

Inferring protein-protein interactions through high-throughput interaction data from diverse organisms

Yin Liu¹, Nianjun Liu², Hongyu Zhao^{2,3,*}

¹Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT

06520, USA

²Department of Epidemiology and Public Health, Yale University School of Medicine,

New Haven, CT 06520, USA

³Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Keywords: domain, likelihood, high-throughput data, protein-protein interaction

ABSTRACT

Motivation: Identifying protein-protein interactions is critical for understanding cellular processes. Because protein domains represent binding modules and are responsible for the interactions between proteins, computational approaches have been proposed to predict protein interactions at the domain level. The fact that protein domains are likely evolutionarily conserved allows us to pool information from data across multiple organisms for the inference of domain-domain and protein-protein interaction probabilities.

Results: We use a likelihood approach to estimating domain-domain interaction probabilities by integrating large-scale protein interaction data from three organisms, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*. The estimated domain-domain interaction probabilities are then used to predict protein-protein interactions in *S. cerevisiae*. Based on a thorough comparison of sensitivity and specificity, Gene Ontology term enrichment and gene expression profiles, we have demonstrated that it may be far more informative to predict protein-protein interactions from diverse organisms than that based on a single organism.

Availability: The program for computing the protein-protein interaction probabilities and supplementary material are available at <http://bioinformatics.med.yale.edu/interaction>

Contact: hongyu.zhao@yale.edu

INTRODUCTION

Protein-protein interactions play critical roles in the control of most cellular processes. Many proteins involved in signal transduction, gene regulation, cell-cell contact and cell cycle control require interacting with other proteins or cofactors for activity in those processes (Papin et al. 2004; Tucker et al. 2001; Wang 2002). Recently, systematic identifications of protein interactions in *S. cerevisiae* have been conducted using high-throughput techniques such as yeast two-hybrid screening methods (Ito et al. 2001; Uetz et al. 2000) or affinity purification coupled with mass spectroscopy (Gavin et al. 2002; Ho et al. 2002). Although these experimental approaches have generated enormous amounts of data and valuable resources for studying protein interactions, these methods suffer from high false positive and false negative rates due to the limitations of these techniques (Mrowka et al. 2001; Von Mering et al. 2002). For example, the false negative rate of the yeast two-hybrid assay used to construct *S. cerevisiae* interaction maps has been estimated to be larger than 70% (Deng et al. 2002). Therefore, there is a great need to develop complementary computational methods capable of accurately predicting interactions between proteins through integrated analysis of data from multiple sources.

A number of computational approaches have been proposed to predict protein-protein interactions, including those based on genomic information (Enright et al. 1999; Tsoka et al. 2000), three-dimensional structural information (Lu et al. 2003; Aloy et al. 2004), integration of multiple genomic datasets (Jansen et al. 2003; Lin et al. 2004; Iossifov et al. 2004), and literature mining (Marcotte et al. 2001). Protein-protein interactions can also be predicted on the basis of evolutionary relationship. It is shown that interacting proteins often exhibit coordinated evolution, so that proteins with similar phylogenetic trees are

more likely to interact with each other (Pazos et al. 2001; Goh et al. 2002; Ramani et al. 2003). In addition, the concept of “interologs” has been proposed based on the idea that a pair of interacting proteins are co-evolving so that their respective orthologs in other organisms tend to interact as well (Walhout et al. 2000).

Several methods have been proposed to predict protein interactions in *S. cerevisiae* on the basis of another important principle, namely domain-domain interactions. Protein domain as a unit of structure, function, and evolution, also serves as a unit for protein-protein interactions. Therefore, it is important to take into account domain-domain interactions when we infer plausible interacting protein pairs. In these methods, proteins are characterized by one or more domains and each domain is responsible for a specific interaction with another domain. Sprinzak and Margalit (2001) identified the domain pairs that are highly correlated with interacting protein pairs using protein-protein interaction data from *S. cerevisiae* as training data. The information was further used to predict interacting protein pairs that contain an interacting domain pair. Similarly, Gomez et al. (2001, 2003) and Deng et al. (2002) estimated the probabilities of domain-domain interactions using protein-protein interaction data from *S. cerevisiae* as training data, and then the estimated domain-domain interaction probabilities can be used to infer protein-protein interaction probabilities. These methods highly depend on the accuracy of the training data and have been mostly applied to protein-protein interaction data from a single organism only, which may be inferior to the methods that can incorporate more information in estimating domain-domain interaction probabilities.

