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, S32).

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 (S33, S34): 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 (S35-S38) 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 (S39). 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 only weakly conditionally dependent. 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.

Table S1 shows the conditional probabilities and likelihood ratios for the individual features that make up the PIP, whereas table S2 shows conditional probabilities and likelihood ratios for different combinations of the protein-protein interaction datasets that make up the PIE.

Distribution of likelihood ratios

The value of integrating multiple data sources can be shown by comparing histograms of the likelihood ratios in the individual datasets with those in the probabilistic interactomes (table S3). Clearly, there are many more protein pairs with high likelihood ratios in the probabilistic interactomes than in the individual datasets. Protein pairs with high likelihood ratios provide leads for further experimental investigation of proteins that potentially form complexes.

Cross-validation

The number of true positives and negatives (with respect to the gold-standard reference) shown in figure 2b were computed by seven-fold cross-validation. For this procedure, the gold-standard positives and negatives were each divided into seven equally sized subsets. Then, each of the seven subsets was set aside once as a test set, while the remaining six subsets were used for training. Each time the conditional probabilities and likelihood ratios were computed from the training set. We then looked for protein pairs in the testing set that superseded the cutoff $L_{cut} = 600$, generating true positives if they were gold-standard positives and false positives if they were gold-standard negatives. The numbers of true and false positives in each of the seven test sets were then summed up to give the total number of true and false positives. The ratio of the total number of true positives to the total number of false positives is what is shown in figure 2b.

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, S40, S41).

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.

