Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data

Naoki Nariai[1]*, Eric D. Kolaczyk[2], Simon Kasif[1,3]

1 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America, 2 Department of Mathematics and Statistics, Boston University, Boston, Massachusetts, United States of America, 3 Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America

Dramatic improvements in high throughput sequencing technologies have led to a staggering growth in the number of predicted genes. However, a large fraction of these newly discovered genes do not have a functional assignment. Fortunately, a variety of novel high-throughput genome-wide functional screening technologies provide important clues that shed light on gene function. The integration of heterogeneous data to predict protein function has been shown to improve the accuracy of automated gene annotation systems. In this paper, we propose and evaluate a probabilistic approach for protein function prediction that integrates protein-protein interaction (PPI) data, gene expression data, protein motif information, mutant phenotype data, and protein localization data. First, functional linkage graphs are constructed from PPI data and gene expression data, in which an edge between nodes (proteins) represents evidence for functional similarity. The assumption here is that graph neighbors are more likely to share protein function, compared to proteins that are not neighbors. The functional linkage graph model is then used in concert with protein domain, mutant phenotype and protein localization data to produce a functional prediction. Our method is applied to the functional prediction of *Saccharomyces cerevisiae* genes, using Gene Ontology (GO) terms as the basis of our annotation. In a cross validation study we show that the integrated model increases recall by 18%, compared to using PPI data alone at the 50% precision. We also show that the integrated predictor is significantly better than each individual predictor. However, the observed improvement vs. PPI depends on both the new source of data and the functional category to be predicted. Surprisingly, in some contexts integration hurts overall prediction accuracy. Lastly, we provide a comprehensive assignment of putative GO terms to 463 proteins that currently have no assigned function.

Citation: Nariai N, Kolaczyk ED, Kasif S (2007) Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data. PLoS ONE 2(3): e337. doi:10.1371/journal.pone.0000337

INTRODUCTION

Functional annotation of genes is a fundamental problem in computational and experimental biology. The problem can be solved at various levels of resolution ranging from identifying high level processes where a given protein might be associated with, to discovery of the cell specific protein-ligand interaction targets of a protein in different biological conditions. The most established and reliable methods for protein function prediction are based on sequence similarity using BLAST [1] and profile methods such as PFAM [2], and PSI-BLAST [1]. Other still evolving methods that are too numerous to list include gene fusion information [3], and phylogenetic profiling [4,5]. Emergent methods that elucidate function from a variety of high-throughput experimental screens have become particularly attractive recently due to the reduced cost of conducting genome-wide functional screens. Genomic and proteomic data sets, including gene expression and protein-protein interaction (PPI) data, are becoming increasingly available for a growing array of organisms. Driven by the hypothesis that co-expressed genes might participate in related biological processes, clustering gene expression profiles across diverse conditions can be used to assign protein function [6–8]. Using PPI data to assign protein function has been extensively studied. These algorithms are often based on the "guilt by association" principle that suggests that interacting neighbors in protein-protein interaction (PPI) networks might also share a function [9–11]. Since such genome-wide data sets are inherently noisy, and each type of data captures only one aspect of cellular activity (e.g. gene expression data measure mRNA levels of transcriptionally induced genes, and PPI data suggest a feasible physical interaction between proteins), it is appealing to combine such heterogeneous data in an effort to improve the coverage and accuracy of protein function prediction.

Bayesian network methodologies for data integration have been explored [12–14] in a number of systems for predicting protein-protein interactions and protein function similarity. These approaches calculate the posterior probability that each pair of genes *i* and *j*, has a functional relationship, given the various types of genome-wide data. These algorithms output a functional linkage graph [3,15] in which an edge between two nodes (genes) represents functional similarity with a reliability score (probability) assigned to each edge. However, using these probabilistic networks to produce a functional assignment remains a hard computational problem. For instance, one approach for protein function annotation based on Markov random fields (MRFs) has been previously investigated [10]. An integrated MRF approach that includes network structures (PPI network and co-expression network) and protein domain information to predict protein function has also been proposed [16]. There, the authors used

Academic Editor: John Storey, University of Washington, United States of America

Received December 13, 2006; **Accepted** March 8, 2007; **Published** March 28, 2007

Funding: This work was supported by NSF grant ITR-048715 and NHGRI grant R01 HG003367-01A1.

Competing Interests: The authors have declared that no competing interests exist.

*** To whom correspondence should be addressed.** E-mail: nariai@bu.edu

Gibbs sampling to estimate the probability that a protein has a particular function. Machine learning methods based on support vector machines have been investigated in several projects [17,18]. In fact, it is rather obvious that if we treat the prediction of function based on each modality as an expert, then any of the popular classification methods (decision trees, boosting, and weighted majority) can in principle be used for "integration" of these predictions. However, given the currently sparse data using complex representations for prediction might lead to overfitting.

