BMC Bioinformatics

BioMed Central

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

A direct comparison of protein interaction confidence assignment schemes

BMC Bioinformatics 2006, **7**:360 doi:10.1186/1471-2105-7-360

Silpa Suthram (ssuthram@ucsd.edu)
Tomer Shlomi (shlomito@post.tau.ac.il)
Eytan Ruppin (ruppin@post.tau.ac.il)
Roded Sharan (roded@post.tau.ac.il)
Trey Ideker (trey@bioeng.ucsd.edu)

ISSN 1471-2105

Article type Research article

Submission date 18 April 2006

Acceptance date 26 July 2006

Publication date 26 July 2006

Article URL http://www.biomedcentral.com/1471-2105/7/360

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

http://www.biomedcentral.com/info/authors/

© 2006 Suthram *et al.*, licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A direct comparison of protein interaction confidence assignment schemes

Silpa Suthram[1,2], Tomer Shlomi[3], Eytan Ruppin[3], Roded Sharan[3], Trey Ideker[1,2§]

[1]Department of Bioengineering, University of California, San Diego, CA 92093, USA
[2] Program in Bioinformatics, University of California, San Diego, CA 92093, USA
[3] School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel

[§]Corresponding author

Email addresses:
SS: ssuthram@ucsd.edu
TS: shlomito@post.tau.ac.il
ER: ruppin@post.tau.ac.il
RS: roded@post.tau.ac.il
TI: trey@bioeng.ucsd.edu

Abstract

Background

Recent technological advances have enabled high-throughput measurements of protein-protein interactions in the cell, producing large protein interaction networks for various species at an ever-growing pace. However, common technologies like yeast two-hybrid may experience high rates of false positive detection. To combat false positive discoveries, a number of different methods have been recently developed that associate confidence scores with protein interactions. Here, we perform a rigorous comparative analysis and performance assessment among these different methods.

Results

We measure the extent to which each set of confidence scores correlates with similarity of the interacting proteins in terms of function, expression, pattern of sequence conservation, and homology to interacting proteins in other species. We also employ a new metric, the Signal-to-Noise Ratio of protein complexes embedded in each network, to assess the power of the different methods. Seven confidence assignment schemes, including those of Bader *et al.*, Deane *et al.*, Deng *et al.*, Sharan *et al.*, and Qi *et al.,* are compared in this work.

Conclusions

Although the performance of each assignment scheme varies depending on the particular metric used for assessment, we observe that Deng *et al.* yields the best performance overall (in three out of four viable measures). Importantly, we also find

that utilizing any of the probability assignment schemes is always more beneficial than assuming all observed interactions to be true or equally likely.

Background

Systematic elucidation of protein-protein interaction networks will be essential for understanding how different behaviors and protein functions are integrated within the cell. Recently, the advent of high-throughput experimental techniques like yeast two-hybrid (Y2H) assays [1] and co-immunoprecipitation (co-IP) screens [2] has led to the elucidation of large-scale protein interaction networks in different species, including *S. cerevisiae* (yeast) [2-5], *D. melanogaster* (fly) [6], *C. elegans* (worm) [7] and *H. sapiens* (human) [8, 9]. These networks, while incorporating thousands or tens of thousands of measured interactions, have so far only partially covered the complete repertoire of protein interactions in an organism, and they have been determined to contain a significant number of false-positive interactions depending on the study [10]. However, recent years have also seen an increase in the accumulation of other sources of biological data such as whole genome sequence, mRNA expression, protein expression and functional annotation. This is particularly advantageous as some of these data sets can be utilized to reinforce true (physical) protein interactions while downgrading others. For instance, biologically relevant protein interactions have been shown to have high mRNA expression correlation for the proteins involved [11].

As a result, many integrative bioinformatic approaches have been developed to unearth true protein-protein interactions. These can be mainly divided into two categories: (1) methods that assign reliability measurements to previously observed interactions; and (2) methods that predict interactions *ab initio*. For category (1), Deane *et al.* [12] and Deng *et al.* [13] introduced methods to tackle the problem of assigning reliabilities to interactions using similarity in mRNA expression profiles. Subsequently, Bader *et al.* [14] used additional features of interacting proteins,

including functional similarity and high network clustering [15], to assign confidence scores to protein interactions. For category (2), Marcotte *et al.* [16], von Mering *et al.* [17], Myers *et al.* [18] and Jansen *et al.* [19] were among the first to predict new protein interactions by incorporating a combination of different features like high mRNA expression correlation, functional similarity, co-essentiality, and co-evolution. These schemes calculate a log-likelihood score for each interaction. As yet another approach in this category, Qi *et al*. [20] predicted new protein interactions using a method based on random forests. Presumably, the relative performance of each of these approaches versus the others involves a combination of factors such as the types of evidence used as inputs, the efficacy of each classification algorithms, and the sets of true and false interactions used as gold standards during training. Very recently, a second work by Qi *et al.* [21] studied the effect of the underlying classification algorithm by comparing the accuracies of different classifiers such as naïve Bayes, logistic regression, and decision trees.