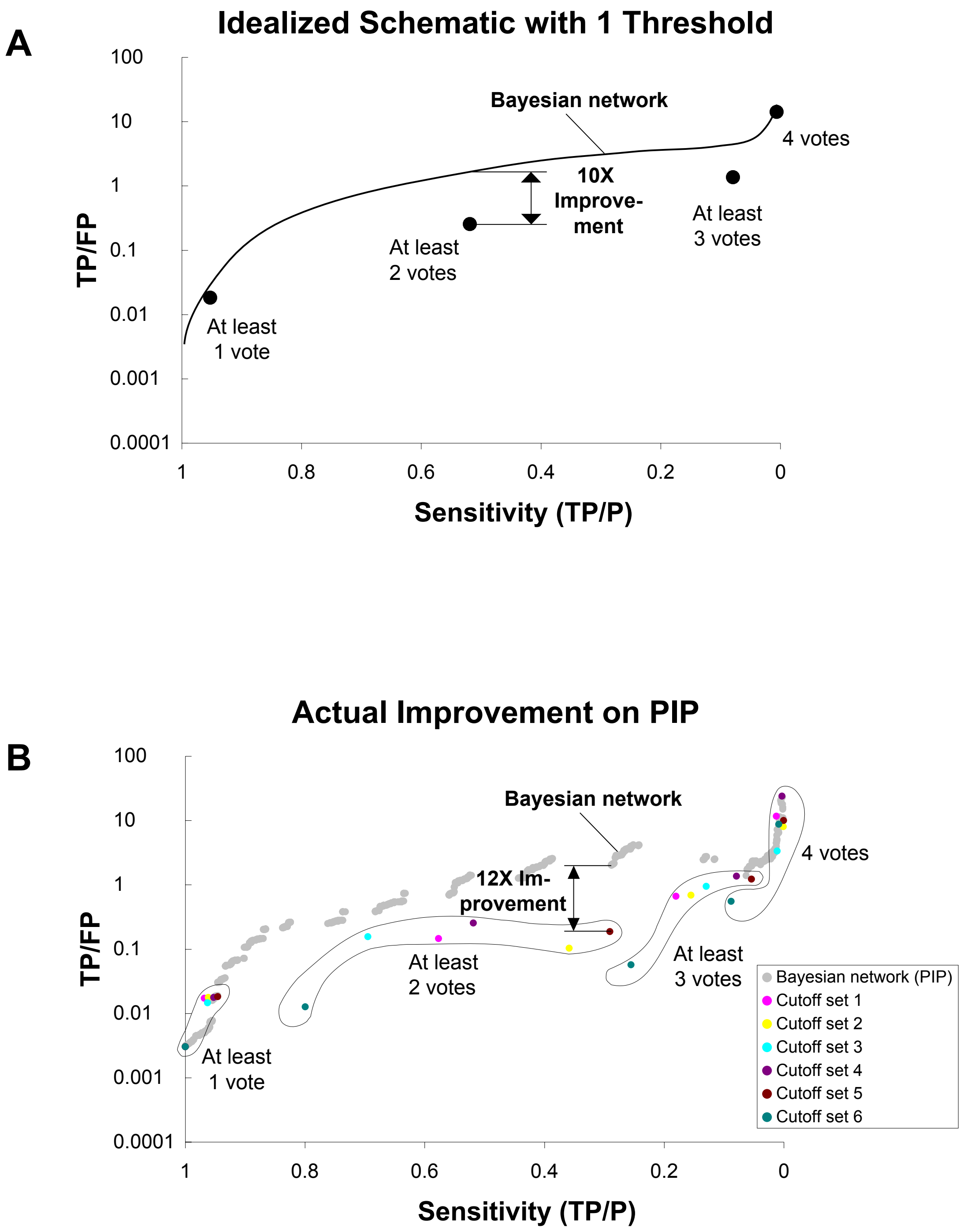
Mitochondrial ribosome

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 (S42). 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* (S43, S44) and the sequence of YGL068W is 52% identical to the L7/L12 ribosomal protein in *E. coli* (S45). Therefore, our predictions for YGL068W and YNL081C seem to provide another level of evidence for annotation of these proteins as mitochondrial ribosomal proteins.

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 (S46), 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 (S47) 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 (S48) with the corresponding tandem mass spectrum to determine the probability of each match.

