161,805	
	Essentiality		8,130,528	
Gold standards	Positives	Proteins in the same MIPS complex	8,250	Training & testing
	Negatives	Proteins separated by localization	2,708,746	

b



c

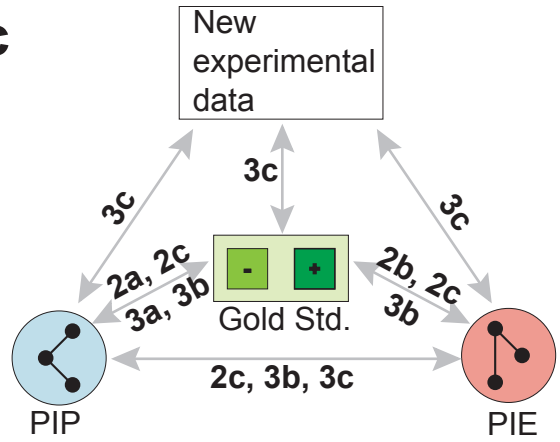


Figure 2

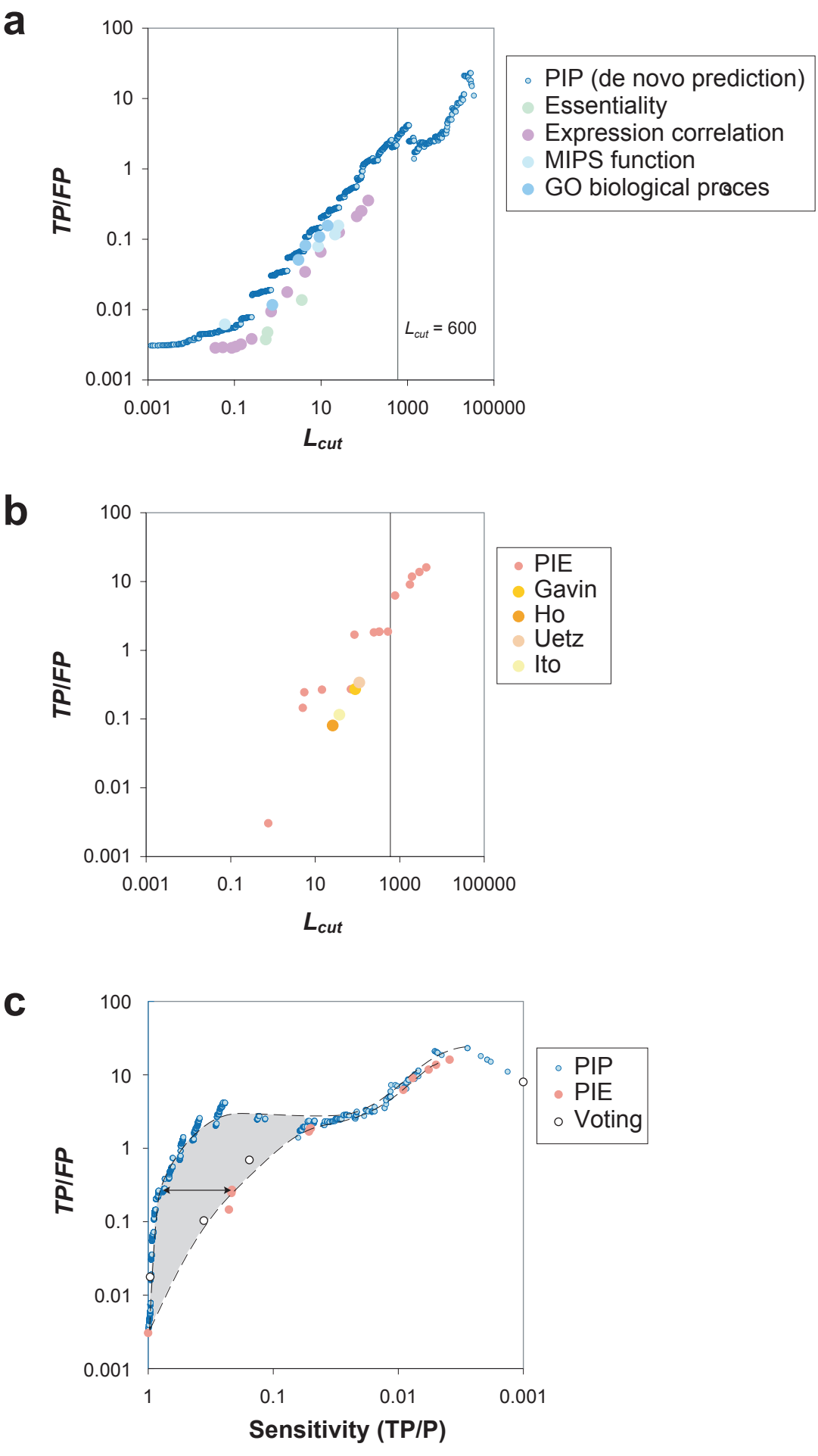

a
100
10
1
0.1
0.01
0.001
TP/FP
0.001
0.1
10
1000
100000
L_{cut}
L_{cut} = 600
PIP (de novo prediction)
Essentiality
Expression correlation
MIPS function
GO biological proces
b
100
10
1
0.1
0.01
0.001
TP/FP
0.001
0.1
10
1000
100000
L_{cut}
PIE
Gavin
Ho
Uetz
Ito
c
100
10
1
0.1
0.01
0.001
TP/FP
1
0.1
0.01
0.001
Sensitivity (TP/P)
PIP
PIE
Voting