Here, we perform a benchmarking analysis to evaluate the published interaction confidence schemes versus one another. Rather than isolate every factor that could influence a scheme's performance, we take a practical approach and evaluate the overall accuracy of each set of confidence scores as reported in the literature and available from the authors' websites. We limit ourselves to works that have assigned confidence scores to a common set of experimentally-observed interactions in yeast; this includes all of the category (1) schemes above, as well as the Qi. *et al.* scheme from category (2). The remaining *ab initio* schemes are concerned with predicting new interactions and do not assign confidences to those interactions that have already been experimentally observed. We also assess the performance of a "null hypothesis", a uniform scheme that considers the same probability for all interactions.

To compare the quantitative accuracy of the methods, we examine the correlations between the confidence estimates and different biological attributes such as function and expression. As a further comparison criterion, we apply the signal processing concept of 'Signal-to-Noise Ratio' (SNR) to evaluate the significance of protein complexes identified in the network based on the different schemes [22]. The discovery of these complexes depends on the connectivity of the interaction network which, in turn, is influenced by the underlying interaction probabilities [22, 23] .

Results

Interaction Confidence Assignment Schemes

Although large-scale protein interaction networks are being generated for a number of species, *S. cerevisiae* is perhaps the best studied among them and is associated with the largest variety and quantity of protein interaction data. Hence, most of the interaction probability schemes have been developed using the yeast protein interaction network as a guide. As the probability schemes were previously computed for different subsets of yeast protein-protein interactions, we compiled a test set of 11,883 yeast interactions common to all schemes. These yeast interactions were derived from both yeast two-hybrid [4, 5] and mass-spectrometry-based [2, 3] screens.

In total, we considered seven interaction probability assignment schemes, including Bader *et al.* [14] (2 schemes), Deane *et al.* [12], Deng *et al.* [13], Sharan *et al.* [23], Qi *et al.* [20] and a default scheme, where all interactions are assigned the same probability. Bader *et al.*, Sharan *et al.* and Qi *et al.* have assigned specific probabilities to every yeast interaction, while Deane *et al.* and Deng *et al.* have grouped yeast interactions into high/medium/low confidence groups. All of the above

schemes define and use some set of gold standard positive and negative interaction examples for the probability estimation.

Bader et al. (BADER_LOW / BADER_HIGH)

As a gold standard positive training data set, Bader *et al.* [14] used interactions determined by co-IP, in which the proteins were also one or two links apart in the Y2H network. The negative training data set was selected from interactions reported either by co-IP or Y2H, but whose distance (after excluding the interaction) was larger than the median distance in Y2H or co-IP respectively. Using these training data, they constructed a logistic regression model that computes the probability of each interaction based on explanatory variables including data source, number of interacting partners, and other topological features like network clustering. We refer to this scheme as Bader *et al.* (low) or BADER_LOW in our analysis.

Initially, the authors used measures based on Gene Ontology (GO) [24] annotations, co-expression, and presence of genetic interactions as measures to validate their data. However, they also combined these measurements into the probability score to bolster their confidence of true interactions. We consider these new confidence scores in our analysis as Bader *et al.* (high) or BADER_HIGH.

Deane et al. (DEANE)

Deane *et al.* [12] estimated the reliability of protein-protein interactions using the expression profiles of the interacting partners. Protein interactions observed in small-scale experiments that were also curated in the Database of Interacting Proteins (DIP) [25] were considered as the gold standard positive interactions. As a gold standard negative, they randomly picked protein pairs from the yeast proteome that were not

reported in DIP. The authors used this information to compute the reliabilities of groups of interactions (obtained from an experiment or a database). Higher reliabilities were assigned to groups whose combined expression profile was closer to the gold standard positive than the gold standard negative interactions. Specifically, reliabilities were assigned to the whole DIP database, the set of all protein interactions generated in any high-throughput genome screen, and protein interactions generated by Ito *et al.* [4].