Figure S1



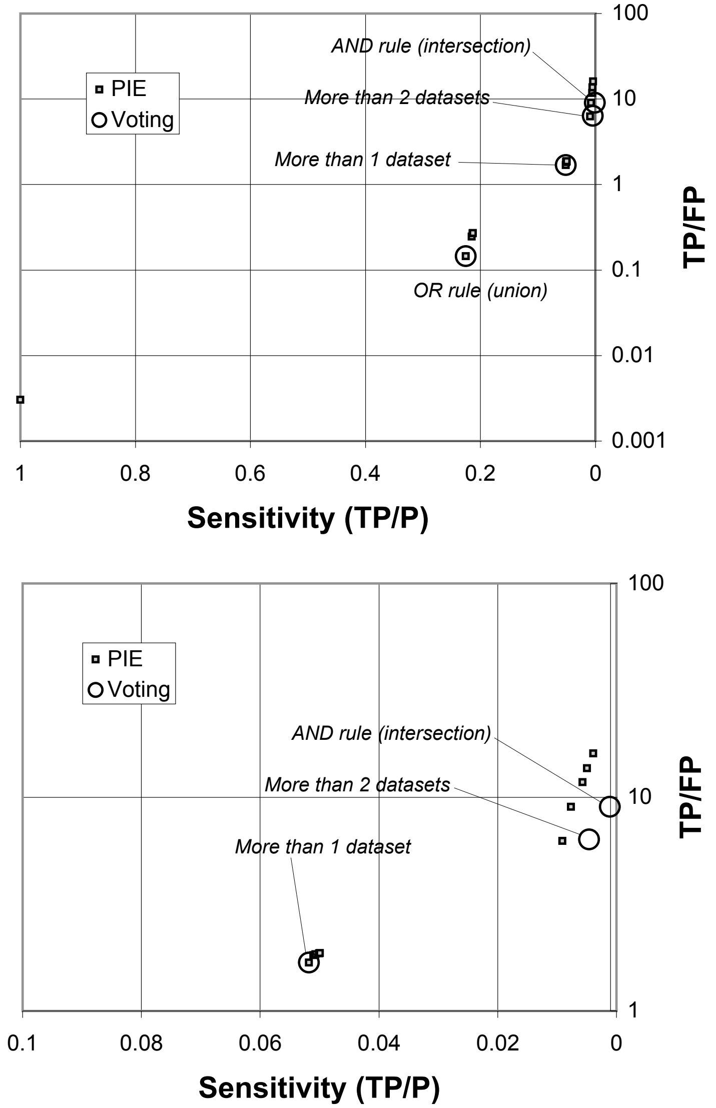
Idealized Schematic with 1 Threshold

Actual Improvement on PIP

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

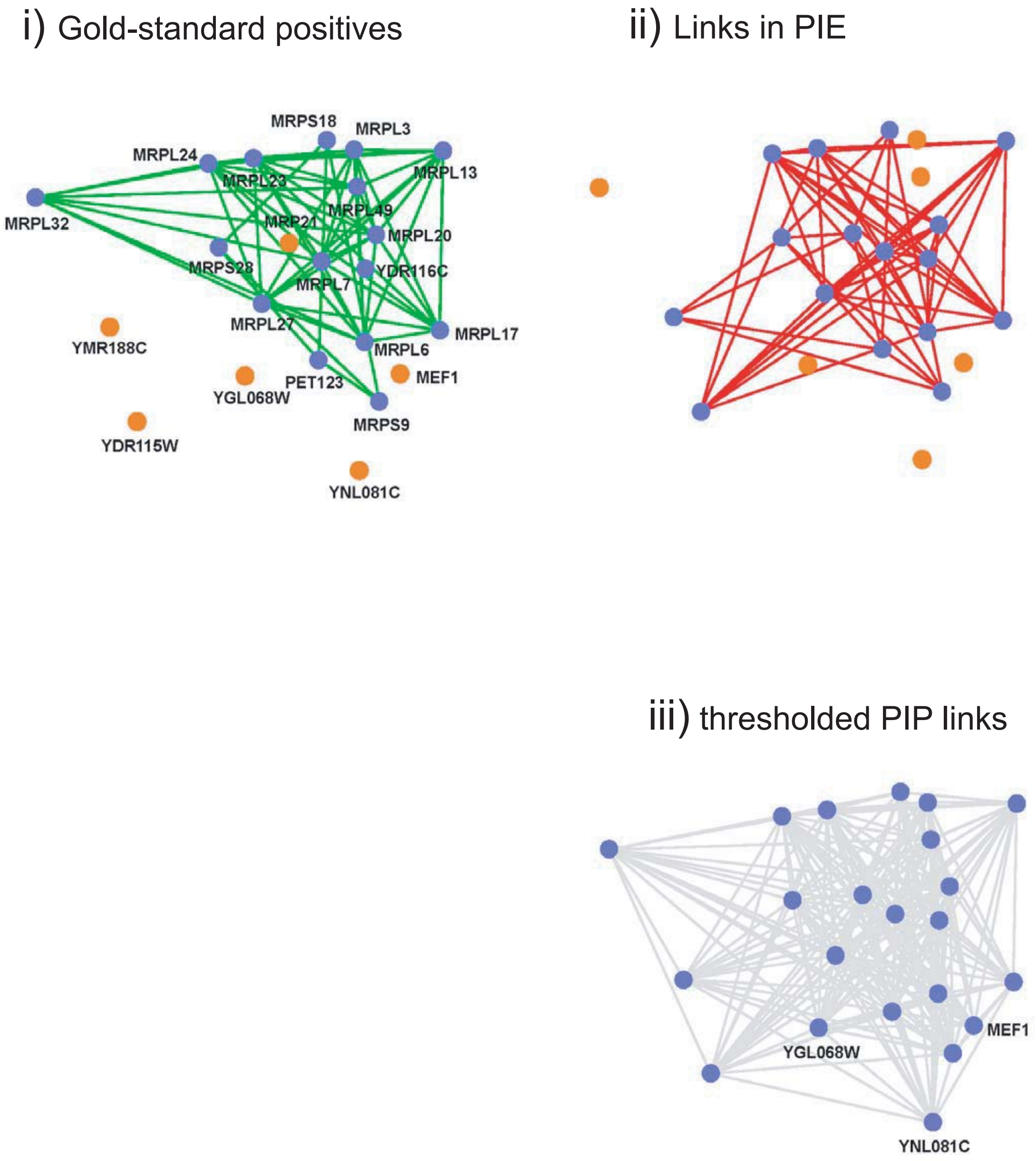


▫ PIE
○ Voting
AND rule (intersection)
More than 2 datasets
More than 1 dataset
OR rule (union)
100
10
1
0.1
0.01
0.001
TP/FP
1
0.8
0.6
0.4
0.2
0
Sensitivity (TP/P)
▫ PIE
○ Voting
AND rule (intersection)
More than 2 datasets
More than 1 dataset
100
10
1
TP/FP
0.1
0.08
0.06
0.04
0.02
0
Sensitivity (TP/P)

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 intervals and the functional similarity counts into 5 intervals. 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 the 2,708,746 gold-standard negatives. The last three columns on the right give the conditional probabilities of the feature values -- *P*(*feature value*|*pos*) and *P*(*feature value*|*neg*) -- and the likelihood ratio *L*, the ratio of these two conditional probabilities.