Figure 3a

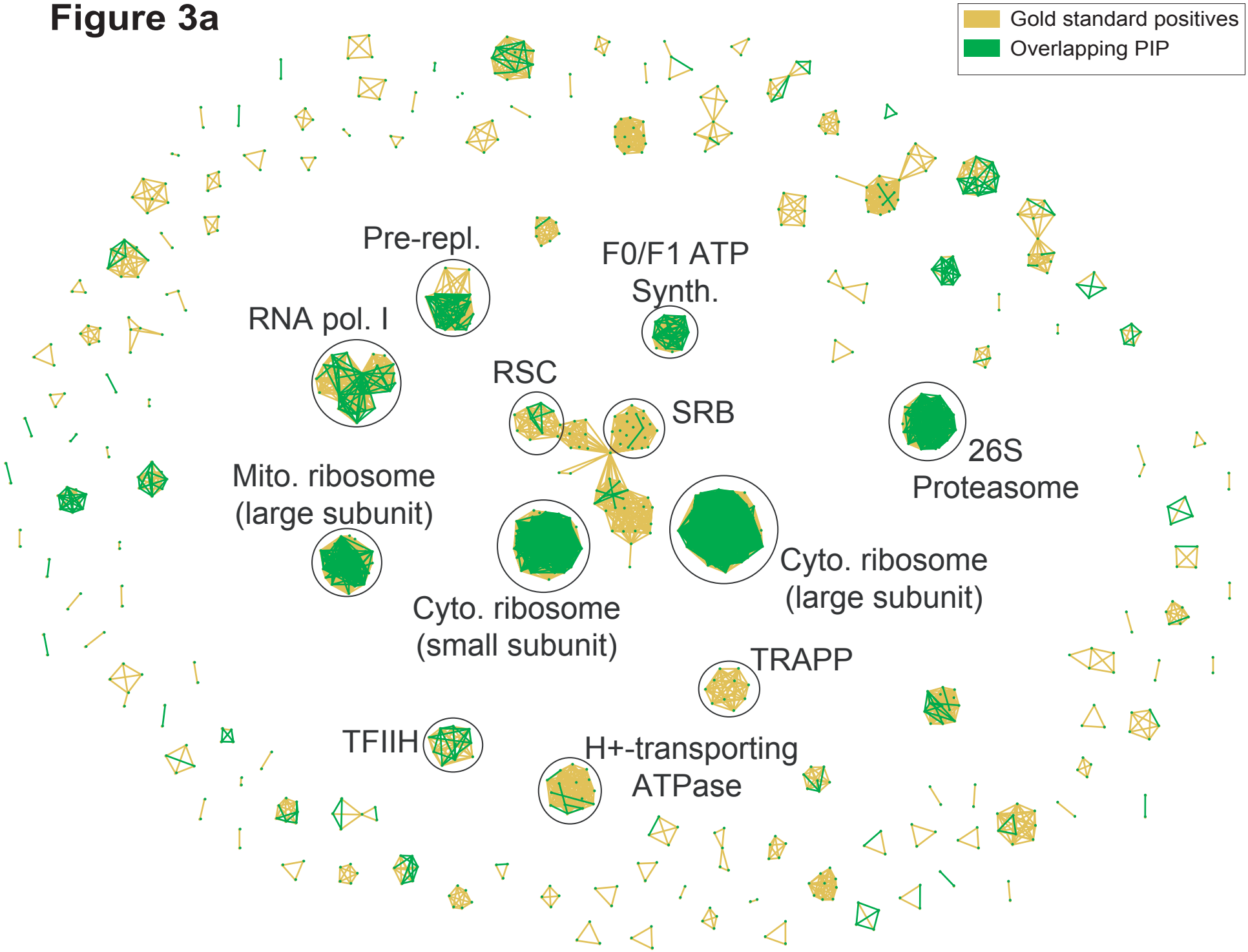

Gold standard positives
Overlapping PIP
Pre-repl.
F0/F1 ATP Synth.
RNA pol. I
RSC
SRB
26S Proteasome
Mito. ribosome (large subunit)
Cyto. ribosome (large subunit)
Cyto. ribosome (small subunit)
TRAPP
TFIIH
H+-transporting ATPase

Figure 3b

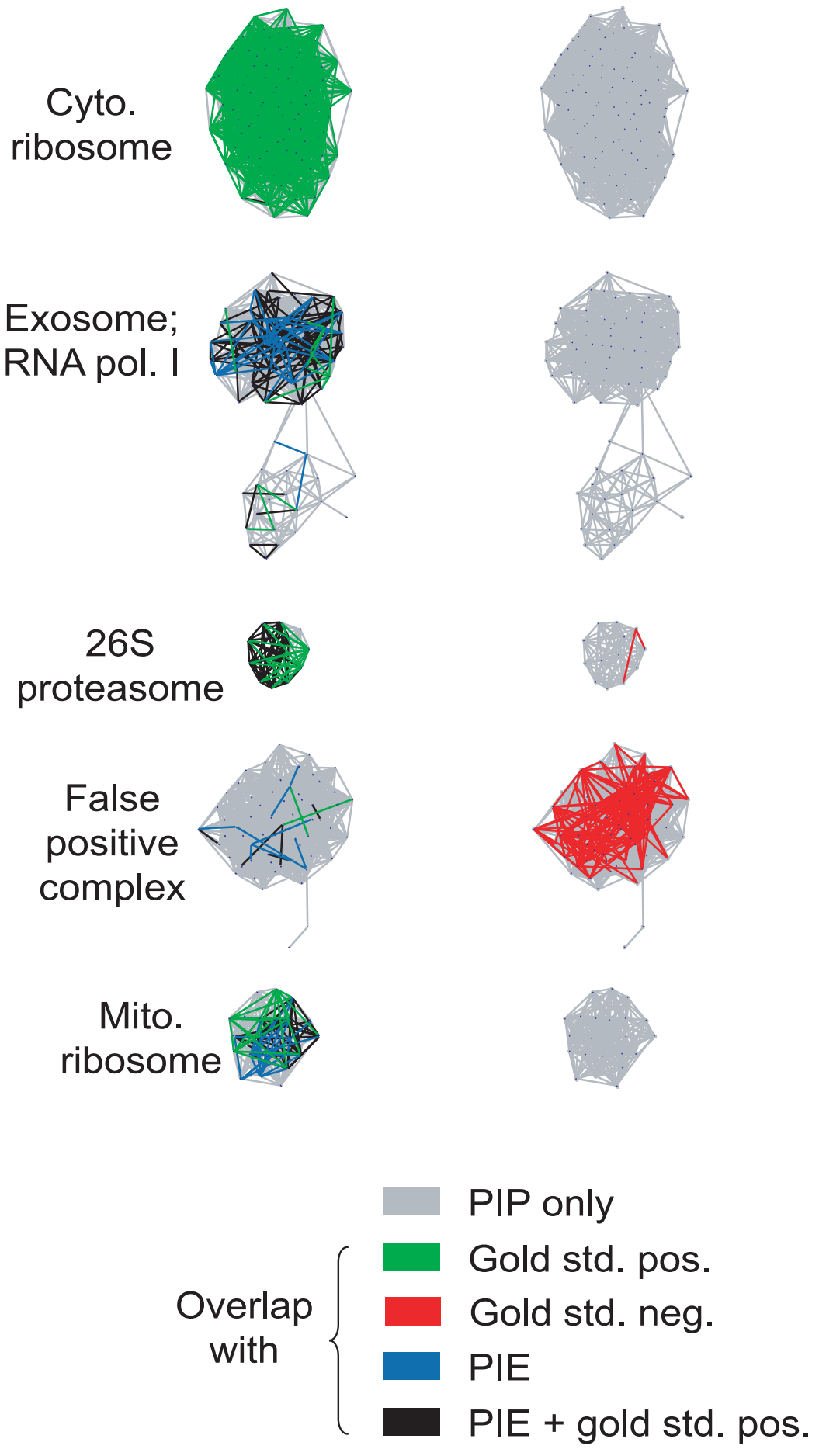

Cyto. ribosome
Exosome; RNA pol. I
26S proteasome
False positive complex
Mito. ribosome
Overlap with
PIP only
Gold std. pos.
Gold std. neg.
PIE
PIE + gold std. pos.