In this paper, our contribution is twofold. First, we propose a simple and relatively transparent probabilistic model for protein function prediction that allows us to efficiently calculate the posterior probability that each gene has a particular function, given various types of genome-wide data. Second, we analyze the effect of combining the heterogeneous data sources in a substantially more comprehensive manner than has been done to date, with the goal of better understanding just which types of genes benefit most from the integration of which types of data sources. In particular, we develop a relatively simple yet useful method to integrate functional linkage graphs with categorical information. The functional linkage graphs are constructed from PPI data and gene expression data. As usual the assumption here is that physically interacting proteins or co-expressed genes are more likely to share protein functions than a randomly selected pair of proteins [10]. Categorical features for each protein, including protein motifs, knockout phenotype, and localization information are captured based on predictive sources of evidence available from the MIPS database [19]. Using Bayesian networks framework, this categorical information is then combined with functional linkage graphs constructed from PPI data and gene expression data to generate functional predictions. Our method is applied to the functional prediction of proteins in yeast (*Saccharomyces cerevisiae*). Our methodology combines PPI data, gene expression data, protein motif information, mutant phenotype data, and protein localization data, while using Gene Ontology (GO) "biological processes" terms [20] as the basis for functional annotation. The long term goal of this research is to develop a probabilistic language to specify which proteins might be active in a given biological process based on the type of interacting partners they have, protein motifs, or transcriptional profiles.

By combining five types of data, the number of correctly recovered known gene-term associations is increased by 18% at the same precision (50%), compared to using PPI data alone. We specifically focused on certain points on the ROC curve in our analysis that we believe are potentially feasible for follow-ups on the prediction in experimental labs. We show that by adding different types of genome-wide data, different types of the GO terms that are specific for the type of information are newly recovered. Also, by conducting robustness analysis of the integration model to PPI edge removal, we provide a novel perspective on the amount of PPI data necessary to obtain high prediction accuracy by the integration model. In that analysis, we find some conditions where integration actually hurts performance rather than improving accuracy. Plausible functions are assigned to 463 currently unannotated proteins by our method, and we discuss some of these novel assignments.

METHODS

2.1 Data preparation

2.1.1 Protein-protein interaction data From the GRID database [21], 31201 non-redundant protein-protein interactions among 5151 yeast *Saccharomyces cerevisiae* genes are extracted. We eliminated self-self interactions and duplicated protein interaction pairs from the database to construct a PPI functional linkage graph [3,15], in which an edge between two nodes (proteins) represents evidence for protein function similarity.

2.1.2 Gene expression data Four types of gene expression data, the Rosetta compendium data [22], cell cycle data [23], stress-response data [24], and DNA-damage data [25] are used in this paper. For each type of gene expression data, Pearson correlation coefficients for all combinations of genes are calculated, and gene pairs whose correlation coefficient is larger than 0.85 are selected as "co-expressed pairs" for each type of gene expression data. We obtained 1783, 645, 10654 and 31827 gene pairs from the Rosetta data, cell cycle data, stress-response data and DNA damage data, respectively. The false discovery rate (FDR) [26,27] for each threshold is less than 10^{-10}, indicating that for each experiment we are only using a set of declared co-expressed pairs for which a false declaration is exceedingly unlikely. Finally, 38151 non-redundant gene pairs are obtained from the combined gene pairs to construct a co-expression functional linkage graph.

2.1.3 Protein motif information From the MIPS [19] database, 2678 protein-motif associations (e.g. YCR065W protein has "Fork head domain signatures and profile" motif) are extracted, covering 2179 proteins across 992 motif categories. If a protein has a specific protein motif, this can increase the probability that the protein has a specific protein function. We describe how to integrate this category information for protein function prediction in Section 2.3.

2.1.4 Gene knock-out phenotype data From the MIPS [19] database, 3013 protein-phenotype associations (e.g. YPR185W deletion mutant exhibits "Starvation sensitivity") are obtained, covering 1460 proteins across 175 mutant phenotype categories.

2.1.5 Protein localization data From the MIPS [19] database, 5191 protein-localization data (e.g. YPR191W protein localizes at "Mitochondrial inner membrane") are obtained, covering 4076 proteins across 41 cellular location categories.

2.1.6 GO term data From the 06/03/2006 version of the Yeast SGD database [28], 107636 gene-term GO assignments are obtained, in which there are 6289 genes and 1965 'biological process' terms in total. For each gene-term association, we expanded the label in the GO hierarchy to include all 'is-a' and 'part-of' ancestors of each GO term. Labels that appear more than 300 times among the 6289 genes are excluded for further analysis, on the assumption that such terms are too broad for protein function prediction. Also labels that appear less than five times among the genes are excluded, since they do not constitute a sufficiently large enough sample to make reliable predictions.

From PPI data and gene expression data, two different functional linkage graphs are obtained. Here, an edge in each functional linkage graph shows that the two nodes (proteins) are a member of the constructed pairs in each data set. For each GO label t and for each functional linkage graph l, we calculate $p_1^{(l)}$, the probability that a protein has label t, given that the interacting partner has label t. This $p_1^{(l)}$ is expected to be higher than $p_0^{(l)}$, the probability that the protein has label t, given that the interacting partner does not have label t. Here, a X^2 test was performed to ensure that $p_1^{(l)}$ and $p_0^{(l)}$ were statistically different using a Bonferroni-corrected p-value of $0.001/T$, where T is the number of terms tested in each data set.

2.2 Categorical features of proteins

Proteins can be associated with categorical features according to different types of categorical information. The categorical features that are used in our predictive methodology are defined below.