The column "sum(pos)" shows how many gold-standard positives are among the protein pairs with likelihood ratio greater than or equal to *L*, which can be computed by summing up the values in the column "pos" to the left. Note that "sum(pos)" is not exactly the same as (but similar to) the number of true positives *TP* that are used for generating figure 2b. This is because *TP* in figure 2b is computed from seven-fold cross-validation (see section *Cross-validation*), whereas sum(pos) is just shown here to illustrate how the number of true positives arises.

"Sum(neg)" (similar to the number of *FP* in figure 2b) shows the number of gold-standards negatives among the protein pairs with likelihood ratio greater than or equal to

L. Finally, “sum(pos)/sum(neg)” is similar to the *TP*/*FP* ratio shown in figure 2b.

Table S1

Essentiality		# protein pairs	Gold-standard overlap					*P(Ess|pos)*	*P(Ess|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	EE	384,126	1,114	81,924	1,114	81,924	0.014	5.18E-01	1.43E-01	3.6
	NE	2,767,812	624	285,487	1,738	367,411	0.005	2.90E-01	4.98E-01	0.6
	NN	4,978,590	412	206,313	2,150	573,724	0.004	1.92E-01	3.60E-01	0.5
Sum		8,130,528	2,150	573,724	-	-	-	1.00E+00	1.00E+00	1.0