Deng et al. (DENG)

Deng *et al*. [13] estimated the reliabilities of different interaction data sources in a manner similar to Deane *et al.* [12]. They separately considered experiments that report pair-wise interactions like Y2H and those that report complex membership like mass spectrometry. Curated pair-wise interactions from the literature and membership in protein complexes from Munich Information center for Protein Sequences (MIPS) [26] were used as the gold standard positive set in each case. Randomly chosen protein pairs formed the gold standard negative data set. Reliabilities for each data source were computed using a maximum likelihood scheme based on the expression profiles of each data set. The authors evaluated reliabilities for Y2H data sources like Uetz *et al.* [5] and Ito *et al.* [4], and protein complex data sources like Tandem Affinity Purification (TAP) [2] and High-throughput Mass Spectrometric Protein Complex Identification (HMS-PCI) [3]. In addition to assigning reliabilities to each dataset, the authors also provided a conditional probability scheme to compute probabilities for groups of interactions observed in two or more data sources. This calculation results in assigning a high probability

(0.99) to yeast interactions observed in more than 1 data source. We use the probabilities generated by this method for our comparative analysis.

Sharan et al. (SHARAN)

Recently, Sharan *et al.* [23] also implemented an interaction probability assignment scheme similar to the one proposed by Bader *et al.* The scheme assigned probabilities to interactions using a logistic regression model based on mRNA expression, interaction clustering and number of times an interaction was observed in independent experiments. Here, we use a modification of this scheme, assigning probabilities to interactions based only on direct experimental evidence. Specifically, interactions with at least two literature references or those that had a distance ≤ 2 in both the co-IP and Y2H networks were defined as the gold standard positives. Conversely, proteins at a distance > 4 in the entire network (after removing the interaction in question) were defined as the gold standard negatives. Binary variables were used to denote whether the interaction was reported in a co-IP data set, Y2H data set, a small-scale experiment or a large–scale experiment. Interaction probabilities were then estimated using logistic regression on the predictor parameters similarly to Bader *et al.* [14].

Qi et al. (QI)

In this study, the authors used interactions that were observed in small-scale experiments and reported by either DIP or Bader *et al.* as their gold standard positive training data [20]. Randomly picked protein pairs were used as the gold standard negative training data. The method incorporates direct evidence such as the type of experiment used to generate the data and indirect evidence like gene expression, existence of synthetic lethal interactions, and domain-domain interactions to construct

a random forest (a collection of decision trees). The resulting forest is then used to calculate the probability that two proteins interact.

Equal Probabilities (EQUAL)

Finally, we also considered the case in which all observed interactions were considered to be equally true. We refer to this case as EQUAL in the analysis.

A summary of all attributes used as inputs to the different probability schemes is provided in Table 1. It should be noted that even though the different probability schemes utilize some of the same types of inputs (e.g., experiment type, expression similarity), they may incorporate these inputs in different ways. For instance, both SHARAN and DENG use "experiment type" as input, but SHARAN explicitly includes each type of experiment as a separate indicator variable in its logistic regression function, while DENG pools data from each experimental type and assigns a single confidence level to the interactions in each pool.

We also compared global statistics such as the average and median probability assigned by each scheme (see Additional File 1). We found that most probability schemes had an average probability in the range of [0.3 - 0.5]. In contrast, Deane *et al.* (DEANE) had a very high average and median probability: over half of the interactions in the test set were assigned a probability of 1. We also computed Spearman correlations among the different probability schemes to measure their levels of inter-dependency (Table 2). The maximum correlation was seen between BADER_LOW and BADER_HIGH, as might be expected since both schemes were reported in the same study and BADER_HIGH was derived from BADER_LOW. On the other hand, Qi *et al.* (QI) had very low Spearman correlation with any of the

probability schemes. The low correlation may reflect an inherent difference between schemes that assign probabilities to experimentally observed interactions and ones that predict protein interactions *ab initio*. The probabilities assigned by the schemes can be obtained from the Supplementary website [27].