- Protein motif (domain): Random variable d_i is associated with a protein where $d_i = 1$ if the protein contains domain d_i, and $d_i = 0$ otherwise. A feature vector $\mathbf{d} = (d_1, d_2,\ldots, d_{qd})^T$ is defined for each protein, where q_d is the total number of protein motif features ($q_d = 992$ in our case).
- Phenotype: Random variable p_i is associated with a protein where $p_i = 1$ if the gene knockout exhibits phenotype p_i, and $p_i = 0$ otherwise. A feature vector $\mathbf{p} = (p_1, p_2,\ldots, p_{qp})^T$ is defined for each protein, where q_p is the total number of phenotype features ($q_p = 175$ in our case).
- Protein localization: Random variable l_i is associated with a protein where $l_i = 1$ if the protein localizes in l_i, and $l_i = 0$ otherwise. A feature vector $\mathbf{l} = (l_1, l_2,\ldots, l_{ql})^T$ is defined for each protein, where q_l is the total number of localization features ($q_l = 41$ in our case).

Naturally, a protein can have several features at the same time. Our aim is to integrate these sources of evidence in a smooth fashion to improve the accuracy and coverage of the functional predictors based on the assumption that if a protein has specific features, then this can increase the probability to infer specific protein functions.

2.3 Computing the posterior probability of function using graphs and features

For each protein i and GO term t, a Boolean random variable $L_{i,t}$ is associated, where $L_{i,t} = 1$ if i is labeled with t, and $L_{i,t} = 0$ otherwise. We want to calculate the probability of $L_{i,t} = 1$ for all combinations of i and t, given the structure of functional linkage graphs constructed above, and the category features that the protein i has, and all the assignments of GO terms to the other proteins. We assume that probability distribution for the labeling $L_{i,t} = 1$ is conditionally independent of all other nodes given the functional annotation of the neighbors and category information of the protein.

We want to calculate the posterior probability given functional linkage graphs and category features of a protein $P(L_{i,t} = 1 \mid \mathcal{N}_i^{(1)},\ldots, \mathcal{N}_i^{(m)}, k_{i,t}^{(1)},\ldots, k_{i,t}^{(m)}, \mathbf{c}_i^{(1)},\ldots, \mathbf{c}_i^{(n)})$, where $\mathcal{N}_i^{(l)} (l = 1,\ldots, m)$ is the number of graph neighbors (excluding unannotated neighbors) of gene i in a functional linkage graph l, $k_i^{(l)} (l = 1,\ldots, m)$ is the number of the neighbors of gene i which are labeled with term t in the graph l, m is the number of different types of functional linkage graphs, $\mathbf{c}_i^{(j)} (j = 1,\ldots, n)$ is the feature vector that the gene i has for a category feature type j, and n is the number of different types of categories (not the number of features). For example, in this paper, $\mathcal{N}_i^{(1)}$ and $k_i^{(1)}$ is the number of neighbors of gene i, and the neighbors that have term t in a PPI network, respectively. $\mathcal{N}_i^{(2)}$ and $k_i^{(2)}$ is the number of the neighbors and the neighbors that have term t in a co-expression network, respectively. $\mathbf{c}_i^{(1)}$, $\mathbf{c}_i^{(2)}$ and $\mathbf{c}_i^{(3)}$ are feature vectors that gene i has for the three types of categories, i.e., protein motif feature vector $\mathbf{d}$, mutant phenotype feature vector $\mathbf{p}$, and localization feature vector $\mathbf{l}$, respectively (Section 2.2).

Applying Bayes' theorem, the posterior probability that gene i has function t $P(L_{i,t} = 1 \mid \mathcal{N}_i^{(1)},\ldots, \mathcal{N}_i^{(m)}, k_{i,t}^{(1)},\ldots, k_{i,t}^{(m)}, \mathbf{c}_i^{(1)},\ldots, \mathbf{c}_i^{(n)})$ can be rewritten (with omitting subscript i and t) as:

$$
\begin{aligned}
&P(L|N^{(1)},\ldots,N^{(m)},k^{(1)},\ldots,k^{(m)},\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(n)})\\
&=\frac{P(k^{(1)},\ldots,k^{(m)},\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(n)}|L,N^{(1)},\ldots,N^{(m)})\cdot P(L|N^{(1)},\ldots,N^{(m)})}{P(k^{(1)},\ldots,k^{(m)},\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(n)}|N^{(1)},\ldots,N^{(m)})}\\
&=\frac{\prod_{l=1}^{m}P(k^{(l)}|L,N^{(1)},\ldots,N^{(m)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|L,N^{(1)},\ldots,N^{(m)})\cdot P(L)}{P(L)\cdot P(k^{(1)},\ldots,k^{(m)},\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(n)}|L,N^{(1)},\ldots,N^{(m)})+P(\overline{L})\cdot P(k^{(1)},\ldots,k^{(m)},\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(n)}|\overline{L},N^{(1)},\ldots,N^{(m)})} \qquad (1)\\
&=\frac{\prod_{l=1}^{m}P(k^{(l)}|L,N^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|L)\cdot P(L)}{P(L)\cdot\prod_{l=1}^{m}P(k^{(l)}|L,N^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|L)+P(\overline{L})\cdot\prod_{l=1}^{m}P(k^{(l)}|\overline{L},N^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|\overline{L})}.
\end{aligned}
$$