Figure 3c

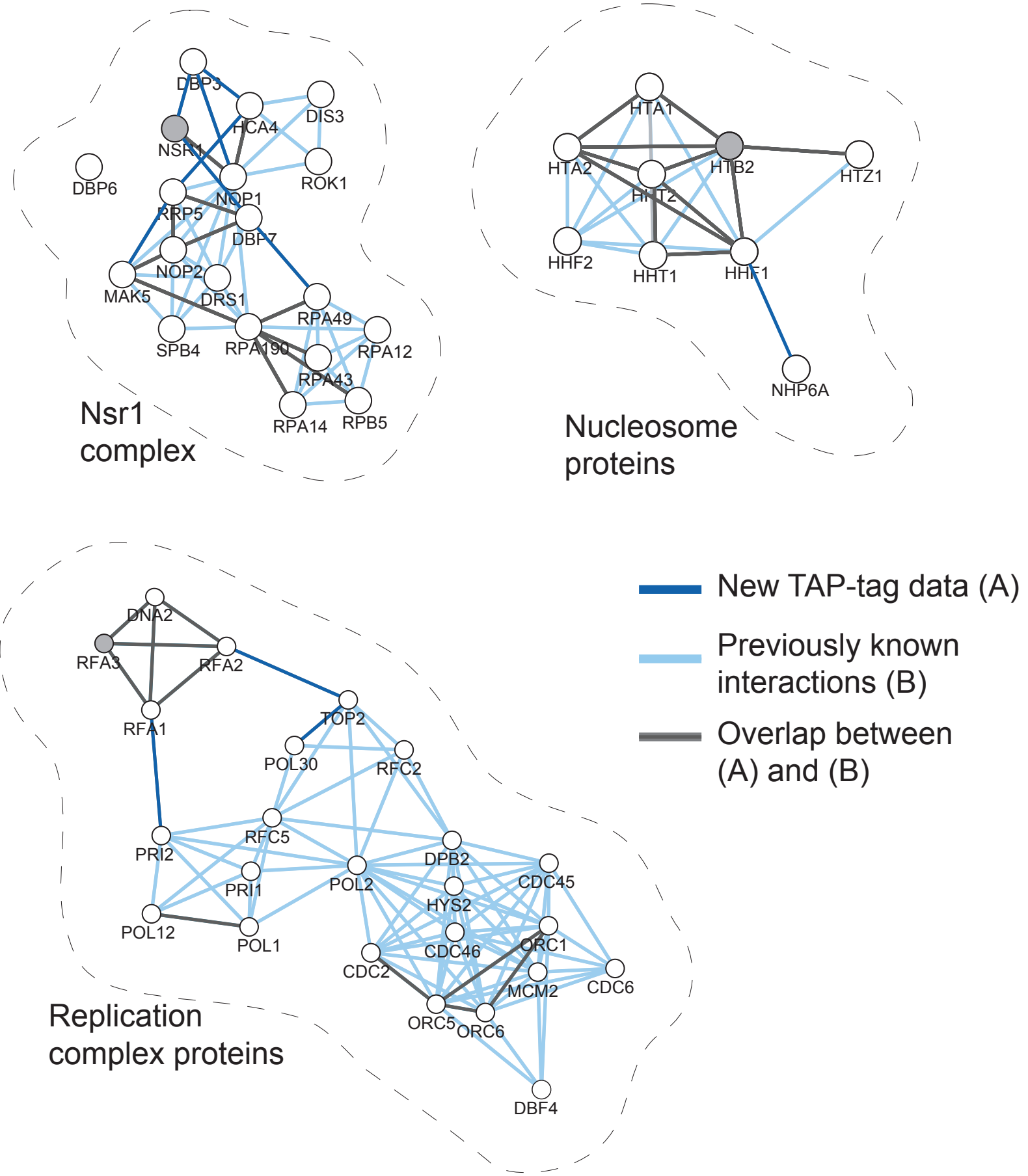
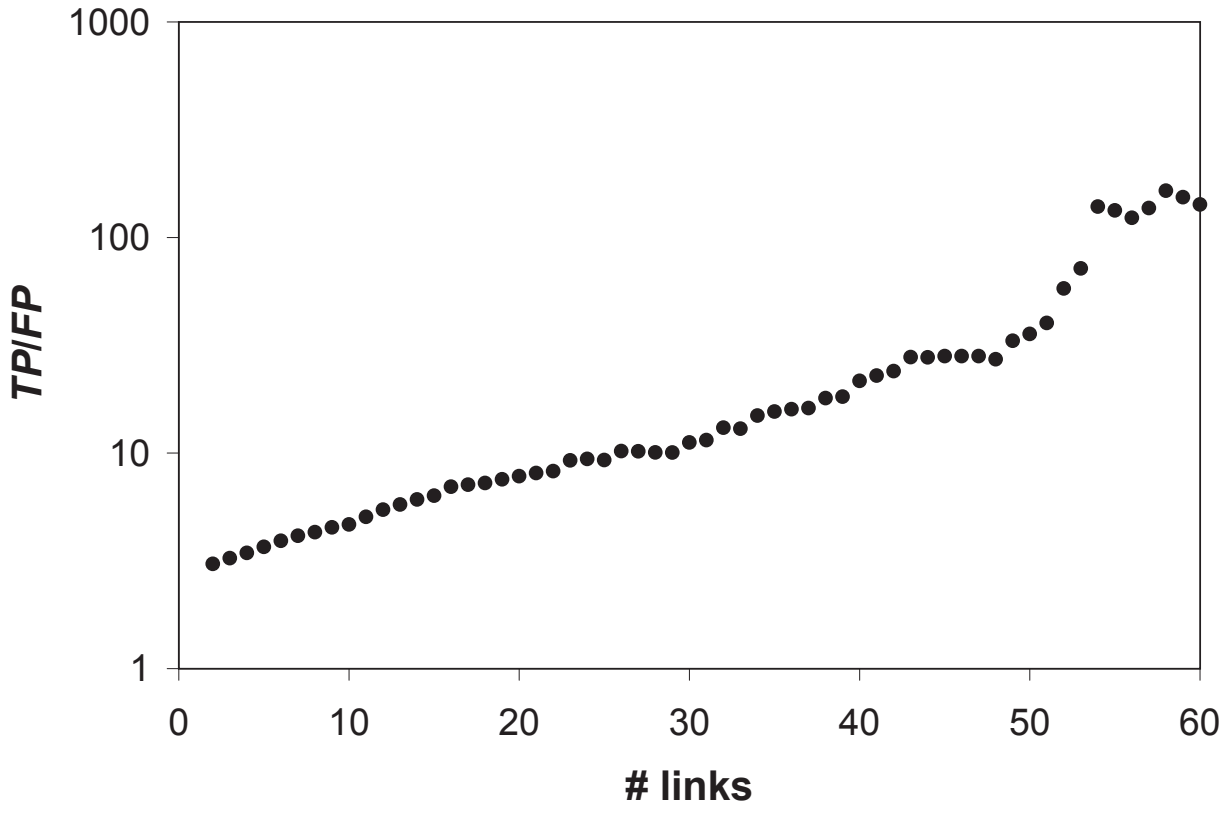

DBP3
DIS3
HCA4
NSR1
DBP6
ROK1
NOP1
RRP5
DBP7
NOP2
MAK5
DRS1
RPA49
SPB4
RPA190
RPA12
RPA43
RPA14
RPB5
Nsr1 complex
HTA1
HTA2
HTB2
HTZ1
HHT2
HHF2
HHT1
HHF1
NHP6A
Nucleosome proteins
DNA2
RFA3
RFA2
RFA1
TOP2
POL30
RFC2
RFC5
PRI2
DPB2
PRI1
POL2
CDC45
HYS2
POL12
POL1
CDC46
ORC1
CDC2
MCM2
CDC6
ORC5
ORC6
DBF4
Replication complex proteins
New TAP-tag data (A)
Previously known interactions (B)
Overlap between (A) and (B)

Figure 4



Supplementary online material

Jansen et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data

Materials and methods

Datasets

Genomic features for computation of the PIP

mRNA expression

We use publicly available expression data, in particular, a time course of expression fluctuations during the yeast cell cycle and the Rosetta compendium, consisting of the expression profiles of 300 deletion mutants and cells under chemical treatments (S1, S2). This data can be used for the prediction of protein-protein interaction because proteins in the same complex are often co-expressed (S3-S6). We computed the Pearson correlation for each protein pair for both the Rosetta and cell cycle datasets. For predicting protein-protein interactions, the Rosetta correlation and the cell cycle correlation represent strongly correlated evidence (see discussion below). We circumvented this problem by computing the first principal component of the vector of the two correlations. Then we used this first principal component as one independent source of evidence for the protein-protein interaction prediction. This first principal component is a stronger predictor of protein-protein interactions than either of the two expression correlation datasets by themselves. We divided this first principal component of expression correlations into 19 bins. For each bin we assessed its overlap with the gold-standard (table S1).