Quality Assessment

One of the most objective ways to assess the performance of the different confidence assignment schemes would be to compare their success at correctly classifying a gold standard set of true protein interactions. However, all of the schemes considered in this analysis had already used the available gold standard sets of known yeast interactions in the training phase of their classifiers and, consequently, assigned high confidence scores to them. As an alternative approach, we employed five measures that had been shown to associate with true protein interactions [11, 22, 28, 29] to gauge the performance of the schemes. One caveat of this approach is that, in some cases, one of the measures used to assess a scheme's performance had already been used (in full or in part) as an input to assigning its probabilities. To avoid circularity, this measure was used only for gauging the performance of the remaining schemes. For each of the five measures, two ways were used to estimate the level of association: Spearman correlation and weighted average (see Methods). Importantly, by using the Spearman correlation coefficient, we are in fact comparing how the schemes rank the interactions, not the absolute scores that are assigned. Note that the EQUAL probability scheme results in Spearman correlation of 0, by definition.

Presence of Conserved Interactions in Other Species

Presence of conserved interactions across species is believed to be associated with biologically meaningful interactions [29]. As our benchmark, we used yeast protein interactions that were conserved with measured *C. elegans* and *D. melanogaster* interactions obtained from the Database of Interacting Proteins (DIP). An interaction was considered conserved if homologs of the interacting yeast proteins were also interacting in another species. Homologs were based on amino-acid sequence similarity computed using BLAST [30], thus allowing a protein to possibly match with multiple proteins in the opposite species (if interacting yeast proteins were homologous to any pair of homologs with an observed interaction, the yeast interaction was counted as conserved). In particular, we allow interactions whose interacting proteins are themselves homologs, but filter cases where both the interacting proteins pointed to the same protein in the other species. We evaluated the weighted average and Spearman correlation between the probability assignment for each yeast interaction and the number of conserved interactions across worm and fly (0, 1, or 2). We used an E-value cut-off of $1X10^{-10}$ to make the homology assignments (Table 3). We observed that SHARAN and BADER_HIGH had the highest weighted average and Spearman correlation. Not surprisingly, EQUAL had the lowest weighted average. Note that the conserved interactions test is a very strong filter for true interactions as it heavily depends on the level of completeness of the interaction networks of other species being considered. However, as the underlying set of interactions is the same across the different probability schemes, this filter affects all schemes similarly.

Expression Correlation

Yeast expression data for ~790 conditions were obtained from the Stanford Microarray Database (SMD) [31]. For every pair of interacting proteins, we computed the Pearson correlation coefficient of expression. We then calculated the Spearman correlation and weighted average between the expression correlation coefficients of interacting proteins and their corresponding probability assignments in the different schemes (see Table 3 and Additional File 2). We found significant association between expression correlations and probabilities in the case of BADER_HIGH, BADER_LOW, QI and DENG. This result is expected as these schemes, with the exception of BADER_LOW, utilize expression similarity for interaction probability calculation. Surprisingly, DEANE probabilities showed very little correlation with expression, even though mRNA expression profiles were used as input in the prediction process reflecting the difference in the way expression similarity is incorporated in this method. In particular, DEANE is the only method where expression similarity between two interacting proteins is taken into account as the Euclidean distance between their expression profiles versus other methods which incorporated the Pearson correlation coefficient of expression. On the other hand, BADER_LOW had a higher Spearman correlation than SHARAN, though both had very similar weighted averages and did not utilize expression data in the training phase.

Gene Ontology (GO) Similarity

As a first measure, we adopted the common notion that two interacting proteins are frequently involved in the same process and hence should have similar GO assignments [24]. The Gene Ontology terms are represented using a directed acyclic

graph data structure in which an edge from term 'a' to term 'b' indicates that term 'b' is either a more specific functional type than term 'a', or is a part of term 'a'. As a result, terms that appear deeper in the graph are more specific. Moreover, specific terms also have fewer proteins assigned to them or their descendants.

Let "P_i" and "P_j" be two proteins that have been observed to interact with each other. To measure their functional similarity, we evaluated the size (number of proteins assigned to the term), represented as "S_{ij}", of the deepest common GO term assignment (deepest common ancestor in graph) shared between them. Thus, a smaller value of S_{ij} indicates a greater functional similarity between P_i and P_j. In addition, we also found that known yeast interactions generally have lower values for S_{ij} than random background (see Additional File 3). To ensure that higher values of our GO measure correspond to higher performance (as is the case for other quality assessment metrics below), we use the negative of S_{ij} (or $-S_{ij}$) to represent the overall GO similarity.