Here, we assume that the probability distribution of $L_{i,t}$ is independent of the number of graph neighbors $\mathcal{N}^{(1)},\ldots, \mathcal{N}^{(m)}$, hence $P(L|\mathcal{N}^{(1)}, \ldots, \mathcal{N}^{(m)}) = P(L)$. Also, we assume that $k^{(1)},\ldots, k^{(m)}$, $\mathbf{c}^{(1)},\ldots, \mathbf{c}^{(n)}$ are conditionally independent of each other, given L and $\mathcal{N}^{(1)},\ldots, \mathcal{N}^{(m)}$. This assumption is similar to the case in a Naive Bayes classifier. It is natural to assume that $k^{(l)}$ is conditionally independent of $\mathcal{N}^{(1)},\ldots, \mathcal{N}^{(m)}$ except $\mathcal{N}^{(l)}$, given L. Also, $\mathbf{c}^{(j)}$ is conditionally independent of $\mathcal{N}^{(1)},\ldots, \mathcal{N}^{(m)}$, given L. Now we need to calculate each decomposed product in (1).

$P(k^{(l)} \mid L,\mathcal{N}^{(l)})$ is the probability that $k^{(l)}$ neighbors are labeled with t out of $\mathcal{N}^{(l)}$ neighbors in a graph l, given that the gene i is labeled with t. Here we assume a binomial distribution [10] and calculate this probability as $P(k^{(l)}|L,N^{(l)})=B(N^{(l)},k^{(l)},p_1^{(l)})=\binom{N^{(l)}}{k^{(l)}}\cdot(p_1^{(l)})^{k^{(l)}}\cdot(1-p_1^{(l)})^{N^{(l)}-k^{(l)}}$, where $p_1^{(l)}$ is the probability that a protein i has label t, given that an interacting partner has label t within a functional linkage graph l, which is pre-calculated by training data (Section 2.1.6). Similarly, $P(k^{(l)}|\overline{L},N^{(l)})=B(N^{(l)},k^{(l)},p_0^{(l)})=\binom{N^{(l)}}{k^{(l)}}\cdot(p_0^{(l)})^{k^{(l)}}\cdot(1-p_0^{(l)})^{N^{(l)}-k^{(l)}}$, where $p_0^{(l)}$ is the probability that a protein i has label t, given that an interacting partner does not have label t within a functional linkage graph l.

$P(\mathrm{c}^{(j)} \mid L)$ is the probability that a gene i has feature vector $\mathrm{c}^{(j)}$ given that the gene i has term t. $P(L)$ is the prior probability that the gene i has term t. This is calculated as $P(L) = f$, where f is the frequency of term t among genes.

Hence the neighborhood function (1) becomes:

$$
(1)=\frac{f\cdot\prod_{l=1}^{m}B(N^{(l)},k^{(l)},p_1^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|L)}{f\cdot\prod_{l=1}^{m}B(N^{(l)},k^{(l)},p_1^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|L)+\overline{f}\cdot\prod_{l=1}^{m}B(N^{(l)},k^{(l)},p_0^{(l)})\cdot\prod_{j=1}^{n}P(\mathbf{c}^{(j)}|\overline{L})}
$$

$$
=\frac{f\cdot\prod_{l=1}^{m}\frac{B(N^{(l)},k^{(l)},p_1^{(l)})}{B(N^{(l)},k^{(l)},p_0^{(l)})}\cdot\prod_{j=1}^{n}\frac{P(\mathbf{c}^{(j)}|L)}{P(\mathbf{c}^{(j)}|\overline{L})}}{f\cdot\prod_{l=1}^{m}\frac{B(N^{(l)},k^{(l)},p_1^{(l)})}{B(N^{(l)},k^{(l)},p_0^{(l)})}\cdot\prod_{j=1}^{n}\frac{P(\mathbf{c}^{(j)}|L)}{P(\mathbf{c}^{(j)}|\overline{L})}+\overline{f}}=\frac{f\cdot\prod_{l=1}^{m}\alpha^{(l)}\cdot\prod_{j=1}^{n}\beta^{(j)}}{f\cdot\prod_{l=1}^{m}\alpha^{(l)}\cdot\prod_{j=1}^{n}\beta^{(j)}+\overline{f}}, \qquad (2)
$$

where $\alpha^{(l)}=B(N^{(l)},k^{(l)},p_1^{(l)})/B(N^{(l)},k^{(l)},p_0^{(l)})$, and $\beta^{(j)}=\frac{P(\mathbf{c}^{(j)}|L)}{P(\mathbf{c}^{(j)}|\overline{L})}=\prod_{x}\frac{P(c_x^{(j)}|L)}{P(c_x^{(j)}|\overline{L})}$, assuming conditional independence between the feature vectors. Here, $P(c_x^{(j)} = 1 \mid L) =$ (# of t-labeled genes that have a feature $c_x^{(j)}$)/(# of t-labeled genes), and $P(c_x^{(j)} = 1 \mid \bar{L}) =$ (# of genes that are not labeled with t and have a feature $c_x^{(j)}$)/(# of genes that are not labeled with t). Since we do not want lack of information to affect protein function prediction, we assume $\frac{P(L|c_x^{(j)}=0)}{P(\overline{L}|c_x^{(j)}=0)}=\frac{P(c_x^{(j)}=0|L)}{P(c_x^{(j)}=0|\overline{L})}\cdot\frac{P(L)}{P(\overline{L})}=\frac{P(L)}{P(\overline{L})}$, hence $P(c_x^{(j)} = 0 \mid L)/P(c_x^{(j)} = 0 \mid \bar{L}) = 1$. For example, suppose that the current genome-wide data do not contain information that a protein is localized in mitochondria. However, this does not necessary mean that the protein does not localize in mitochondria. The reason might be that just lack of the currently available data.