Expression correlation		# protein pairs	Gold standard overlap					*P(exp|pos)*	*P(exp|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	0.9	678	16	45	16	45	0.36	2.10E-03	1.68E-05	124.9
	0.8	4,827	137	563	153	608	0.25	1.80E-02	2.10E-04	85.5
	0.7	17,626	530	2,117	683	2,725	0.25	6.96E-02	7.91E-04	88.0
	0.6	42,815	1,073	5,597	1,756	8,322	0.21	1.41E-01	2.09E-03	67.4
	0.5	96,650	1,089	14,459	2,845	22,781	0.12	1.43E-01	5.40E-03	26.5
	0.4	225,712	993	35,350	3,838	58,131	0.07	1.30E-01	1.32E-02	9.9
	0.3	529,268	1,028	83,483	4,866	141,614	0.03	1.35E-01	3.12E-02	4.3
	0.2	1,200,331	870	183,356	5,736	324,970	0.02	1.14E-01	6.85E-02	1.7
	0.1	2,575,103	739	368,469	6,475	693,439	0.01	9.71E-02	1.38E-01	0.7
	0	9,363,627	894	1,244,477	7,369	1,937,916	0.00	1.17E-01	4.65E-01	0.3
	-0.1	2,753,735	164	408,562	7,533	2,346,478	0.00	2.15E-02	1.53E-01	0.1
	-0.2	1,241,907	63	203,663	7,596	2,550,141	0.00	8.27E-03	7.61E-02	0.1
	-0.3	484,524	13	84,957	7,609	2,635,098	0.00	1.71E-03	3.18E-02	0.1
	-0.4	160,234	3	28,870	7,612	2,663,968	0.00	3.94E-04	1.08E-02	0.0
	-0.5	48,852	2	8,091	7,614	2,672,059	0.00	2.63E-04	3.02E-03	0.1
	-0.6	17,423	-	2,134	7,614	2,674,193	0.00	0.00E+00	7.98E-04	0.0
	-0.7	7,602	-	807	7,614	2,675,000	0.00	0.00E+00	3.02E-04	0.0
	-0.8	2,147	-	261	7,614	2,675,261	0.00	0.00E+00	9.76E-05	0.0
	-0.9	67	-	12	7,614	2,675,273	0.00	0.00E+00	4.49E-06	0.0
Sum		18,773,128	7,614	2,675,273	-	-	-	1.00E+00	1.00E+00	1.0

MIPS function similarity		# protein pairs	Gold standard overlap					*P(MIPS|pos)*	*P(MIPS|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	1 -- 9	6,584	171	1,094	171	1,094	0.16	2.12E-02	8.33E-04	25.5
	10 -- 99	25,823	584	4,229	755	5,323	0.14	7.25E-02	3.22E-03	22.5
	100 -- 1000	88,548	688	13,011	1,443	18,334	0.08	8.55E-02	9.91E-03	8.6
	1000 -- 10000	255,096	6,146	47,126	7,589	65,460	0.12	7.63E-01	3.59E-02	21.3
	10000 -- Inf	5,785,754	462	1,248,119	8,051	1,313,579	0.01	5.74E-02	9.50E-01	0.1
Sum		6,161,805	8,051	1,313,579	-	-	-	1.00E+00	1.00E+00	1.0

GO biological process similarity		# protein pairs	Gold standard overlap					*P(GO|pos)*	*P(GO|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	1 -- 9	4,789	88	819	88	819	0.11	1.17E-02	1.27E-03	9.2
	10 -- 99	20,467	555	3,315	643	4,134	0.16	7.38E-02	5.14E-03	14.4
	100 -- 1000	58,738	523	10,232	1,166	14,366	0.08	6.95E-02	1.59E-02	4.4
	1000 -- 10000	152,850	1,003	28,225	2,169	42,591	0.05	1.33E-01	4.38E-02	3.0
	10000 -- Inf	2,909,442	5,351	602,434	7,520	645,025	0.01	7.12E-01	9.34E-01	0.8
Sum		3,146,286	7,520	645,025	-	-	-	1.00E+00	1.00E+00	1.0