Table 3 shows the relationship between GO similarity and the interaction probabilities for each scheme. Of the schemes that did not use functional annotations as inputs, DENG and SHARAN both had a very high Spearman correlation with GO (with DENG slightly higher than SHARAN). However, one potential concern was that GO functional assignments could incorporate evidence of co-expression which was used as an input by the DENG scheme. This potential circularity can be addressed by use of the partial correlation coefficient to factor out the dependency of GO on co-expression (see Additional File 4). However, the partial correlation is almost certainly an overcorrection since GO similarity and co-expression (and in fact any two lines of evidence) are expected to have some correlation if they are both predictive of true interactions. Regardless, with or without the correction, DENG and

SHARAN scored within 2% of each other; thus the two schemes are practically indistinguishable by the GO metric.

Signal-to-Noise Ratio of Protein Complexes

Most cellular processes involve proteins that act together by assembling into functional complexes. Several methods [23, 32-35] have been developed to identify complexes embedded within a protein interaction network, in which a complex is typically modeled as a densely-connected protein sub-network. Recently, we showed that the quality of these identified protein complexes could be estimated by computing their signal-to-noise ratio (SNR), a standard measure used in information theory and signal processing to assess data quality (see Methods) [22]. Essentially, SNR evaluates the density of complexes found in the protein interaction network against a randomized version of the same network.

As the SNR is independent of the number of complexes reported, its value can be directly compared across the different probability schemes. For discovery of protein complexes, we applied a previously-published algorithm [23] which includes interaction probabilities in the complex identification process. SNR was then computed on the set of complexes identified by each probability scheme. Results are shown in Table 4; out of all of the schemes, DENG had the highest SNR of protein complex detection.

Conservation Rate Coherency

Interacting proteins have been shown to evolve at similar rates, probably due to selection pressure to maintain the interaction over time [28]. For every pair of interacting proteins, P_i and P_j, let "r_i" and "r_j" be their respective rates of evolution.

We then computed a "conservation rate coherency score" (CR_{ij}) as the negative absolute value of the difference between the evolutionary rates of the two corresponding genes: $CR_{ij} = -| r_i - r_j |$. The negative absolute value was used to ensure that higher values represent higher performance, consistent with other metrics.

Evolutionary rates were obtained from Fraser *et al*. [36] and estimated using nucleotide substitution frequencies. We calculated the Spearman correlation between the values of CR for the interacting proteins and their corresponding probability assignments in the different schemes (see Table 4). For all probability assignment schemes we obtained a statistically significant correlation (p-value < 0.05) between the conservation rate coherency scores and the corresponding probabilities, indicating that proteins with high probability interactions tend to have similar conservation rates. The highest correlation was obtained for DENG.

Discussion

A brief review of the performance results suggests that the DENG method (Deng *et al.*) emerges as the clear winner, with top scores in three out of four non-circular quality metrics. Comprising a 'second tier' are BADER_HIGH, BADER_LOW (the two Bader methods) and SHARAN, which perform very similarly across most metrics with some differences in conservation coherency or gene expression (for which SHARAN performs better or worse, respectively). BADER_LOW, which considers experiment type and interaction clustering as inputs, has a higher expression score than SHARAN, which considers experiment type only, implying that interaction clustering helps capture expression similarity. Interestingly, BADER_HIGH, which incorporates more input attributes than BADER_LOW or SHARAN, does not have

substantially higher rankings. Thus, in this case, adding more inputs to a probability assignment scheme does not appear to strongly enhance its quality.

As for the remaining schemes with lower overall performance (DEANE and QI), it is interesting to note that these were arguably the least and most sophisticated schemes, respectively. The DEANE method relied on only a single evidence type for assigning confidences, that of gene expression, whereas it appears that other factors may have been more informative (Table 1). In contrast to DEANE, QI had the largest number of inputs for assigning confidences and, among these, included data on both co-expression and experiment type. However, it is well known that classifier accuracy can be degraded by including many irrelevant input variables [37], and perhaps this is the reason for QI's lower performance. As an alternative explanation, in Qi *et al.*'s evaluation of classification schemes, they concluded that their method was very successful in predicting co-complex membership, but performed poorly when considering physical interactions [21]. In our analysis, all interactions (even co-complex membership) were treated as pair-wise protein interactions, and this assumption may have contributed to the poor performance of Qi *et al.* Certainly, their classification method was among the most sophisticated of the schemes that we evaluated, and as such it is worthy of future exploration (perhaps with different sources of input data) regardless of its performance in the present study.

Finally, EQUAL almost always scored lowest, regardless of quality metric. Thus, utilizing any probability scheme is better than considering all observed interactions to be true or equally probable.