RESULTS

The integration algorithm described in the Methods section is evaluated on the task of predicting protein functions for *Saccharomyces cerevisiae*. The algorithm fuses probabilities obtained from diverse data sources including PPI, gene expression, protein motif information, gene knock-out phenotype, and protein localization. The functional annotations used to train our models are based on the GO category and are obtained from Yeast SGD database [28]. The algorithm is first validated on known protein-term associations by a 5-fold cross validation methodology. We also conduct robustness analysis to understand the effect of removal of PPI edges on the accuracy of the prediction with one or more data sources. Finally, we predict protein function of unannotated genes.

3.1 Cross Validation Analysis of Prediction Accuracy

First, we attempted to predict known protein-term associations by 5-fold cross validation. For each gene g and term t, the probability that gene g has term t is calculated based on equation (2), given that we know every other gene-term associations in the training set. It is predicted that gene g has term t if the probability exceeds a specified threshold. A positive g-t association set is obtained from the GO "biological processes" data, and negative g-t association set is defined as follows: If the association is not in the positive set, and g is annotated with at least one biological process t, and t is neither ancestor nor descendant of the known function in the GO hierarchy. Figure 1 presents a ROC curve of function prediction by different combinations of each data sources. Sensitivity is defined as #TP/(#TP+#FN), which corresponds to recall, and specificity is defined as #TN/(#FP+#TN), which corresponds to precision. For varying posterior probability cut-off, 1-specificity and sensitivity is plotted. The result shows that by combining the five specific types of data described above, protein functions can be predicted more accurately, compared to each data source alone.

Figure 2 summarizes the impact that data integration has on protein function prediction sensitivity at a fixed precision (50% and 80%). The error bar shows the standard deviation of 10 independent cross-validation experiments. At the 50% precision, 14906 known protein-term associations can be recovered on average by combining five types of data. On the other hand, when we use PPI data alone, 12662 associations can be recovered on average. Our integrated method thus realizes an 18% increase in the number of functional predictions for genes at the 50% precision. At the 80% precision, the combination of all data (PPI, gene expression, protein motif, mutant phenotype, and protein localization) works better than any other combinations and other single source of data. However, at the 80% precision, combining PPI data with one other data source shows a little improvement in predictive accuracy, suggesting that PPI data is particularly informative.

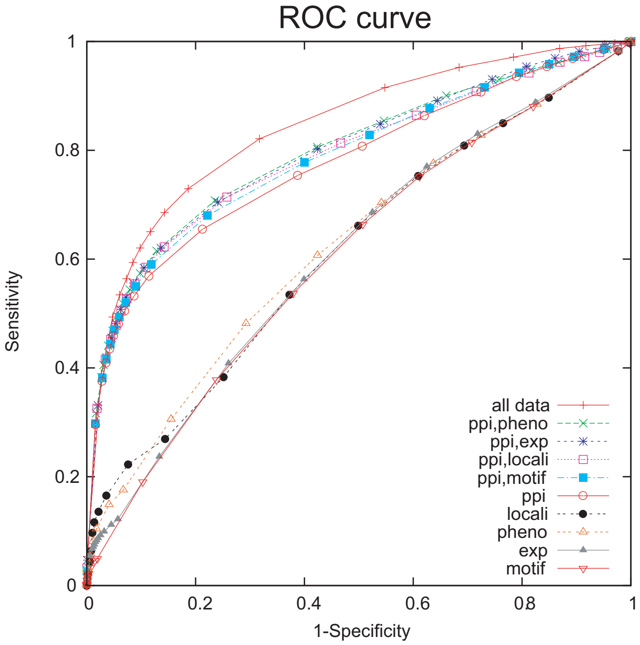

Figure 1. The ROC curve of recall experiment by 5-fold cross validation. Sensitivity is defined as #TP/(#TP+#FN), and specificity is defined as #TN/(#FP+#TN).
doi:10.1371/journal.pone.0000337.g001