Because domains are likely conserved evolutionarily, information from multiple organisms may be integrated together to improve the estimation of domain-domain interaction probabilities. In our study, we incorporate information from three organisms, *S. cerevisiae*, *C. elegans*, and *D. melanogaster* to effectively utilize the domain information as the evolutionary connection among these model organisms. The protein-domain relationship can be extracted from relevant databases, such as PFAM and SMART (Bateman et al. 2004; Letunic et al. 2004). By integrating large-scale protein-protein interaction data from these three organisms, we have extended a likelihood approach proposed by Deng et al. (2002) to estimate the probabilities of domain-domain interactions based on information from all three organisms. Considering each protein as a collection of domains, we can then estimate the probabilities of protein-protein interactions in *S. cerevisiae* based on the inferred domain-domain interaction probabilities. The protein pairs with their interaction probabilities above a certain threshold can then be predicted to be interacting with each other. In order to assess the performance of our method, we first apply it to the interaction data from *S. cerevisiae* only and compare its performance with that of three other methods that predict protein interactions based on the domain composition of proteins in the cross-validation measurement, and we demonstrate that our method provide comparable performance with others. Then, we compare our prediction results based on all three organisms with those based on *S. cerevisiae* alone. We find that the integrated analysis provides more reliable inference of protein-protein interactions than the analysis from a single organism based on the analysis of sensitivity and specificity, Gene Ontology term enrichment and gene expression profiles.

METHODS

Data Sources

In our study, the high-throughput yeast two-hybrid data from three organisms, *S. cerevisiae*, *C. elegans* and *D. melanogaster* are used to infer domain-domain interaction probabilities. For *S. cerevisiae*, we use a combined dataset from two independent studies (Ito et al. 2000; Uetz et al. 2000), which includes a total of 5,295 interactions. For *C. elegans*, 4,714 interactions were reported from yeast two-hybrid experiments (Li et al. 2004). For *D. melanogaster*, results from two-hybrid experiments yielded a total of 20,349 interaction pairs (Giot et al. 2003). The protein-domain relationships for each protein in *S. cerevisiae*, *C. elegans*, and *D. melanogaster* are extracted from PFAM (Bateman et al. 2004) and SMART (Letunic et al. 2004).

Maximum Likelihood Estimation of Domain-Domain and Protein-Protein Interaction Probabilities

We estimate the probabilities of domain-domain interactions through the extension of a likelihood approach proposed by Deng et al. (2002) so that it can incorporate information from all three organisms. In this model, we make the following assumptions: 1) Domain-domain interactions are independent, so whether two domains interact or not does not depend on the interactions among other domains. 2) The probability that two domains m and n interact is the same among all the three organisms. 3) Two proteins i and j interact if and only if at least one pair of domains from the two proteins interact.

With these assumptions, we have $\Pr(P_{ijk} = 1) = 1 - \prod_{(D_{mn} \in P_{ijk})} (1 - \lambda_{mn})$, where P_{ijk} represents the protein pair i and j in species k , $P_{ijk} = 1$ if protein i and protein j in species k interact with each other, and $P_{ijk} = 0$ otherwise. Here, $k = 1, 2, 3$ represents species *S. cerevisiae*, *C. elegans*, and *D. melanogaster* respectively, λ_{mn} represents the probability that domain m interacts with domain n , and the notation $(D_{mn} \in P_{ijk})$ denotes all pairs of domains from protein pair i and j in species k . The probability that proteins i and j in species k are observed to be interacting in the experiments is:

$\Pr(O_{ijk} = 1) = \Pr(P_{ijk} = 1)(1 - fn) + (1 - \Pr(P_{ijk} = 1))fp$, where $O_{ijk} = 1$ if interaction between protein i and j is observed in species k , and $O_{ijk} = 0$ otherwise. Here, fn and fp represent the false negative rate and false positive rate of the protein interaction data. It was estimated that total number of interactions between all yeast proteins is about 20,000~30,000 (Bader et al. 2004). Therefore, for *S. cerevisiae*, we have

$$\begin{aligned} fn &= \Pr(O_{ijk} = 0 | P_{ijk} = 1) = 1.0 - \frac{\Pr(O_{ijk} = 1, P_{ijk} = 1)}{\Pr(P_{ijk} = 1)} \\ &\geq 1.0 - \frac{\Pr(O_{ijk} = 1)}{\Pr(P_{ijk} = 1)} = 1.0 - \frac{\text{Number of observed interacting pairs}}{\text{Number of real interacting pairs}} \geq 1.0 - \frac{5295}{20000} \geq 0.74. \end{aligned}$$

We obtained a total of 5,717 proteins from SWISS-PROT and TrEMBL, therefore,

$$\begin{aligned} fp &= \Pr(O_{ijk} = 1 | P_{ijk} = 0) = \frac{\Pr(O_{ijk} = 1, P_{ijk} = 0)}{\Pr(P_{ijk} = 0)} \\ &\leq \frac{\Pr(O_{ijk} = 1)}{\Pr(P_{ijk} = 1)} = \frac{\text{Number of observed interacting pairs}}{\text{Total protein pairs} - \text{Number of real interacting pairs}} \\ &\leq \frac{5295}{5717 * (5717 + 1) / 2 - 30000} \leq 3.3E - 4 \end{aligned}$$

Similarly, for *C. elegans*, the fn is about 0.90 by mapping the observed interactions to a benchmark data set (Li et al. 2004) and we estimate fp to be less than 3E-5. For *D. melanogaster*, the fn is about 0.80 (Giot et al. 2003) and we estimate fp to be less than 3.6E-4.