Beyond these broad rankings, is it possible to synthesize data from five largely independent metrics to arrive at an overall quantitative index of performance? As one approach, we normalized the scores for each metric as a fraction of the best score

achieved within that metric over all confidence assignment schemes (i.e., for each metric, the highest score was fixed to 1 and the scores of the remaining schemes were converted to fractional values between 0 and 1). Table 5 summarises the fractional scores for the six probability schemes and five quality assessment measures. Note that expressing scores as fractional values is an intermediate normalization which preserves the score distribution but compresses its range; although potentially more informative than the non-parametric analysis above based only on ranks, it must also be interpreted with more caution. However, in this case, the fractional scores reinforce the findings reported above based on rank.

Conclusion

We have compared and contrasted seven probability assignment schemes for yeast protein interactions. Surprisingly, Deng *et al.* performs significantly better than others while being one of the least sophisticated. It assigns discrete probability scores to large groups of interactions rather than to individuals, and it inputs just two lines of evidence, experiment type and expression similarity, rather than many. Generalizing these observations, more complex approaches are so far unable to outperform simpler variants. Thus, we arrive at a somewhat unexpected conclusion: At least in interaction confidence assignment, sometimes less means more.

Methods

GO Databases

The Gene Ontology annotations for yeast proteins were obtained from the July 5th, 2005 download of the Saccharomyces Genome Database (SGD) [38]; the graph of relations between terms was obtained from the Gene Ontology consortium (http://www.geneontology.org/).

Weighted Average

The weighted average is given by $WA = \frac{\sum_{i=1}^{N} p_i * m_i}{\sum_{i=1}^{N} p_i}$, where p_i is the probability of a given interaction and m_i is the value of one of the five measures for the interaction.

Signal to noise ratio (SNR)

To compute SNR, a search for dense interaction complexes is initiated from each node (protein) and the highest scoring complex from each is reported. This yields a distribution of complex scores over all nodes in the network. A score distribution is also generated for 100 randomized networks, which have identical degree distribution to the original network. SNR is computed using these original and random score distributions (representing signal and noise, respectively) according to the standard formula [39] using the root-mean-square (rms):

$$\mathrm{SNR} = \log_{10} \frac{\mathrm{rms(original\ complex\ scores)}}{\mathrm{rms(random\ complex\ scores)}}, \quad \text{where } \mathrm{rms}(x_1 \cdots x_M) = \sqrt{\frac{1}{M}\sum_{i=1}^{M} x_i^2}$$

where M denotes the total number of complexes (in this case, equal to the number of nodes) and x_i represents the score of an individual complex.

Authors' contributions

The conception and design of this study was done by SS, TI and RS. The conservation coherency calculations were carried out by TS and the calculations for other metrics were done by SS. The drafts of the manuscript were prepared by SS and TI. The revisions and critical analysis of the manuscript were carried out by RS, ER and TS. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge the following funding support for this research: the National Center for Research Resources (RR018627, SS, TI); the National Institute of General Medical Sciences (GM070743-01, TI); a David and Lucille Packard Fellowship award (TI); the Alon fellowship (RS); and the Tauber Fund (TS).

References

1. Fields S, Song O: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**:245-246.

2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, 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 MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**:141-147.

3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, 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 AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**:180-183.

4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**:4569-4574.

5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, 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, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**:623-627.

6. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of Drosophila melanogaster**. *Science* 2003, **302**:1727-1736.

7. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M: **A map of the interactome network of the metazoan C. elegans**. *Science* 2004, **303**:540-543.

8. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome**. *Cell* 2005, **122**:957-968.

9. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M: **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* 2005, **437**:1173-1178.

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

11. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae**. *Nucleic Acids Res* 2001, **29**:3513-3519.

12. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations**. *Mol Cell Proteomics* 2002, **1**:349-356.

13. Deng M, Sun F, Chen T: **Assessment of the reliability of protein-protein interactions and protein function prediction**. *Pac Symp Biocomput* 2003:140-151.

14. Bader JS, Chaudhuri A, Rothberg JM, Chant J: **Gaining confidence in high-throughput protein interaction networks**. *Nat Biotechnol* 2004, **22**:78-85.

15. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world**. *Proc Natl Acad Sci U S A* 2003, **100**:4372-4376.

16. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**:83-86.

17. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms**. *Nucleic Acids Res* 2005, **33**:D433-437.

18. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG: **Discovery of biological networks from diverse functional genomic data**. *Genome Biol* 2005, **6**:R114.

19. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**:449-453.