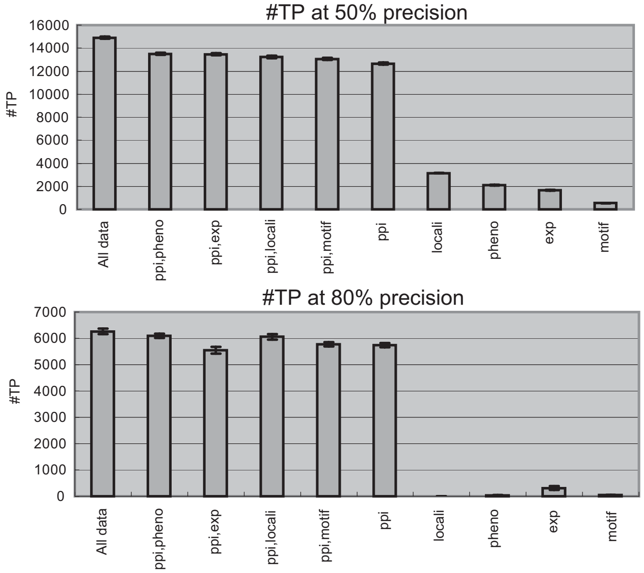

Figure 2. #TP at 50% precision (Upper) and #TP at 80% precision (Lower). Here, *ppi*, *exp*, *motif*, *pheno*, and *locali* corresponds to PPI, gene expression, protein motif, phenotype, and localization data, respectively.
doi:10.1371/journal.pone.0000337.g002

These two levels of precision, i.e., 50% and 80%, were chosen as being reasonably representative of the range of possible improvements observed in our study. In addition to the performance characteristics just described, we also examined the issue of falsely predicted proteins, as a function of the threshold applied to posterior probabilities. Using the method of [29], the rate of false discoveries, for the classifier integrating all data sources, was estimated to be 0.13 and 8.4×10^{-6}, respectively, at the 50% and 80% precision levels.

Next, we analyzed whether prediction accuracy depends upon the functional category to be predicted. It is expected that the prediction performance of specific GO terms depends on what kinds of data sources one uses. Table S1 (in Supporting Information) shows the list of GO terms, which are improved by adding gene expression data in addition to PPI data at the 50% precision for each GO term prediction. Here GO terms are listed, of which the number of #TP is increased by at least 10, compared to that of using PPI data alone. We can see from the result that many of the improved terms are metabolism related (e.g. "amino acid and derivative metabolism", "nitrogen compound biosynthesis", and etc.). Since metabolic reactions are often interactions between enzymes and compounds, and such proteins (enzymes) do not necessarily have protein-protein interactions between enzymes in a same pathway, it might be difficult to reconstruct and hence predict such metabolic pathways by using PPI data alone. In this sense, measuring gene expression of enzymes and identifying co-expressed genes will be supplementary information for capturing functionally related genes. Here, the result suggests that the gene expression data actually helps to identify such metabolic components that are working in a same pathway. Table S2 shows the list of GO terms, which are improved by adding protein motif information in addition to PPI data. Here, it is interesting to see

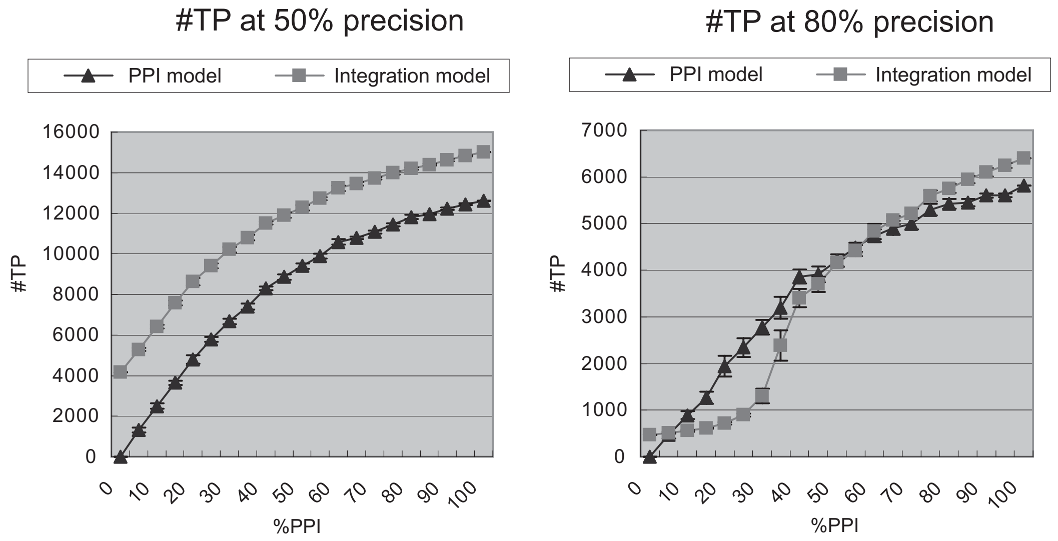

Figure 3. #TP at 50% precision (Left) and #TP at 80% precision (Right) with varying amount of PPI edges. At 50% precision, the integration model always wins over the PPI model. However, at 80% precision, the integration model wins only when more than 50% of original PPI edges are present.
doi:10.1371/journal.pone.0000337.g003

that GO terms "phosphorylation" and "phosphate metabolism" are most strikingly improved. Also, among newly recovered genes that have the GO term "transcription from RNA polymerase II promoter" by adding motif information, nine proteins have a protein motif "Zinc finger, C2H2 type, domain". Since protein kinases and transcription factors often have specific binding domains, protein motif information are particularly useful for predicting these terms. Table S3 shows the list of GO terms, which are improved by adding phenotype data in addition to PPI data. Among the improved GO terms, "cell wall organization and biogenesis", "cell budding", "reproduction", and "external encapsulating structure organization and biogenesis" might be related to phenotypes of a cell, and only listed in this table. Table S4 shows the list of GO terms, which are improved by adding localization data in addition to PPI data. Among the improved GO terms, "ion transport", "Golgi vesicle transport", and "vesicle-mediated transport" are cellular location specific GO terms, and these terms are only listed in this table. We can conclude from these results that by adding different types of genome-wide data, different types of GO terms that are specific for the data type can newly be predicted.