Biological function

Interacting proteins often function in the same biological process (S7-S9). This means that two proteins acting in the same biological process are more likely to interact than two proteins involved in different processes.

We collected information from two catalogs of functional information about proteins, the MIPS functional catalog (S10) – which is separate from the MIPS complexes catalog -- and the data on biological processes from Gene Ontology (GO) (S11). We used the following procedure to quantify functional similarity between two proteins: We first consider which set of functional classes two proteins share, given one of the functional classification systems. Then we count how many of the ~18 million protein pairs in yeast share the exact same functional classes as well (yielding a count between 1 and ~18 million). In general, the smaller this count, the more similar and specific is the functional description of the two proteins, while large counts indicate a very non-specific functional relationship between the proteins. We found that low counts (i.e., high functional similarity) correlate with a higher chance of two proteins being in the same complex (table S1).

Essentiality

We considered whether proteins are essential or non-essential (S10). It should be more likely that both of two proteins in a complex are essential or non-essential, but not a mixture of these two attributes. This is because a deletion mutant of either one protein should by and large produce the same phenotype: They both impair the function of the same complex. Indeed we find such a relationship supported by the data (table S1).

Finally, in principle, our approach could have been extended to a number of other features related to interactions (e.g. phylogenetic occurrence, gene fusions, gene neighborhood) (S12-19).

Gold-standard

For the validation and prediction of protein complexes, we need to have reference datasets that serve as gold-standards of positives (proteins that are in the same complex) and negatives (proteins that do not interact).

For reliable data about existing protein complexes we took the MIPS complexes catalog as a reference in its version from November 2001 (S10). It consists of a list of known protein complexes based on data collected from the biomedical literature (most of these are derived from small-scale studies in contrast to the high-throughput experimental interaction data (S7, S20-S24). We only considered classes that contain single complexes. For instance, the MIPS class 'translation complexes' contains the subclasses 'mitochondrial ribosome', the 'cytoplasmic ribosome' and a number of other subclasses related to translation-related complexes; we only considered pairs among proteins in those subclasses as positives. Overall, this yielded a filtered set of 8250 protein pairs that are within the same complex.

There is no direct information about which proteins do not interact. However, protein localization data provides indirect information if we assume that proteins in different compartments do not to interact. We compiled a list of 2,691,903 protein pairs in

different compartments from the current yeast localization data. In compiling this list, we attributed proteins to one of five compartments as has been done previously (S25-S27).

Ideally, the positive gold-standard and the negative gold-standard should be mutually exclusive. In practice, this is not precisely the case. Of the 8,250 protein pairs in the positive gold-standard, the subcellular localization is known for both proteins in 6,133 cases. Of these 6,133 protein pairs, 124 intersect with the set of gold-standard negatives (representing a fraction of 2% = 124/6,133). This is very small compared to the randomly expected size of the intersection (65%), which can be computed by randomly shuffling the subcellular localization of the proteins in the positives set. Thus, although the gold-standard sets are not ideal, they provide a good practical approximation.

One reason for the small intersection between the gold-standards positives and negatives is that some proteins change their subcellular localization. Several of the 124 protein pairs in the intersection are in transcription-factor complexes. This is plausible, given that transcription factors must be translated in the cytoplasm before they are transported to the nucleus; thus, they are at least transiently located in the cytoplasm.

Computational methods (Bayesian networks)

The need for integrating data from a variety of sources has been emphasized recently in computational biology (S26, S28-S30). Bayesian networks are particularly suitable for the task of combining evidence from heterogeneous data sources (S31).

Bayesian networks are a representation of the joint probability distribution among multiple variables (which could be datasets or information sources). Formally, they can

be described as follows (S32, S33): We define as 'positive' a pair of proteins that are in the same complex. Given the number of positives among the total number of protein pairs, the 'prior' odds of finding a positive are:

$$O_{prior} = \frac{P(pos)}{P(neg)} = \frac{P(pos)}{1 - P(pos)}$$

In contrast, the 'posterior' odds are the odds of finding a positive after we consider N datasets with values $f_1 \ldots f_N$:

$$O_{post} = \frac{P(pos \mid f_1 ... f_N)}{P(neg \mid f_1 ... f_N)}$$

(The terms 'prior' and 'posterior' refer to the situation before and after knowing the information in the N datasets.)

The likelihood ratio L defined as

$$L(f_1 ... f_N) = \frac{P(f_1 ... f_N \mid pos)}{P(f_1 ... f_N \mid neg)}$$

relates prior and posterior odds according to Bayes' rule:

$$O_{post} = L(f_1 ... f_N) O_{prior}$$

In the special case that the N features are conditionally independent (i.e., they provide uncorrelated evidence), the Bayesian network is a so-called 'naïve' network, and L can be simplified to:

$$L(f_1 ... f_N) = \prod_{i=1}^{N} L(f_i) = \prod_{i=1}^{N} \frac{P(f_i \mid pos)}{P(f_i \mid neg)}$$

L can be computed from contingency tables relating positive and negative examples with the N features (by binning the feature values $f_1 \ldots f_N$ into discrete intervals, see table S1 and table S2). Determining the prior odds O_{prior} is somewhat arbitrary in that it requires an assumption about the number of positives. However, based on previous estimates (S34-S37) we think that 30,000 positives is a conservative lower bound for the number of positives (i.e., pairs of proteins that are in the same complex). Given that there are approximately 18 million protein pairs in total, the prior odds would then be about 1 in 600. With $L > 600$ we would thus achieve $O_{post} > 1$.

In the naïve Bayesian network the assumption is that the different sources of evidence (i.e., our datasets with information about protein complexes) are conditionally independent. Conditional independence means that the information in the N datasets is independent given that a protein pair is either positive or negative. We have tested this criterion for the different datasets using scatterplots and have found that they are largely conditionally uncorrelated (S38). The only exceptions are the two datasets of expression correlations. (We described above how we circumvented this problem.)

Surprisingly, the two datasets of functional similarity, derived from the MIPS and GO functional catalogs, were also for the most part conditionally independent. We would have expected that the quantification of functional similarities would yield similar results for both catalogs; this, however, was not the case, such that we can basically treat each data source as conditionally independent evidence.

The PIE and PIP data turned out to be conditionally independent, such that they could be combined in a naïve Bayesian fashion to form the PIT.

From a computational standpoint, the naïve Bayesian network is easier than the fully connected network. The more conditional independence relationships there are between variables, the easier it is generally to compute the parameters in a Bayesian network.