The likelihood function that characterizes the probability of the observed protein interaction data across all three organisms is: $L = \prod \Pr(O_{ijk} = 1)^{O_{ijk}} (1 - \Pr(O_{ijk} = 1))^{1-O_{ijk}}$.

We can see that the likelihood function L is a function of parameter λ_{mn} if we specify fixed values for fn and fp . To obtain the MLEs of the parameters, we propose to use the EM algorithm (Dempster et al. 1977) that consists of the expectation (E) step and the maximization (M) step. In the E-step, we need to calculate the expectations of the complete data given the observed data. Here, the complete data include all the domain-domain interactions for each protein-protein pair i and j of each of the three organisms, denoted by $D_{mn}^{(ij)}$. We have

$$E(D_{mn}^{(ij)} | O_{ijk} = o_{ijk}, \lambda_{mn}) = \frac{\lambda_{mn}^{(t-1)} (1 - fn)^{o_{ijk}} fn^{1-o_{ijk}}}{\Pr(O_{ijk} = o_{ijk} | \lambda_{mn}^{(t-1)})}.$$

With the expectations of the complete data, in the M-step, we update the λ_{mn} by,

$$\lambda_{mn}^{(t)} = \frac{\lambda_{mn}^{(t-1)}}{N_{mn}} \sum \frac{(1 - fn)^{o_{ijk}} fn^{1-o_{ijk}}}{\Pr(O_{ijk} = o_{ijk} | \lambda_{mn}^{(t-1)})},$$

where N_{mn} is the total number of protein pairs containing domain (m, n) across the three organisms, and the summation is over all these protein pairs.

We update the parameter estimates of the λ_{mn} by iterating between the E-step and the M-step until convergence to obtain the MLEs of the λ_{mn} for all the domain pairs. The estimated values of the λ_{mn} allow us to compute the protein interaction probabilities so that two proteins with an interaction probability greater than a certain threshold can be predicted to be interacting partners.

Cross-Validated Comparison and ROC Analysis

To compare our likelihood approach with other similar methods that predict protein interactions based on protein domain information, we measure the performance of each prediction using a five-fold cross-validation. As all the other methods predicting protein interaction pairs are applied on the interaction data from *S. cerevisiae* only, we define the training interaction data for the cross-validation as follows: we considered the 3,543 yeast physical interaction pairs in MIPS as positive examples, and the other possible protein pairs, totally 6,895,215 pairs as negative examples. At each iteration of the cross-validation experiments we reserve one fifth of both positives and negatives for testing and use the remaining data for training. The training-test procedure is repeated five times.

The prediction accuracy is measured using the Receiving Operator characteristic (ROC) curve that demonstrates the tradeoffs between sensitivity and specificity. It is a plot of the true positive rate (sensitivity) against the false positive rate (1- specificity) for different thresholds. Here, the true positive rate, denoted as TPF, is calculated as the number of predicted protein pairs that are included in the positive examples divided by 3,543, the total number of positives, and the false positive rate, denoted as FPF, is calculated as the number of predicted protein pairs that are included in the negative examples divided by 6,895,215,

the total number of negatives. The ROC score, calculated as the area under the ROC curve is a measurement of prediction accuracy. The closer the ROC score to 1.0, the better the prediction is. In our study, we repeat the entire cross-validation procedure for three times in order to estimate the variance of the ROC score.

Gene Ontology Analysis

We determine whether the two genes encoding the predicted interacting protein pair have any GO annotation enriched in the biological process ontology by using Saccharomyces Genome Database (SGD) GO TermFinder (<http://search.cpan.org/dist/GO-TermFinder/>). The probability that two genes share the same biological process by chance is calculated through the hypergeometric distribution. The p-value is calculated using the following equation:

$$p - value = \sum_x^n \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}},$$

where N and M represent the total number of genes in the population and the number of genes have a particular biological process category annotation, respectively, and n and x represent the number of genes in the set and the number of genes in the set annotated with the particular biological process, respectively. As each gene set we investigate is a pair of genes, both n and x are equal to 2. The p-value is corrected for multiple testing using Bonferroni correction and a protein pair is considered as have a GO term enriched if the corrected p-value is less than 0.05.

To assess the overall statistical significance of the observed GO term enrichment, we generate randomized protein-domain associations by randomly permuting the domain labels of all proteins while leaving the number of domains associated with each protein untouched. We then run the same prediction procedure on the permuted domain information. This process is repeated 100 times and the number of predicted protein pairs having GO term enriched is recorded for each permutation. The empirical p-value for the observed GO term enrichment is calculated as the fraction of the permutations having a larger number of GO term enriched protein pairs than that based on the observed data.

RESULTS

The protein-domain relationships are extracted from PFAM and SMART, and there are a total of 3,317 domains associated with the proteins of the three organisms (*S. cerevisiae*, *C. elegans*, and *D. melanogaster*). The distribution of these domains across the three organisms is shown in a Venn diagram in Figure 1.