20. Qi Y, Klein-Seetharaman J, Bar-Joseph Z: **Random forest similarity for protein-protein interaction prediction from multiple sources**. *Pac Symp Biocomput* 2005:531-542.

21. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: **Evaluation of different biological data and computational classification methods for use in protein interaction prediction**. *Proteins* 2006, **63**:490-500.

22. Suthram S, Sittler T, Ideker T: **The Plasmodium protein network diverges from those of other eukaryotes**. *Nature* 2005, **438**:108-112.

23. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species**. *Proc Natl Acad Sci U S A* 2005, **102**:1974-1979.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25-29.

25. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res* 2002, **30**:303-305.

26. Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F: **MIPS: a database for protein sequences, homology data and yeast genome information**. *Nucleic Acids Res* 1997, **25**:28-30.

27. **Supplementary Website. http://*chianti.ucsd.edu/Suthram2006***.

28. Pagel P, Mewes HW, Frishman D: **Conservation of protein-protein interactions - lessons from ascomycota**. *Trends Genet* 2004, **20**:72-76.

29. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs**. *Genome Res* 2004, **14**:1107-1118.

30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

31. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G: **The Stanford Microarray Database accommodates additional microarray platforms and data formats**. *Nucleic Acids Res* 2005, **33**:D580-582.

32. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.

33. Hu H, Yan X, Huang Y, Han J, Zhou XJ: **Mining coherent dense subgraphs across massive biological networks for functional discovery**. *Bioinformatics* 2005, **21 Suppl 1**:i213-i221.

34. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proc Natl Acad Sci U S A* 2003, **100**:11394-11399.

35. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *Proc Natl Acad Sci U S A* 2003, **100**:12123-12128.

36. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network**. *Science* 2002, **296**:750-752.

37. J.Weston SM, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik: **Feature selection for SVMs.** *Avd. In Neural Inf. Proc Syst. 13* 2001.

38. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM: **Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms**. *Nucleic Acids Res* 2004, **32**:D311-314.

39. Shanmugan KS: *Digital and analog communication systems*. New York: Wiley; 1979.

Tables

Table 1 - Summary of input attributes for the different probability schemes.

Prob. Scheme	Experiment Type	Number of Experimental Observations	Protein-DNA binding	Gene / Protein Expression	Interaction Clustering	SL*	GO*	DDI*	Gene Fusion / Co-occur / Nbrhd*
BADER_LOW	X	X			X				
BADER_HIGH	X	X		X	X	X	X		
DEANE		X		X					
DENG	X	X		X					
SHARAN	X	X							
QI	X		X	X	X	X	X	X	X
EQUAL									

*SL: Synthetic Lethal; GO: Gene Ontology; DDI: Domain-domain Interactions; Nbrhd: Neighborhood

Table 2 - Correlation of different probability schemes*.

	BADER_HIGH	DEANE	DENG	SHARAN	QI
BADER_LOW	0.923	0.655	0.633	0.626	0.095
	BADER_HIGH	0.672	0.644	0.665	0.151
		DEANE	0.718	0.847	-0.090
			DENG	0.680	0.185
				SHARAN	-0.013

*p-values of all correlation measurements were significant (p-value $\leq 2\times10^{-16}$).

Table 3 - Correlation of interaction probabilities with the GO similarity measure, mRNA expression correlation and interaction conservation.*

Prob. Scheme	GO Annotation		Expression Correlation		Interaction Conservation	
	SC	WA	SC	WA	SC#	WA#
BADER_LOW	0.424	-5.850	**0.185**	**0.494**	0.132	0.147
BADER_HIGH	*0.501*	*-5.680*	*0.223*	*0.503*	**0.136**	**0.158**
DEANE	0.385	-5.910	*0.016*	*0.481*	0.098	0.139
DENG	**0.490**	**-5.620**	*0.185*	*0.511*	0.102	0.147
SHARAN	**0.471**	**-5.710**	0.050	0.492	**0.134**	**0.158**
QI	*0.425*	*-6.040*	*0.269*	*0.495*	0.080	0.125
EQUAL	—	-6.320	—	0.482	—	0.102

*Bold values indicate the scheme that performs the best. Italicized values indicate potential circularity, i.e., schemes that use GO annotations or mRNA expression profiles for confidence scoring that are similar to those used here for comparative assessment. *P*-values for all the Spearman correlation measurements are significant. SC: Spearman Correlation; WA: Weighted Average.

All measurements were done at an E-value cut-off of $1X10^{-10}$.

Table 4 - Associations of conservation rate coherency scores and SNR with interaction probabilities.