Experimental methods (TAP-tagging)

Frozen cell pellets from 3 L yeast cultures grown in YPD medium to an OD_{600} of 1.0-1.5 were broken with dry ice in a coffee grinder. Tagged complexes were purified on IgG and calmodulin columns from extracts as previously described (S39), except that the buffers for the calmodulin column contained no detergent and the elution buffer for the calmodulin column contained 100 mM ammonium bicarbonate in place of 100 mM NaCl. The purified proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) on gels containing 10% polyacrylamide and the proteins were visualized by silver staining. The protein bands were reduced, alkylated and subjected to in-gel tryptic digestion. Peptide samples were then spotted onto a target plate with a matrix of α-cyano-4-hydroxycinnamic acid (Fluka). MALDI TOF mass spectrometry analysis was conducted utilizing a Reflex IV (Bruker Daltonics, Billerica, MA) instrument in positive ion reflectron mode. For LC-MS/MS, a portion of the purified protein preparation was concentrated by evaporation and resuspended in 100mM NH_4HCO_3/1mM $CaCl_2$ buffer, pH 8.5 and digested overnight at 37°C with 2mL of immobilized Poros trypsin beads (PerSeptive). The entire digest was fractionated as described (S40) on a 7.5 cm (100 um ID) reverse phase C18 capillary column attached in-line to a ThermoFinnigan LCQ-Deca ion trap mass spectrometer by ramping a linear gradient from 2 to 60% solvent B in 90 min. Solvent A consisted of 5% acetonitrile,

0.5% acetic acid and 0.02% HFBA and solvent B consisted of 80:20 acetonitrile/water containing 0.5% acetic acid and 0.02% HFBA. The flow rate at the tip of the needle was set to 300 nL/min by programming the HPLC pump and using a split line. The mass spectrometer cycled through four scans as the gradient progressed. The first was a full mass scan followed by successive tandem mass scans of the three most intense ions. A dynamic exclusion list was used to limit collection of tandem mass spectra for peptides that eluted over a long period of time. All tandem mass spectra were searched using the SEQUEST computer algorithm against a complete yeast protein sequence database (6/2000). Each high-scoring peptide sequence was evaluated using STATQUEST (S41) with the corresponding tandem mass spectrum to determine the probability of each match.

Comparison of Bayesian networks with voting

A simpler integration method than Bayesian networks would be a voting procedure, in which each dataset contributes an additive vote towards classification of a protein pair as positive. One can compute likelihood and *TP*/*FP* ratios depending on how many datasets agree. One extreme of this procedure is to accept every protein pair as positive that has at least one vote (i.e., the union of all datasets, OR rule), whereas the other extreme is to limit positives to only those pairs that have votes from all datasets (i.e., the intersection, AND rule). Both approaches have previously been applied to protein-protein interaction data (S7, S42, S43).

Comparison of voting with the PIP

One limitation of a voting procedure is that it requires the input datasets to be binary in format, meaning that a protein interaction is either 'present' or 'absent' in a dataset. 'Present' can then be counted as a positive vote in a voting procedure.

The situation is different for the datasets that we used for our *de novo* prediction (PIP). For instance, the mRNA expression dataset contains expression correlations of protein pairs that range on a continuous scale from –1.0 to +1.0. In order to transform these data into binary format, it is necessary to first set an arbitrary cutoff value (for instance, such that correlations greater than 0.7 can be counted as a positive vote). Similarly, the GO process dataset and the MIPS function dataset are not binary in that they contain integer values ranging between 1 and ~18,000,000, representing the similarity of function; in the essentiality dataset, there are three different values. We tried different combinations of cutoffs and then compared the results with the performance of the Bayesian network (figure S1).

Loss of information in the voting procedure

The setting of cutoffs to transform the datasets used in the *de novo* prediction into a binary format naturally involves a loss of information. Another complication of the voting procedure is that different cutoffs change the results of the voting procedure, but there is no immediately obvious procedure for setting cutoffs in an optimal fashion. Given that the Bayesian network can take into account the full information contained in the input datasets, it is not surprising that it exhibits a better prediction performance than

the voting procedure. The Bayesian network can accommodate datasets of multiple formats, such as those containing continuous variables and other non-binary formats.

Treatment of data sources with different reliability

An additional advantage of the Bayesian network over the voting procedure is that it is inherently probabilistic in nature. This lets it easily handle data sources of unequal reliability, whereas simple voting can only give equal weighting to each source.

Comparison of voting with the PIE

Since the four experimental protein-protein interaction datasets that make up the PIE have binary format, it is very straightforward to apply a voting procedure to them.

The advantages of the Bayesian network are less obvious in this situation, although it provides a more fine-grained way of combining the data and tends to have a slightly higher sensitivity for a given level of accuracy than the voting procedure (figure S2). This is because the different subsets can overlap quite differently with the positives and negatives of the gold standards, even if the number of datasets agreeing with each other is the same. For instance, among the subset of protein pairs that are present in the two large-scale two-hybrid datasets (S7, S20-S22), but not the two in-vivo pull-down datasets (S23, S24), 6 overlap with the positives and 23 with the negatives in the gold-standards. Conversely, for the subset of proteins that are only present in two pull-down datasets, the corresponding numbers are 337 positives and 209 negatives in the gold-standards (table S2).

In summary, the Bayesian network performed slightly better than voting procedure with regard to the PIE. In the de novo prediction (PIP), the accuracy of the Bayesian network was about an order of magnitude higher than that of the voting procedure. Since the Bayesian network can take into account more of the information that is contained in the input datasets than the voting procedure, the advantages of the Bayesian network are more evident in a situation where the input datasets are non-binary.

Mitochondrial ribosome

One of the large complexes found in the thresholded PIP is the mitochondrial ribosome (figure 3b). Figure S3 shows this complex in more detail. The de novo prediction overlapped with data from both the gold-standard and the PIE, but, in addition, the de novo prediction added three proteins to this complex (MEF1, YNL081C, and YGL068W). MEF1 is a translation elongation factor and should thus be transiently associated with the mitochondrial ribosome (S44). For the other two proteins there is no direct experimental evidence of their function. However, the sequence of YNL081C is 40% identical to a 30S ribosomal subunit in *Thermus thermophilus* (S45, S46) and the sequence of YGL068W is 52% identical to the L7/L12 ribosomal protein in *E. coli* (S47). Therefore, our predictions for YGL068W and YNL081C seem to provide another level of evidence for annotation of these proteins as mitochondrial ribosomal proteins.

Figures and tables

Figure S1: Comparison of voting and Bayesian network applied to the PIP

Part A: Schematized for simplicity. **Part B**: Actual data (in the same framework for comparison). We measure prediction performance in two ways: first, in terms of sensitivity, represented on the abscissa -- the fraction of true positives (*TP*) among the positives in the gold standard reference ($P = TP + FN$) – and, second, the ratio of true to false positives (TP/FP), represented on the ordinate. The sensitivity is a measure of coverage and the TP/FP ratio a measure of accuracy of the prediction methods.

Part A: The black dots represent the outcomes of a particular voting procedure, while the solid line represents the results of the Bayesian network. Note that the voting procedure leads to four discrete outcomes. This is because the input datasets need to be transformed into a binary format for the voting procedure ("positive vote" or "no vote"). The Bayesian network does not require such a coarse transformation of the input datasets, but can take more of the information into account, leading to a more continuous set of results.

Part B: Since different cutoffs affect the results of the voting procedure, we computed the results for a range of different cutoff sets (a 'cutoff set' contains the four cutoffs applied to each of the four input datasets). Each cutoff set produces four different outcomes that are represented by the same color. Solid lines enclose regions of the same number of votes. Gray dots represent the Bayesian network results. The Bayesian network has a larger set of possible outcomes (reflecting the fact that it takes into account

more of the input information), leading to improved prediction. For instance, at 50% sensitivity, the Bayesian network has a *TP*/*FP* ratio that is about an order of magnitude greater than that of a voting procedure.