Here is an example how the combination of different types of data helps to predict protein function more specifically. Genes YKR055W, YIL118W and YJL128C have a GO term "intracellular signaling cascade", but neither the PPI data nor the protein motif information alone can predict the GO term for the proteins. When PPI data alone is used, a GO term "signal transduction", which is a parent of "intracellular signaling cascade" in the GO hierarchy and hence a broader term, can be predicted. However, when both PPI data and protein motif information are used, the GO term can be predicted correctly. In this case, information that the proteins have a protein motif "protein kinases signatures and profile" or "prenyl group binding site" helps to predict more specific term "intracellular signaling cascade" correctly.

3.2 Robustness analysis of the integration model

In the recall experiment in Section 3.1, we showed that PPI data is the strongest source of evidence for protein function prediction in our model, compared to other data sources. Here, we want to know whether our integration model works well or not when the amount of PPI data is limited. In this experiment, a certain fraction of the PPI edges are randomly removed from the original PPI network, and then protein function is predicted using our integration model. Figure 3 shows the result of prediction at 50% precision (left) and 80% precision (right). Here, *x*-axis shows how much of PPI edges are present, compared to the original PPI network. For example, at $x = 50$, half of PPI edges are randomly removed from the original PPI network. The error bar shows the standard deviation of 10 independent experiments. At 50% precision, the integration model always wins, regardless of the number of PPI edges present. However, interestingly, at 80% precision, the integration model wins only when more than 50% of PPI edges are present. In other words, in order to obtain high prediction accuracy (80% precision) by the integration model, certain amount of PPI data is necessary (more than 50% of original PPI edges in this case). This result suggests that the combination of gene expression, protein motif, mutant phenotype and protein localization data is still a weak indicator of protein function, and hence need to have certain amount of PPI information (strong indicator of protein function) in order to obtain high prediction accuracy (80% precision).

3.3 Prediction of function unknown proteins

By integrating five types of data, we assign plausible GO terms to 463 proteins among 1481 currently unannotated yeast *Saccharomyces cerevisiae* proteins (complete list available in Supporting Information, Table S5). The threshold probability for function annotation is 0.470, in which we expect 50% precision for the protein function prediction from the cross validation experiment in Section 3.1.

Among the predicted function of unannotated proteins, recent literature reported [30] that YBL028C, YBR271W, YCR016W, YJR003C, YDL167C, YDR361C, YIL096C, YIL127C, YLR449W, YMR310C, YNL022C, YNL132W, YNL175C, YGR187C, YGR283C and YOR021C as rRNA and ribosome biosynthesis (RRB) regulon. It has also been reported [31] that YLR051C encodes a protein involved in pre-rRNA processing, confirming our prediction of "ribosome biogenesis" (or related terms). Other than ribosomal proteins, recent literature reported that YAL053W participates in a cell wall biosynthesis process [32]. Since our prediction for YAL053W is "cell wall organization and biogenesis", we can say that there is an experimental validation for the prediction. Also, it is reported [33] that YBR280C encodes a protein, which targets Aah1p for proteasome-dependent degradation. Here, our prediction for the protein "SCF-dependent proteasomal ubiquitin-dependent protein catabolism" is quite consistent with the literature.

It is confirmed here that 20 out of 463 function predictions for unannotated proteins are quite consistent with the conclusion from the recent publications. We expect that many of our predictions will turn out to be true after validation experiments.

All the biological data and a Perl program used in this analysis are available at: http://genomics10.bu.edu/nariai/yeast_func/.

DISCUSSION

In this paper, we propose a probabilistic method to predict protein function from multiple types of genome-wide data. Pair-wise information between proteins, such as PPI data or co-expression information is converted into a functional linkage graph, in which an edge between nodes represents evidence for protein function similarity. Category information, such as protein motif information, mutant phenotype data, and protein localization data is combined with the functional linkage graphs using a unified probabilistic framework. We showed in our 5-fold cross validation experiment that our method successfully improved prediction accuracy and coverage by integrating five types of genome-wide data. Also, by conducting robustness analysis of the integration model to PPI edge removal, we showed that there is a certain amount of PPI data necessary to obtain high prediction accuracy by the integration model. We proposed functional predictions for 463 currently unannotated proteins. One subjective aspect of our method is in the choice 0.85 in thresholding the correlation coefficients in constructing our co-expression functional linkage graph. However, we have found our results to be quite robust to this choice; for example, even much higher thresholds yield qualitatively quite similar results. In principle, a more objective choice of threshold could be made through the use of cross-validation, but this would come at the cost of an increased computational burden. Other limitations are that we assume probabilistic conditional independence between different types of functional linkage graphs and each informational category. Of course, this assumption might not always be correct in a biological sense. For example, some of physically interacting protein pairs are also co-expressed. However, previous literature has reported that Naive Bayes frequently tends to work well, and frequently better than more sophisticated classifiers, when the data are sparse compared to the dimensionality of the problem, even when the features (e.g., in our case, the functional linkage graphs and category feature vectors) are not truly conditionally independent [34,35]. Hence, we anticipate that our method may in fact prove to be a fairly strong contender in competition, for the types of data we use, with more sophisticated methods that may follow. In addition, we assume independence between non-ancestral GO terms. Since the GO terms comprise a hierarchical structure, and there would be dependencies among the GO terms, one might want to take the dependency between the nodes into account. Also in our method, within a functional linkage graph, non-neighbors (two nodes whose distance are more than one) are not considered for functional similarity. Application of a Markov Random Field model in conjunction with belief propagation and/or sampling may address these limitations and is the subject of on-going investigation.