Sensitivity and Specificity

In this study, we have extended a likelihood approach by Deng et al. (2002) to integrate information from diverse organisms to infer protein-protein interaction probabilities. We compare the performance of the likelihood approach with three other methods that have also been used for protein interaction prediction, including the sequence-signature method proposed by Sprinzak and Margalit (2001), the attraction-only model (Gomez et al. 2001), and the attraction-repulsion model (Gomez et al. 2003). All these four methods explore the experimental protein interaction data to assign the probability or score for each protein pair, and make predictions of interacting protein pairs based on a selected decision threshold. To

compare the performance of each prediction method, we apply these methods to the same training interaction data obtained from a single organism - *S. cerevisiae* only, and measure the performance of each method using the five-fold cross-validation. For different thresholds, the sensitivity and specificity of each prediction method are calculated and the ROC scores that measure the accuracy of prediction for each method are obtained (See the Methods section). The results in Figure 2 clearly demonstrate that with only the information from a single organism, the prediction performance of the likelihood approach, with a ROC score of 0.628 ± 0.005 , is comparable with that of the attraction-repulsion model, and is significantly better than that of the attraction-only model and the sequence-signature method.

The advantage of our extended likelihood approach is that it allows us to incorporate the large-scale protein-protein interaction data from diverse organisms. In order to assess the benefit of simultaneous analysis of multiple organisms, we investigate the information gain from the joint analysis of all three organisms compared with the analysis solely based on *S. cerevisiae*. Because information from *C. elegans* and *D. melanogaster* can affect (and hopefully improve) the estimated domain-domain interaction probabilities in *S. cerevisiae*, the predicted protein-protein interactions differ between the two methods. Taking the 3,543 protein-protein physical interactions recorded in MIPS as true positives, we estimate the sensitivity and specificity for each threshold of the two methods either based on information from all three organisms or based on information from *S. cerevisiae* alone. The results are summarized in the ROC curves in Figure 3. The improvement based on the joint analysis of three organisms can be easily seen from this figure.

Evaluation of GO Term Enrichment

In order to evaluate the quality of our predicted protein interactions, we investigate whether two genes encoding a predicted interacting protein pair are functionally related. Because genes more likely share the same biological process if they are functionally related (Vazquez et al. 2003), we determine whether these two genes have any GO annotation enriched in the biological process ontology compared to what would be expected by chance from a random pair of genes. We observe that out of the top 1,000 predicted interacting protein pairs based on the information of all three organisms, 203 pairs have at least one GO term enriched, while only 91 pairs out of the top 1,000 predicted pairs based on the information of yeast alone have a GO term enriched. To assess the statistical significance of these results, we compare these predictions with those based on randomized protein-domain associations (see the Methods section). We find that the observed 203 GO term enriched pairs based on the information from all three species are statistically significant (empirical p-value is 0), whereas the observed 91 Go term enriched pairs based on *S. cerevisiae* alone are not statistically significant (empirical p-value is 0.06).

Gene Expression Profiles

Interacting proteins are more likely to be co-expressed than a random pair of genes and this fact has been used for experimental validation of the predicted protein-protein interactions (Ge et al. 2001; Kemmeren et al. 2002). In our study, we test whether there is statistical evidence suggesting that gene expression profiles are more similar between the predicted protein pairs, where the similarity is defined by the Pearson correlation coefficient between the gene expression profiles of these two genes. For gene expression profiles, we use publicly available gene expression data, including a time course study during the yeast cell

cycle (Spellman et al. 1998) and the Rosetta “compendium” set that is composed of 300 diverse mutations and chemical treatments (Hughes et al. 2000).

To test whether the correlation coefficients of gene expressions for the predicted interacting protein pairs are significantly higher than those for random gene pairs, we compare the distribution of the correlation coefficients between the predicted interacting protein pairs with a probability threshold of 0.1, the physical interaction protein pairs from MIPS, the predicted interacting pairs excluding those pairs from MIPS, and random pairs. We find the distribution of the correlation coefficients of the predicted protein pairs is similar to that of the annotated interacting protein pairs in MIPS, which are verified interacting proteins. Compared with random protein pairs, the predicted protein pairs have higher mean correlation coefficient (Supplementary Data). In addition, we compare the mean expression correlation coefficient for the predicted interacting protein pairs based on information from all three organisms and that based on information from *S. cerevisiae* alone. For this comparison, we first identify the top N predicted interacting pairs based on either method, where N is varied from 100, 500, 1000, 2000, 5000, to 10000. We then calculate the average correlation coefficient for the predicted interacting pairs in the set for each method. As shown in Table 1, as N increases, the mean correlation coefficient decreases due to the inclusion of a larger proportion of false positives in the data set. More importantly, for any given N , the mean correlation coefficient for the predicted interacting protein pairs based on the information from all three organisms is significantly higher than that for protein pairs predicted using the information from *S. cerevisiae* alone. In addition, the distributions of the correlation coefficients for the top 1000 predicted protein pairs based on two different sources are shown in Figure 4. As can be seen from this figure, there is a general shift of the

distribution to higher correlation coefficient values for protein pairs predicted based on the information from all three organisms than those predicted based on *S. cerevisiae* alone, indicating that the prediction based on the information from all three organisms more likely yields more reliable predicted interacting protein pairs.