Prob. Scheme	Conservation Coherency (SC*)	SNR
BADER_LOW	0.090	0.734
BADER_HIGH	0.104	0.735
DEANE	0.113	0.537
DENG	**0.141**	**0.950**
SHARAN	0.126	0.742
QI	0.080	0.706
EQUAL	—	0.657

* SC: Spearman Correlation. Bold values indicate the scheme which performs the best. Note that conservation scores based on weighted averages were omitted as they were very similar across the different confidence assignment schemes.

Table 5 – Fractional scores of the confidence assignment schemes in each of the five quality measures*.

Probability Scheme	Gene Ontology (SC)	Interaction Conservation (SC at 1×10^{-10})	Gene Expression (SC)	SNR	Conservation Coherency (SC)
DENG: Deng *et al.*	**1.00**	0.76	—	**1.00**	**1.00**
BADER_HIGH: Bader *et al.* (high)	—	**1.00**	—	0.77	0.74
BADER_LOW: Bader *et al.* (low)	0.86	0.98	**1.00**	0.77	0.64
SHARAN: Sharan *et al.*	**0.96**	**1.00**	0.27	0.78	0.89
DEANE: Deane *et al.*	0.78	0.73	—	0.57	0.80
QI: Qi *et al.*	—	0.58	—	0.74	0.57

*Fractional scores are between [0,1] with 1 performing the best (indicated in bold for each measure). Cells with a dash (-) indicate circularity, i.e., the measures used as (full or partial) input to the corresponding probability schemes. SC: Spearman Correlation; SNR: Signal to Noise Ratio.

Additional Material

Additional File 1

Format: MS Word Document

Title: Global properties of the probability assignment schemes.

Description: Shows properties like average and median probabilities.

Additional File 2

Format: MS Word Document

Title: Correlation of interaction probabilities with mRNA expression correlation.

Description:

Ribosomal components are among the most co-expressed genes, and could potentially lead to the observed relative importance of co-expression data. To check for the effect of ribosomal proteins, we filtered the yeast interaction set in our analysis to remove all ribosomal proteins and calculated the correlation between co-expression and interaction probability. These results are shown in Additional Table 2. The numbers in the brackets represent the values of Spearman correlation coefficient and weighted average after removing the ribosomal proteins from the interaction data. We observe that removing the ribosomal proteins does not change the values significantly.

Additional File 3

Format: MS Word Document

Title: Histograms of GO similarity scores

Description:

We evaluated the GO similarity scores for known yeast interactions reported in the MIPS database [26]. The histogram of the scores is shown in Additional Figure 1A. We also generated a background distribution by computing the GO similarity scores for 1,000 random interactions (Additional Figure 1B). These random interactions were generated by picking pairs of proteins randomly from the set of interacting proteins in yeast. It is evident from the two figures that true proteins interactions (i.e known yeast interactions reported in MIPS) generally have lower GO similarity scores than the background.

Additional File 4

Format: MS Word Document

Title: Spearman partial correlations for schemes using expression as input

Description:

Spearman Partial Rank Correlation Coefficient

The Spearman partial rank correlation coefficient between two random variables A and X, given the fact that both A and X are correlated to random variable Y, denotes the correlation between A and X, when Y is kept constant. It is calculated as follows:

$$r_{AX,Y} = \frac{r_{AX} - r_{XY} r_{AY}}{\sqrt{(1 - r_{XY}^2)(1 - r_{AY}^2)}}$$

Here, r_{AX}, r_{XY} and r_{AY} represent the Spearman correlation coefficients between A and X, X and Y, and, A and Y respectively. The significance level is given by

$$D_{AX,Y} = 1/2\sqrt{N-4}\ln\left(\frac{1 + r_{AX,Y}}{1 - r_{AX,Y}}\right)$$

$D_{AX,Y}$ has a normal distribution with zero mean and variance one. N represents the size of the data set.

Additional files provided with this submission:

Additional file 4 : Additional File 4.doc : 24Kb
http://www.biomedcentral.com/imedia/6036939111097651/sup4.DOC
Additional file 3 : Additional File 3.doc : 81Kb
http://www.biomedcentral.com/imedia/1494788322109764/sup3.DOC
Additional file 2 : Additional File 2.doc : 28Kb
http://www.biomedcentral.com/imedia/8806399641097646/sup2.DOC
Additional file 1 : Additional File 1.doc : 30Kb
http://www.biomedcentral.com/imedia/1021483372109764/sup1.DOC