Figure S1

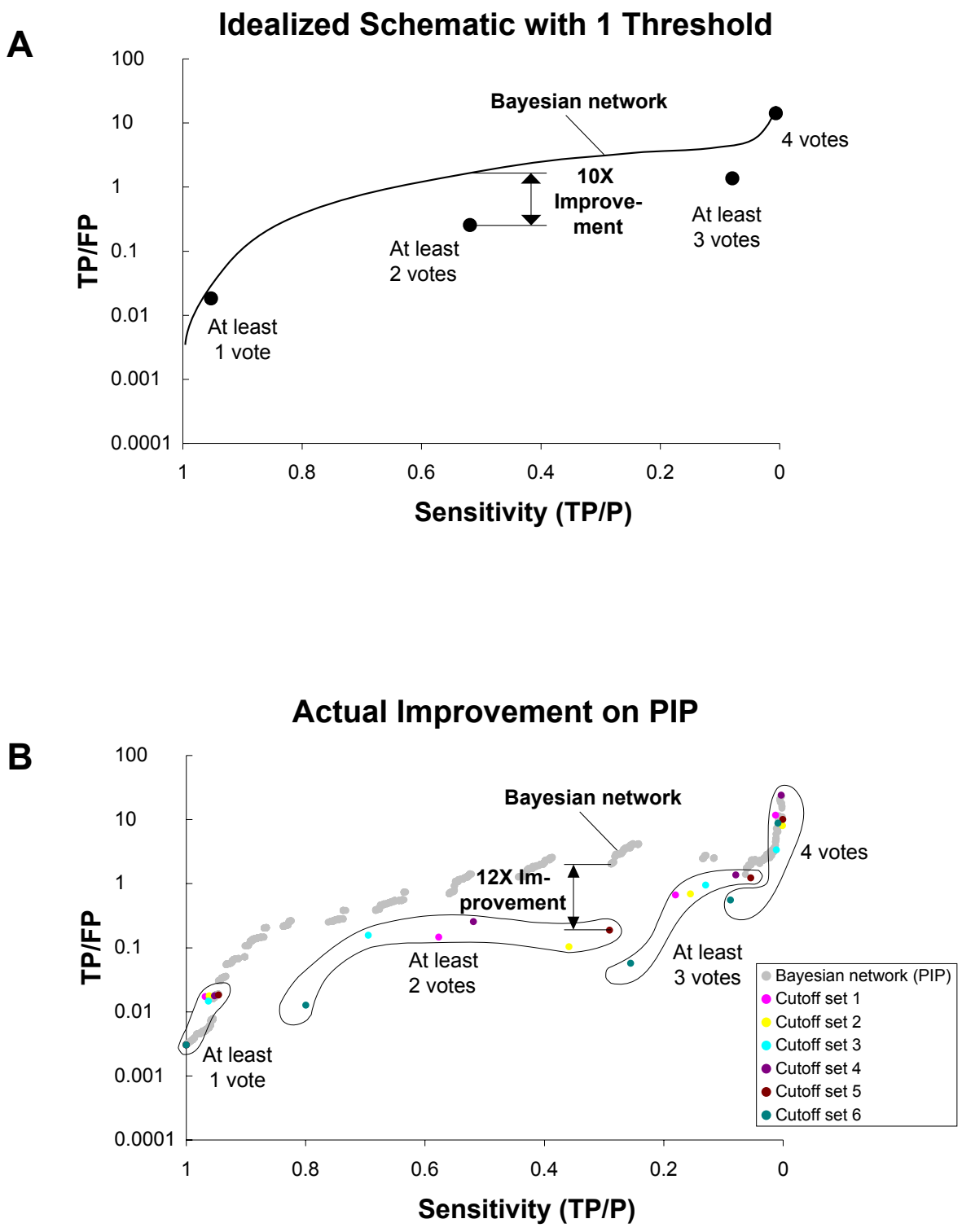
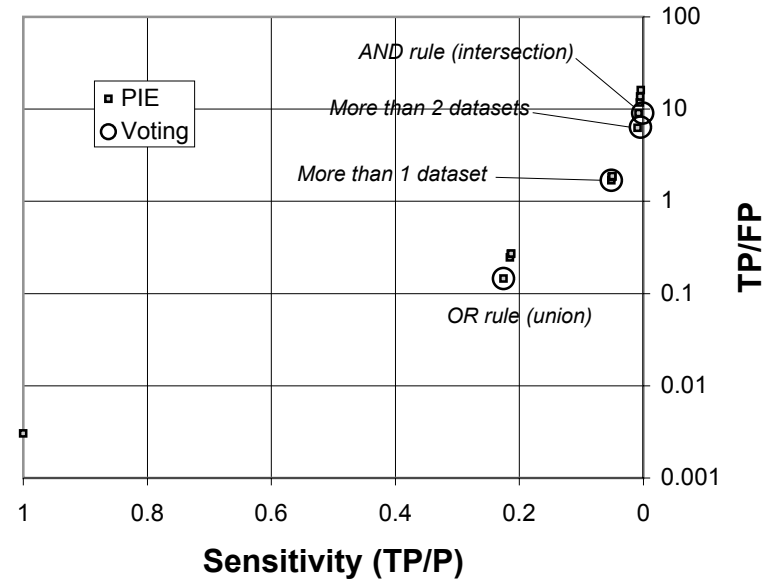

A
Idealized Schematic with 1 Threshold
100
10
1
0.1
0.01
0.001
0.0001
TP/FP
Bayesian network
4 votes
10X
Improve-
ment
At least
3 votes
At least
2 votes
At least
1 vote
1
0.8
0.6
0.4
0.2
0
Sensitivity (TP/P)
B
Actual Improvement on PIP
100
10
1
0.1
0.01
0.001
0.0001
TP/FP
Bayesian network
4 votes
12X Im-
provement
At least
3 votes
At least
2 votes
At least
1 vote
Bayesian network (PIP)
Cutoff set 1
Cutoff set 2
Cutoff set 3
Cutoff set 4
Cutoff set 5
Cutoff set 6
1
0.8
0.6
0.4
0.2
0
Sensitivity (TP/P)

Figure S2: Comparison of voting and Bayesian network applied to PIE

Sensitivity and *TP*/*FP* ratio of the voting procedure and those of the fully connected Bayesian network we used for computing the PIE. The simplest case of a voting procedure is the 'OR' rule, in which a protein pair needs to be in only dataset to be classified as positive. The most stringent case is the 'AND' rule, in which a protein pair needs to be in all datasets to be classified as a positive.