Biological Significance of the Predictions

In this section, we discuss the biological relevance of the predicted interacting protein pairs. Although many of the predicted pairs are in the MIPS database, some of the top ones are not. Table 2 summarizes the top 10 predictions that are not in the MIPS database and all these predictions have their estimated interaction probabilities equal to 1. Table 2 also provides the functional annotation of these genes. Some of our predicted protein pairs include the subunits of the same protein complex, for example, MCD1 and IRR1 are subunits of the yeast cohesin complex. Some other predictions involve interactions between proteins belonging to the same family, such as OCA1 and SNZ1 or between members of two different families, such as the VAC and ECM families. The interactions between the VAC8, a phosphorylated vacuole membrane protein that is required for protein targeting from cytoplasm to vacuole (Scott et al. 2000), and the members of the ECM family, such as ECM15, may indicate that the ECM proteins are required for vacuole formation in the three-dimensional extracellular matrices.

Some of our predictions may be biologically important. For example, it has been shown that the lack of Srp1 export might impair cNLS-dependent nuclear protein import in yeast (Stade et al. 2002). Because the ubiquitin-like modification of some proteins, such as RanGAP1, is required for protein nucleocytoplasmic trafficking (Matunis et al. 1998), the ubiquitin ligase

may be involved in the nuclear protein import. Therefore, it may be reasonable to consider that Srp1 and BUL2, a component of the ubiquitin ligase complex, interact with each other and play a role in the nuclear protein import process together. The interaction between CUP2 and THI4 may indicate that genes activated by the transcription factor CUP2 are involved in the process of thiamine biosynthesis, in which THI4 plays an important role. Another example is the protein-pair DCS1-NTH2. NTH2 is a neutral trehalase, and it has been proposed that the phosphorylation of DCS1 by CaM kinase II would lead its dissociation from the neutral trehalase, thus the activity of the neutral trehalase would be upregulated (Souza et al. 2002). Therefore, the lack of CaM kinase II would down-regulate the neutral trehalase activity due to the interaction between DCS1 and NTH2. In addition, we may predict the functions of some unknown proteins based on their interacting partners. For example, YMR009W is predicted to interact with FUN34, a transmembrane protein that is involved in ammonia production, therefore, we can predict that YMR009W may also be involved in this process.

CONCLUSIONS AND DISCUSSION

In this article, we propose to estimate the probabilities of interactions between domain pairs by pooling information from three organisms - *S. cerevisiae*, *C. elegans*, and *D. melanogaster* based on large-scale protein interaction data. Using the estimated domain-domain interaction probabilities, we can then estimate the probabilities of interactions between each protein pair in a given organism. We focus our attention on predicting the protein interactions in *S. cerevisiae*, and we have found that even based on the information from *S. cerevisiae* only, the likelihood approach is among the best-performing methods considered in our comparisons. Because of the experimental errors of large-scale two

hybrid assays, the domain interactions inferred from one organism may not be reliable, and the incorporation of data from other organisms can indeed improve the estimated domain-domain and protein-protein interactions. The extension of the likelihood approach allows the incorporation of the information from all three organisms, and the prediction results were found better than those obtained based on the information from *S. cerevisiae* alone through the examinations of ROC curves, GO term enrichments, and expression profiles. Therefore, we conclude that the approach proposed in this study outperforms those used for comparison, providing more informative inference of protein interactions.

The results from our approach can be further improved when the domain information is further and more reliably annotated in the future. Currently, only about 2/3 of the *S. cerevisiae* proteins have a defined domain composition, and we have only considered possible interactions between those proteins with annotated domain information. As a result, the predictions based on domain-domain interactions will only be able to capture a portion of all interactions, the number of which is estimated about 20,000~30,000 in *S. cerevisiae*. Our predicted interacting pairs depend on the threshold value used for the estimated interaction probabilities and the number of predicted pairs increases as we reduce the threshold. Due to the unknown number of truly interacting protein pairs as well as the incompleteness of the annotated domain information, it is difficult to set a threshold value to match the expected number of interacting pairs. When we set the threshold at 0.1, 20,088 protein pairs are predicted to interact with each other. At this level, using MIPS physical interaction data as the gold standard, we estimate the sensitivity and specificity to be 38.6% and 99.7% respectively (The list of all the predicted interactions is provided as supplementary information). As the interacting protein pairs included in MIPS are far from

being complete, these values calculated based on the MIPS data could be different from the actual values.