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 from the overlap with the gold-standard positives (“pos”) and negatives (“neg”). Ordering the rows of the table in the order of the likelihood ratio allows computing “sum(pos)” and “sum(neg)”. 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*	*sum(pos)*	*sum(neg)*	*sum(pos)/ sum(neg)*			
1	1	1	0	16	6	0	6	0	-	7.27E-04	0.00E+00	-
1	0	0	1	53	26	2	32	2	16.0	3.15E-03	7.38E-07	4268.3
1	1	1	1	11	9	1	41	3	13.7	1.09E-03	3.69E-07	2955.0
1	0	1	1	22	6	1	47	4	11.8	7.27E-04	3.69E-07	1970.0
1	1	0	1	27	16	3	63	7	9.0	1.94E-03	1.11E-06	1751.1
1	0	1	0	34	12	5	75	12	6.3	1.45E-03	1.85E-06	788.0
1	1	0	0	1920	337	209	412	221	1.9	4.08E-02	7.72E-05	529.4
0	1	1	0	29	5	5	418	227	1.8	6.06E-04	1.85E-06	328.3
0	1	1	1	16	1	1	413	222	1.9	1.21E-04	3.69E-07	328.3
0	1	0	1	39	3	4	421	231	1.8	3.64E-04	1.48E-06	246.2
0	0	1	1	123	6	23	427	254	1.7	7.27E-04	8.49E-06	85.7
1	0	0	0	29221	1331	6224	1758	6478	0.3	1.61E-01	2.30E-03	70.2
0	0	1	0	730	5	112	1763	6590	0.3	6.06E-04	4.13E-05	14.7
0	0	0	1	4102	11	644	1774	7234	0.2	1.33E-03	2.38E-04	5.6
0	1	0	0	23275	87	5563	1861	12797	0.1	1.05E-02	2.05E-03	5.1
0	0	0	0	2702284	6389	2695949	8250	2708746	0.0	7.74E-01	9.95E-01	0.8

Table S3: Distribution of likelihood ratios

The number of protein pairs in the individual datasets and the probabilistic interactomes as a function of the likelihood ratio (top: PIP, bottom: PIE). Likelihood ratios were binned into four different intervals.

Table S3

Likelihood ratio	# protein pairs				
	PIP	Essentiality	mRNA expression	MIPS functional similarity	GO biological process similarity
0 - 10	18524971	8130528	18610532	5874302	3125819
10 - 100	214686	0	65268	287503	20467
100 - 1000	25404	0	678	0	0
>= 1000	8067	0	0	0	0

Likelihood ratio	# protein pairs				
	PIE	Gavin	Ho	Uetz	Ito
0 - 10	27377	0	0	0	0
10 - 100	30074	31304	25333	0	4393
100 - 1000	2038	0	0	981	0
>= 1000	129	0	0	0	0

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. O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, D. Botstein, *Proc Natl Acad Sci USA* **100**, 8348-53 (2003).

S33. J. Pearl, *Probabilistic reasoning in intelligent systems* (Morgan Kaufmann, San Mateo, 1988).

S34. F. V. Jensen, *Bayesian Networks and Decision Graphs* (Springer, New York, 2001).

S35. A. Kumar, M. Snyder, *Nature* **415**, 123-4. (2002).

S36. C. von Mering *et al.*, *Nature* **417**, 399-403. (2002).

S37. G. D. Bader, C. W. Hogue, *Nat Biotechnol* **20**, 991-7. (2002).

S38. A. Grigoriev, *Nucleic Acids Res* **15**, 4157-61. (2003).

S39. http://genecensus.org/intint

S40. M. Gerstein, N. Lan, R. Jansen, *Science* **295**, 284-7. (2002).

S41. A. H. Tong *et al.*, *Science* **295**, 321-4. (2002).

S42. A. Vambutas, S. H. Ackerman, A. Tzagoloff, *Eur J Biochem* **201**, 643-52. (1991).

S43. M. Pioletti *et al.*, *Embo J* **20**, 1829-39. (2001).

S44. F. Schluenzen *et al.*, *Cell* **102**, 615-23. (2000).

S45. M. Leijonmarck, A. Liljas, *J Mol Biol* **195**, 555-79. (1987).

S46. N. J. Krogan *et al.*, *Mol Cell Biol* **22**, 6979-92. (2002).

S47. C. L. Gatlin, G. R. Kleemann, L. G. Hays, A. J. Link, J. R. Yates, 3rd, *Anal Biochem* **263**, 93-101. (1998).

S48. T. Kislinger *et al.*, *Mol Cell Proteomics* **2**, 96-106. (2003).