Figure S2


PIE
Voting
AND rule (intersection)
More than 2 datasets
More than 1 dataset
OR rule (union)
100
10
1
0.1
0.01
0.001
TP/FP
1
0.8
0.6
0.4
0.2
0
Sensitivity (TP/P)

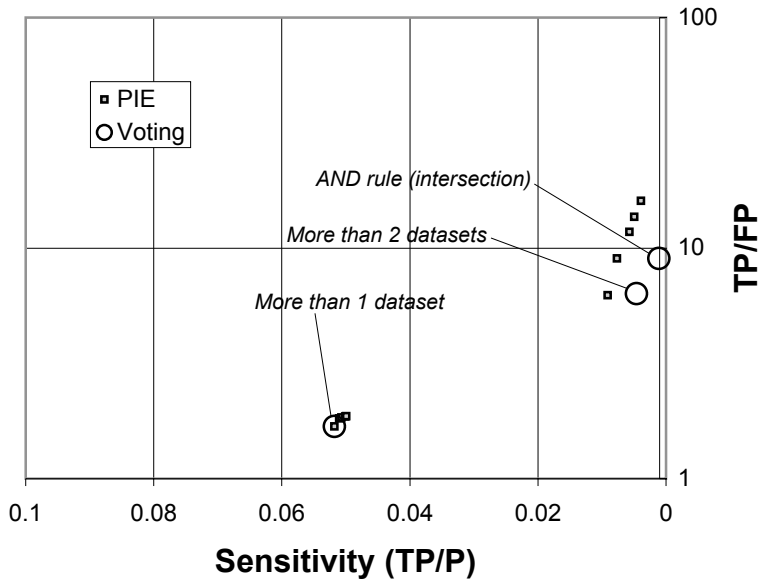
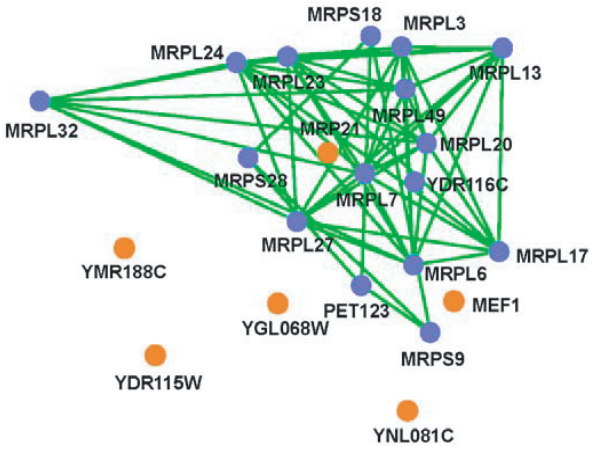

PIE
Voting
AND rule (intersection)
More than 2 datasets
More than 1 dataset
100
10
1
TP/FP
0.1
0.08
0.06
0.04
0.02
0
Sensitivity (TP/P)

Figure S3: Mitochondrial ribosome

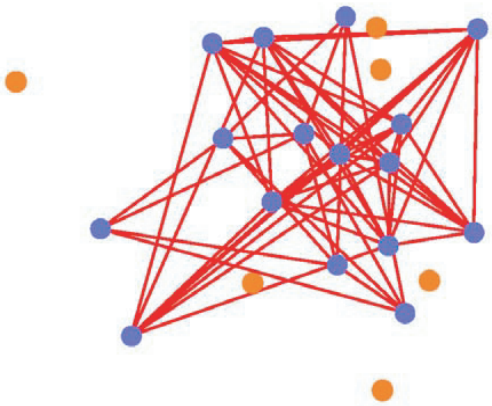
Proteins in the mitochondrial ribosome overlapping with: (i) the gold standard positives (MIPS complexes catalog), (ii) the PIE and (iii) the PIP, which encompasses the data from both (i) and (ii). Blue nodes represent proteins present in each of the three sets, whereas the three orange proteins appeared only in the PIP.

Figure S3

i) Gold-standard positives



ii) Links in PIE



iii) thresholded PIP links

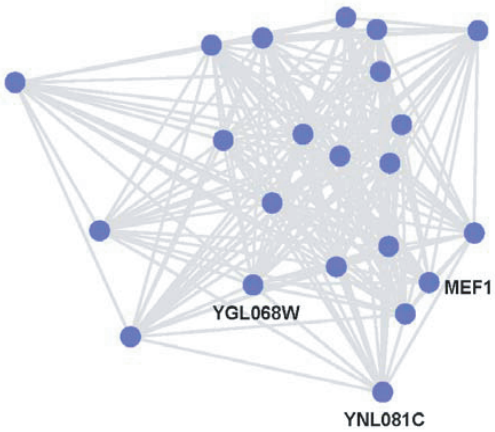


Table S1: Parameters of the naïve Bayesian network (PIP)

The first column describes the genomic feature. Protein pairs in the essentiality data can take on three discrete values (EE, both essential; NN, both non-essential; and NE, one essential and one not), while the values for mRNA expression correlations range on a continuous scale between –1.0 and +1.0; functional similarity counts are integers between 1 and ~18 million. We binned the mRNA expression correlation values into 19 bins and the functional similarity counts into 5 bins. The second column gives the number of protein pairs with a particular feature value (i.e., 'EE') drawn from the whole yeast interactome (~18M pairs). Columns 'pos' and 'neg' give the overlap of these pairs with the 8,250 gold-standard positives and 2,708,746 gold-standard negatives. The final three columns give the conditional probabilities and the likelihood ratio *L*.

Table S1

Essentiality		# protein pairs	Gold-standard overlap		P(Ess|pos)	P(Ess|neg)	L
			pos	*neg*			
Values	EE	301,088	1,114	81,924	5.18E-01	1.43E-01	3.63
	NE	2,481,701	624	285,487	2.90E-01	4.98E-01	0.58
	NN	4,771,865	412	206,313	1.92E-01	3.60E-01	0.53
Sum		7,554,654	2,150	573,724	1.00E+00	1.00E+00	1.00

Expression correlation		# protein pairs	Gold standard overlap		P(exp|pos)	P(exp|neg)	L
			pos	*neg*			
Values	0.9	617	16	45	2.10E-03	1.68E-05	124.93
	0.8	4,127	137	563	1.80E-02	2.10E-04	85.50
	0.7	14,979	530	2,117	6.96E-02	7.91E-04	87.97
	0.6	36,145	1,073	5,597	1.41E-01	2.09E-03	67.36
	0.5	81,102	1,089	14,459	1.43E-01	5.40E-03	26.46
	0.4	189,369	993	35,350	1.30E-01	1.32E-02	9.87
	0.3	444,757	1,028	83,483	1.35E-01	3.12E-02	4.33
	0.2	1,016,105	870	183,356	1.14E-01	6.85E-02	1.67
	0.1	2,205,895	739	368,469	9.71E-02	1.38E-01	0.70
	0	8,118,256	894	1,244,477	1.17E-01	4.65E-01	0.25
	-0.1	2,345,009	164	408,562	2.15E-02	1.53E-01	0.14
	-0.2	1,038,181	63	203,663	8.27E-03	7.61E-02	0.11
	-0.3	399,554	13	84,957	1.71E-03	3.18E-02	0.05
	-0.4	131,361	3	28,870	3.94E-04	1.08E-02	0.04
	-0.5	40,759	2	8,091	2.63E-04	3.02E-03	0.09
	-0.6	15,289	-	2,134	0.00E+00	7.98E-04	0.00
	-0.7	6,795	-	807	0.00E+00	3.02E-04	0.00
	-0.8	1,886	-	261	0.00E+00	9.76E-05	0.00
	-0.9	55	-	12	0.00E+00	4.49E-06	0.00
Sum		16,090,241	7,614	2,675,273	1.00E+00	1.00E+00	1.00

MIPS function similarity		# protein pairs	Gold standard overlap		P(MIPS|pos)	P(MIPS|neg)	L
			pos	*neg*			
Values	1 -- 9	6,584	171	1,094	2.12E-02	8.33E-04	25.50
	10 -- 99	25,823	584	4,229	7.25E-02	3.22E-03	22.53
	100 -- 1000	88,548	688	13,011	8.55E-02	9.91E-03	8.63
	1000 -- 10000	255,096	6,146	47,126	7.63E-01	3.59E-02	21.28
	10000 -- Inf	5,785,754	462	1,248,119	5.74E-02	9.50E-01	0.06
Sum		6,161,805	8,051	1,313,579	1.00E+00	1.00E+00	1.00

GO biological process similarity		# protein pairs	Gold standard overlap		P(GO|pos)	P(GO|neg)	L
			pos	*neg*			
Values	1 -- 9	4,789	88	819	1.17E-02	1.27E-03	9.22
	10 -- 99	20,467	555	3,315	7.38E-02	5.14E-03	14.36
	100 -- 1000	58,738	523	10,232	6.95E-02	1.59E-02	4.38
	1000 -- 10000	152,850	1,003	28,225	1.33E-01	4.38E-02	3.05
	10000 -- Inf	2,909,442	5,351	602,434	7.12E-01	9.34E-01	0.76
Sum		3,146,286	7,520	645,025	1.00E+00	1.00E+00	1.00

Table S2: Calculation of the PIE

The actual computation for the fully connected Bayesian network is simple: The four binary experimental interaction datasets (S7, S20-S23) can be combined in at most 2^4 = 16 different ways (subsets). For each of these 16 subsets, we can compute a likelihood ratio. The format of the table follows that of table S1.