It is well-known that the two-hybrid assays contain many errors and the exact error rates are hard to assess because the actual protein-protein interactions are not yet known. Based on the number of interactions in our training data, we have estimated the range of the false positive rates and false negative rates (See the Methods Section). The estimated value of fn agrees with the literature that publish the dataset, while the estimated value of fp differs with those established in the literature by an order of magnitude because a different definition of false positive (the number of incorrect interaction observed in experiments divided by the total number of observed interactions) is used. We fix the fn and fp rates in our analysis as this approach has been shown to be robust with respect to a range of experimental error rates (Supplementary Data). In our study, we set the error rates to be $fp = 3E-4$ and $fn = 0.85$ for the interaction data for all three organisms to ease the computation, the yielded predictions are used for the GO term enrichments and gene expression analysis. In addition, we have applied our approach to a core interaction dataset including 1,374 interactions from *S. cerevisiae* (Ito et al. 2000; Uetz et al. 2000), 2,135 interactions from *C. elegans* (Li et al. 2004), and 4,625 interactions from *D. melanogaster* (Giot et al. 2003). We set the error rates to be $fp = 0$ and $fn = 0.95$ because the dataset contains only high-confidence interactions. However, the analysis yields a smaller number of predicted interactions, and measured by sensitivity and specificity, the overall performance of the core dataset is not comparable with that of the dataset including all the interactions (Supplementary Data). Given the core dataset only contains ~ 8000 interactions for all the three organisms, which is much smaller than the number of expected interactions, the information included in the core dataset may be

further from being complete than the complete dataset though it has a smaller false positive rate, thus limiting the prediction power of our approach.

We predict protein-protein interactions through the annotated protein domains, which are responsible for protein interactions through direct physical interactions. Therefore, our goal, precisely defined, is to predict whether two proteins have direct physical interactions, not the proteins that are in the same complex. In this study, we have focused on the integration of two-hybrid data from different organisms. The prediction reveals potential protein physical interactions, but some of which may not be biologically relevant in a physiological condition. In principle, other types of data can be integrated into the approach, for example, the integration of data from high-throughput mass spectrometry protein complex purification along with the correlated mRNA expression profiles are expected to extend our prediction, yielding functional related protein pairs.

The basic principle of our approach is based on the fact that domain-domain interactions are likely conserved across different organisms, therefore allowing us to borrow information from diverse organisms to improve the predictions of protein-protein interactions in a given organism. Although our current approach has indeed led to improved predictions, it can be further refined to generate more accurate predictions. For example, we may first improve the predictions of protein-protein interactions within the same organism through integrating diverse data sources from that organism (e.g. Jansen 2003; Lin et al. 2004) and then perform joint analysis across different organisms based on the results from these integrated analyses. The current approach estimates the domain-domain interaction probabilities for each domain-domain pair separately, and these estimated probabilities may be more accurately estimated

by pooling information from domains with similar structures or functions. Finally, a Bayesian approach may be adopted here both to incorporate prior information on domain-domain interactions as well as to better infer domain-domain interaction probabilities.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation grant DMS-0241160 and YL was supported by the NIH Institutional Training Grants for Informatics Research.

REFERENCES

- Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R.B. (2004) Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026-9.
- Bader, J.S., Chaudhuri, A., Rothbergm J.M., and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnology* **22**, 78-85.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-41.
- Dempster, A.P., Laird, N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, **39**, 1C38.
- Deng, M., Mehta, S., Sun, F., and Chen, T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* **12**, 1540-8.
- Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

- Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.
- Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet.* **29**, 482-6.
- Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-36.
- Goh, C.S. and Cohen, F.E. (2002) Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol.* **324**, 177-92.
- Gomez, S.M., Lo, S.H., and Rzhetsky, A. (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* **159**, 1291-8.
- Gomez, S.M., Noble, W.S., and Rzhetsky, A. (2003) Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* **19**, 1875-81.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-3.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26
- Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J.S., White, K.P., and Rzhetsky, A. (2004) Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics* **20**, 1205-13.

- Ito, T., Chiba, T., Ozawa, R., Yoshida M., Hattori, M., et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*. **98**, 4569-74.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53.
- Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell*. **9**, 1133-43.
- Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*. **32**, D142-4.
- Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-3.
- Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics* **5**, 154.
- Lu, L., Arakaki, A.K., Lu, H., and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res*. **13**, 1146-54.
- Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001) Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359-63.

- Matunis, M.J., Wu, J., and Blobel, G. (1998) SUMO-1 modification and its role in targeting the Ran GTPase-activating protein, RanGAP1, to the nuclear pore complex. *Cell Biol*, **140**, 499-509.
- Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*. **32**, D41-4.
- Mrowka, R., Patzak, A., and Herzel, H. (2001) Is there a bias in proteome research? *Genome Res*. **11**, 1971-3.
- Papin, J. and Subramaniam, S. (2004) Bioinformatics and cellular signaling. *Curr Opin Biotechnol*. **15**, 78-81.
- Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*. **14**, 609-14.
- Ramani, A.K. and Marcotte, E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*. **327**, 273-84.
- Scott, S.V., Nice, D.C., Nau, J.J., Weisman, L.S., Kamada, Y., Keizer-Gunnink, I., Funakoshi, T., Veenhuis, M., Ohsumi, Y., Klionsky, D.J. (2000) Apg13p and Vac8p are part of a complex of phosphoproteins that are required for cytoplasm to vacuole targeting. *J Biol Chem*. **275**, 25840-9.
- Souza, A.C., De Mesquita, J.F., Panek, A.D., Silva, J.T., Paschoalin, V.M. (2002) Evidence for a modulation of neutral trehalase activity by Ca²⁺ and cAMP signaling pathways in *Saccharomyces cerevisiae*. *Braz J Med Biol Res* **35**, 11-6.
- Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher. B. (1998) Comprehensive identification of cell cycle-

- regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. **9**, 3273-97.
- Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*. **311**, 681-92.
- Stade, K., Vogel, F., Schwienhorst, I., Meusser, B., Volkwein, C., Nentwig, B., Dohmen, R.J., and Sommer, T. (2002) A lack of SUMO conjugation affects cNLS-dependent nuclear protein import in yeast. *J. Biol. Chem.* **277**, 49554-49561.
- Tsoka, S. and Ouzounis, C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet*. **26**, 141-2.
- Tucker, C.L., Gera, J.F., and Uetz P. (2001) Towards an understanding of complex protein networks. *Trends Cell Biol*. **11**, 102-6.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-7.
- Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat Biotechnology* **21**, 697-700.
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
- Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116-22.
- Wang, J. (2002) Protein recognition by cell surface receptors: physiological receptors versus virus interactions. *Trends Biochem Sci*. **27**, 122-6.

FIGURE LEGENDS

Figure 1. The distribution of the domains in *S. cerevisiae*, *C. elegans*, and *D. melanogaster*

Figure 2. ROC score summary. Error bars indicate the standard deviation over three cross-validation experiments.

Figure 3. ROC curves of the prediction results based on different information sources

Figure 4. Comparisons of the distributions of the Pearson correlation coefficients for the top 1,000 predicted interacting protein pairs based on different information sources. sdc, prediction based on the information from three organisms *S. cerevisiae*, *D. melanogaster* and *C. elegans*.

FIGURE 1.

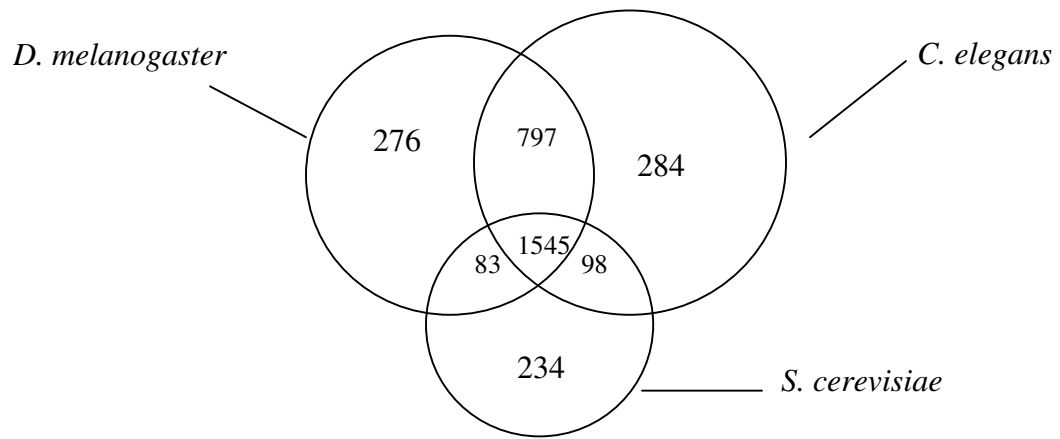


FIGURE 2.

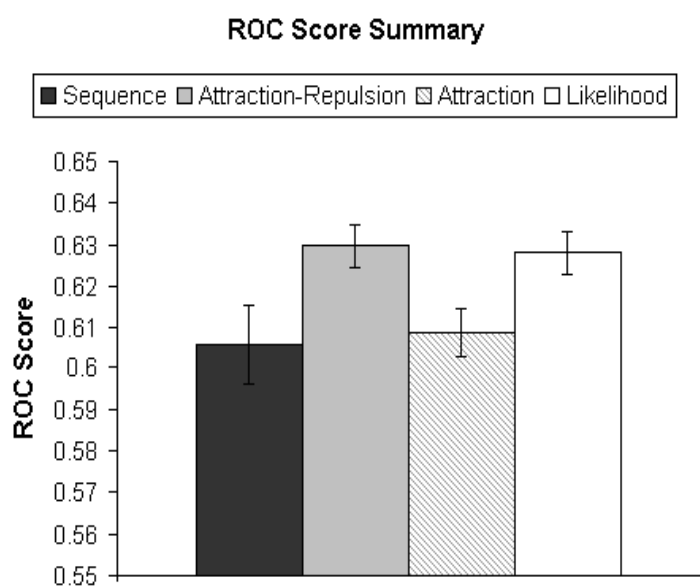


FIGURE 3.