Although the result presented here is a case study for yeast *S. cerevisiae*, we believe that similar advances are possible for fly, worm, mouse and human where analogous resources are being compiled.

SUPPORTING INFORMATION

Table S1 Improved GO terms by adding gene expression data at 50% precision.
Found at: doi:10.1371/journal.pone.0000337.s001 (0.02 MB XLS)

Table S2 Improved GO terms by adding protein motif data at 50% precision.
Found at: doi:10.1371/journal.pone.0000337.s002 (0.02 MB XLS)

Table S3 Improved GO terms by adding phenotype data at 50% precision.
Found at: doi:10.1371/journal.pone.0000337.s003 (0.02 MB XLS)

Table S4 Improved GO terms by adding localization data at 50% precision.
Found at: doi:10.1371/journal.pone.0000337.s004 (0.02 MB XLS)

Table S5 Prediction result of unannotated genes. N1 is the number of neighbors in PPI network, k1 is the number of t-labeled (t is the predicted GO term) neighbors in PPI network, N2 is the number of neighbors in co-expression network, and k2 is the number of t-labeled neighbors in co-expression network.
Found at: doi:10.1371/journal.pone.0000337.s005 (0.23 MB XLS)

ACKNOWLEDGMENTS

We thank Dr. Lingang Zhang, Dr. John Rachlin and Dr. T. M. Murali for helpful discussion, and anonymous reviewers for helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: SK NN EK. Performed the experiments: NN. Analyzed the data: SK NN. Contributed reagents/materials/analysis tools: NN. Wrote the paper: SK NN EK.

REFERENCES

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
2. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247–251.
3. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86.
4. Gaasterland T, Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 3: 199–217.
5. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
6. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. pp 418–429.
7. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
8. Zhou X, Kao MC, Wong WH (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A 99: 12783–12788.

9. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, et al. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci U S A 101: 2888–2893.
10. Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19 Suppl 1: i197–204.
11. Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nat Biotechnol 18: 1257–1261.
12. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
13. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M (2005) Assessing the limits of genomic data integration for predicting protein networks. Genome Res 15: 945–953.
14. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A 100: 8348–8353.
15. Yanai I, DeLisi C (2002) The society of genes: networks of functional links between genes from comparative genomics. Genome Biol 3: research0064.
16. Deng M, Chen T, Sun F (2004) An integrated probabilistic model for functional prediction of proteins. J Comput Biol 11: 463–475.
17. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. Bioinformatics 20: 2626–2635.
18. Wong SL, Zhang LV, Roth FP (2005) Discovering functional relationships: biochemistry versus genetics. Trends Genet 21: 424–427.
19. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, et al. (2002) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 30: 31–34.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
21. Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the General Repository for Interaction Datasets. Genome Biol 4: R23.
22. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109–126.
23. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273–3297.
24. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257.
25. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, et al. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol Biol Cell 12: 2987–3003.
26. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B. pp 289–300.
27. Storey JD (2002) A direct approach to false discovery rates. J Roy Stat Soc B. pp 479–498.
28. Dwight SS, Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, et al. (2004) Saccharomyces genome database: underlying principles and organisation. Brief Bioinform 5: 9–22.
29. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3: e267.
30. Wade CH, Umbarger MA, McAlear MA (2006) The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes. Yeast 23: 293–306.
31. Rempola B, Karkusiewicz I, Piekarska I, Rytka J (2006) Fcf1p and Fcf2p are novel nucleolar Saccharomyces cerevisiae proteins involved in pre-rRNA processing. Biochem Biophys Res Commun 346: 546–554.
32. Protchenko O, Rodriguez-Suarez R, Androphy R, Bussey H, Philpott CC (2006) A screen for genes of heme uptake identifies the FLC family required for import of FAD into the endoplasmic reticulum. J Biol Chem 281: 21445–21457.
33. Escusa S, Camblong J, Galan JM, Pinson B, Daignan-Fornier B (2006) Proteasome- and SCF-dependent degradation of yeast adenine deaminase upon transition from proliferation to quiescence requires a new F-box protein named Saf1p. Mol Microbiol 60: 1014–1025.
34. Domingos P, Pazzani M (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning 29: 103–130.
35. Rachlin J, Kasif S, Salzberg S (1994) Towards a better understanding of memory-based reasoning systems. International Conference on Machine Learning. pp 242–250.