Table S2

Gavin (g)	Ho (h)	Uetz (u)	Ito (i)	# protein pairs	Gold-standard overlap		*P(g,h,u,i | pos)*	*P(g,h,u,i | neg)*	*L*
					pos	*neg*			
0	0	0	0	2702284	6389	2695949	7.74E-01	9.95E-01	0.8
0	1	0	0	23275	87	5563	1.05E-02	2.05E-03	5.1
0	0	0	1	4102	11	644	1.33E-03	2.38E-04	5.6
0	0	1	0	730	5	112	6.06E-04	4.13E-05	14.7
1	0	0	0	29221	1331	6224	1.61E-01	2.30E-03	70.2
0	0	1	1	123	6	23	7.27E-04	8.49E-06	85.7
0	1	0	1	39	3	4	3.64E-04	1.48E-06	246.2
0	1	1	0	29	5	5	6.06E-04	1.85E-06	328.3
0	1	1	1	16	1	1	1.21E-04	3.69E-07	328.3
1	1	0	0	1920	337	209	4.08E-02	7.72E-05	529.4
1	0	1	0	34	12	5	1.45E-03	1.85E-06	788.0
1	1	0	1	27	16	3	1.94E-03	1.11E-06	1751.1
1	0	1	1	22	6	1	7.27E-04	3.69E-07	1970.0
1	1	1	1	11	9	1	1.09E-03	3.69E-07	2955.0
1	0	0	1	53	26	2	3.15E-03	7.38E-07	4268.3
1	1	1	0	16	6	0	7.27E-04	0.00E+00	-

References

S1. R. J. Cho *et al.*, *Mol Cell* **2**, 65-73. (1998).

S2. T. R. Hughes *et al.*, *Cell* **102**, 109-26. (2000).

S3. H. Ge, Z. Liu, G. M. Church, M. Vidal, *Nat Genet* **29**, 482-6. (2001);

S4. R. Jansen, D. Greenbaum, M. Gerstein, *Genome Res* **12**, 37-46. (2002).

S5. P. Kemmeren *et al.*, *Mol Cell* **9**, 1133-43. (2002).

S6. A. Grigoriev, *Nucleic Acids Res* **29**: 3513-9. (2001).

S7. B. Schwikowski, P. Uetz, S. Fields, *Nat Biotechnol* **18**, 1257-61. (2000).

S8. S. Letovsky, S. Kasif, *Bioinformatics* **19**, I197-I204. (2003).

S9. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, *Nat Biotechnol* **21**, 697-700. (2003).

S10. H. W. Mewes *et al.*, *Nucleic Acids Res* **30**, 31-4. (2002).

S11. M. Ashburner *et al.*, *Nat Genet* **25**, 25-9. (2000).

S12. T. Dandekar, B. Snel, M. Huynen, P. Bork, *Trends Biochem Sci* **23**, 324-8. (1998).

S13. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, C. A. Ouzounis, *Nature* **402**, 86-90. (1999).

S14. T. Gaasterland, M. A. Ragan, *Microb Comp Genomics* **3**, 199-217. (1998).

S15. C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, F. E. Cohen, *J Mol Biol* **299**, 283-93. (2000).

S16. S. Tsoka, C. A. Ouzounis, *Genome Res* **11**, 1503-10. (2001).

S17. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc Natl Acad Sci U S A* **96**, 4285-8. (1999).

S18. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev, *In Silico Biol* **1**, 93-108. (1999).

S19. F. Pazos, A. Valencia, *Protein Eng* **14**, 609-14. (2001).

S20. P. Uetz *et al.*, *Nature* **403**, 623-7. (2000).

S21. T. Ito, T. Chiba, M. Yoshida, *Trends Biotechnol* **19**, S23-7. (2001).

S22. T. Ito *et al*., *Proc Natl Acad Sci U S A* **98**, 4569-74. (2001).

S23. A. C. Gavin *et al*., *Nature* **415**, 141-7. (2002).

S24. Y. Ho *et al*., *Nature* **415**, 180-3. (2002).

S25. A. Kumar *et al*., *Genes Dev* **16**, 707-19. (2002).

S26. A. Drawid, M. Gerstein, *J Mol Biol* **301**, 1059-75. (2000).

S27. A. Drawid, R. Jansen, M. Gerstein, *Trends Genet* **16**, 426-30. (2000).

S28. P. Pavlidis, J. Weston, J. Cai, W. S. Noble, *J Comput Biol* **9**, 401-11 (2002).

S29. M. Gerstein, *Nat Struct Biol* **7 Suppl**, 960-3. (2000).

S30. M. Steffen, A. Petti, J. Aach, P. D'Haeseleer, G. Church, *BMC Bioinformatics* **3**, 34. (2002).

S31. V. Pavlovic, A. Garg, S. Kasif, *Bioinformatics* **18**, 19-27. (2002).

S32. J. Pearl, *Probabilistic reasoning in intelligent systems* (Morgan Kaufmann, San Mateo, 1988).

S33. F. V. Jensen, *Bayesian Networks and Decision Graphs* (Springer, New York, 2001).

S34. A. Kumar, M. Snyder, *Nature* **415**, 123-4. (2002).

S35. C. von Mering *et al*., *Nature* **417**, 399-403. (2002).

S36. G. D. Bader, C. W. Hogue, *Nat Biotechnol* **20**, 991-7. (2002).

S37. A. Grigoriev, *Nucleic Acids Res* **15**, 4157-61. (2003).

S38. http://genecensus.org/intint

S39. N. J. Krogan *et al*., *Mol Cell Biol* **22**, 6979-92. (2002).

S40. C. L. Gatlin, G. R. Kleemann, L. G. Hays, A. J. Link, J. R. Yates, 3rd, *Anal Biochem* **263**, 93-101. (1998).

S41. T. Kislinger *et al*., *Mol Cell Proteomics* **2**, 96-106. (2003).

S42. M. Gerstein, N. Lan, R. Jansen, *Science* **295**, 284-7. (2002).

S43. A. H. Tong *et al*., *Science* **295**, 321-4. (2002).

S44. A. Vambutas, S. H. Ackerman, A. Tzagoloff, *Eur J Biochem* **201**, 643-52. (1991).

S45. M. Pioletti *et al*., *Embo J* **20**, 1829-39. (2001)

S46. F. Schluenzen *et al*., *Cell* **102**, 615-23. (2000).

S47. M. Leijonmarck, A. Liljas, *J Mol Biol* **195**, 555-79. (1987).