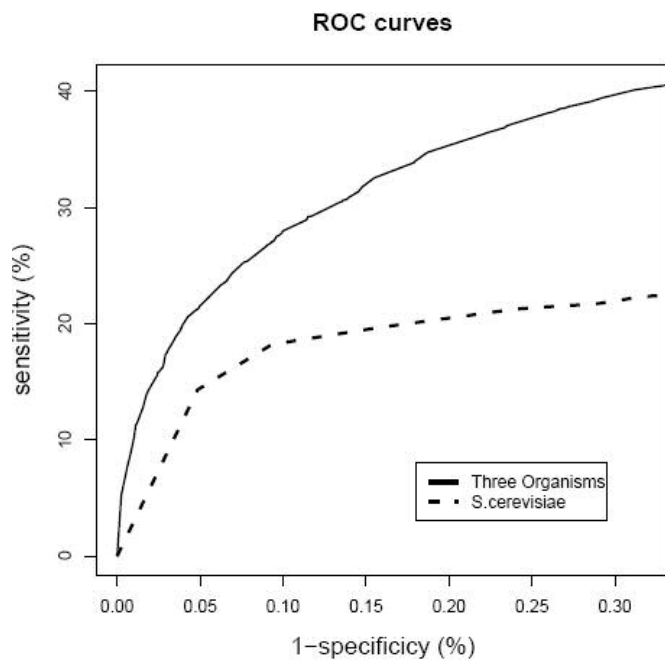
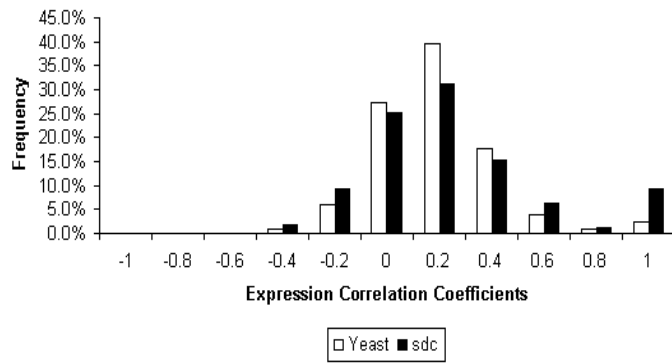


FIGURE 4.

Distribution of Pearson correlation coefficients of gene expression (Cell Cycle)



Distribution of Pearson correlation coefficients of gene expression (Rosetta)

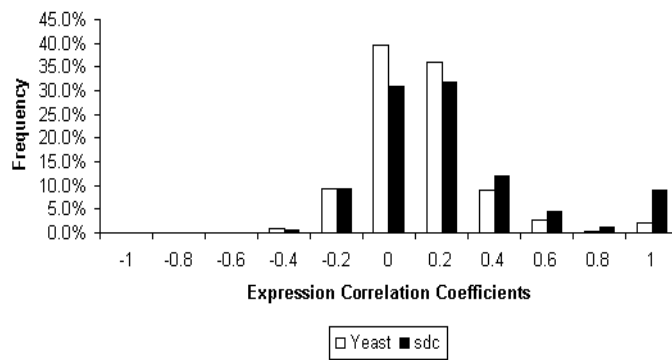


Table 1. Comparison of the Mean Correlation Coefficient for the Selected Predicted Protein Pairs based on two different Information Sources. sdc, prediction based on the information from three organisms *S. cerevisiae*, *D. melanogaster* and *C. elegans*.

Pairs	Cell Cycle			Rosetta		
	Mean(sdc)	Mean (<i>S.cerevisiae</i>)	Empirical p-value	Mean(sdc)	Mean (<i>S.cerevisiae</i>)	Empirical p-value
100	0.28	0.10	0	0.20	0.05	0
500	0.21	0.09	0	0.17	0.04	0
1000	0.15	0.09	0	0.13	0.03	0.0001
2000	0.12	0.08	0.0055	0.08	0.03	0.0129
5000	0.10	0.08	0.0255	0.06	0.02	0.0250
10000	0.09	0.07	0.0264	0.05	0.01	0.0258

Table 2. The top 10 predicted interacting protein pairs that are not included in the MIPS physical interaction dataset. All these pairs have estimated interaction probability equal to 1. Each row represents an interacting protein pair with their corresponding annotated functions. The protein function annotations are obtained from CYGD (Comprehensive Yeast Genome Database).

Protein I	Function	Protein II	Function
MCD1	mitotic Chromosome Determinant	IRR1	nuclear cohesin protein
ECM31	involved in cell wall biogenesis and architecture	VPS9	required for Golgi to vacuole trafficking
CUP2	copper-dependent transcription factor	THI4	involved in thiamine biosynthesis and DNA repair
BUL2	ubiquitin-mediated protein degradation	SRP1	karyopherin-alpha or importin
DCS1	scavenger mRNA decapping enzyme	NTH2	neutral trehalase
SNZ1	member of the stationary phase-induced gene family, involved in response to cell stress	SNZ1	member of the stationary phase-induced gene family, involved in response to cell stress
YMR009W	unknown function localised to cytoplasm and nucleus	FUN34	integral membrane protein, involved in ammonia production
OCA1	putative protein tyrosine phosphatase	OCA1	putative protein tyrosine phosphatase
ECM15	involved in cell wall biogenesis and architecture	VAC8	required for vacuole inheritance and protein targeting from the cytoplasm to vacuole
SPC2	signal peptidase 18 KD subunit	URA3	orotidine-5'-phosphate decarboxylase