An Integrated Probabilistic Model for Functional Prediction of Proteins

Minghua Deng, Ting Chen, Fengzhu Sun

Molecular and Computational Biology Program
Department of Biological Sciences
University of Southern California
1042 West 36th Place, Los Angeles, CA 90089-1113, USA
Contact: fsun@hto.usc.edu or tingchen@hto.usc.edu

ABSTRACT

We develop an integrated probabilistic model to combine protein physical interactions, genetic interactions, highly correlated gene expression network, protein complex data, and domain structures of individual proteins to predict protein functions. The model is an extension of our previous model for protein function prediction based on Markovian random field theory. The model is flexible in that other protein pairwise relationship information and features of individual proteins can be easily incorporated. Two features distinguish the integrated approach from other available methods for protein function prediction. One is that the integrated approach uses all available sources of information with different weights for different sources of data. It is a global approach that takes the whole network into consideration. The second feature is that the posterior probability that a protein has the function of interest is assigned. The posterior probability indicates how confident we are about assigning the function to the protein. We apply our integrated approach to predict functions of yeast proteins based upon MIPS protein function classifications and upon the interaction networks based on MIPS physical and genetic interactions, gene expression profiles, Tandem Affinity Purification (TAP) protein complex data, and protein domain information. We study the sensitivity and specificity of the integrated approach using different sources of information by the leave-one-out approach. In contrast to using MIPS physical interactions only, the integrated approach combining all of the information increases the sensitivity from 57% to 87% when the specificity is set at 57%-an increase of 30%. It should also be noted that enlarging the interaction network greatly increases the number of proteins whose functions can be predicted.

RECOMB'03, April 10–13, 2003, Berlin, Germany.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and genetics*

General Terms

Algorithms

Keywords

Function Prediction, Pfam Domain, Protein-Protein Interaction, Markov Random Field, Gibbs Sampler

1. INTRODUCTION

Protein function prediction is one of the most important problems in the post-genome era. The classical method of protein function prediction is to find homologies between a protein and other proteins in protein databases using programs such as FASTA [34] and PSI-BLAST [1] and then to predict functions based on sequence homologies. Another sequence-based approach is called the "Rosetta stone method" where two proteins are inferred to have similar functions if they are together in another genome [27]. By comparing a number of sequenced genomes, the phylogenetic pattern (the presence and absence of the protein in these sequenced genomes) of a protein can be determined. It is believed that genes with similar phylogenetic patterns are likely to share similar functions. Using this idea, the functional links between genes can be predicted [28] based on phylogenetic patterns.

Recent developments of high-throughput bio-techniques have generated a variety of different sources of data that are useful for the study of protein functions. Clustering analysis of gene expression data can be used to predict functions of unknown proteins based on the idea that co-expressed genes are more likely to have similar functions [3, 11, 33]. Methods have also been developed to predict protein functions based on protein physical or genetic interactions using the idea of guilt-by-association: *the neighborhood-count method* [12, 36] and *the chi-square method* [18]. Protein complex data can be used for protein function prediction based on the idea that proteins in the same complex tend to have similar function. For the variety of different interaction data and how they relate to protein functions, see [29].

We developed a Markov random field (MRF) model for protein function prediction using protein-protein interaction data [6]. Two main features distinguish the MRF-based methods from other guilt-by-association methods. One is that the MRF model uses global information on the entire interaction network instead of the local interaction network. The second is that the MRF model gives the *probability* that a protein has a function of interest instead of predicting whether the protein has or does not have the function. This probability indicates how confident we are about assigning the function to the protein. The method was applied to the prediction of protein function based on "cellular role" using protein functions defined in Yeast Proteome Database (YPD) [5]. The results showed that the MRF-based method outperforms the two guilt-by-association methods.

Features of individual proteins have long been used for protein function prediction. A feature here refers to an observation about a protein. It can be the presence or absence of a motif signal, the protein's isoelectric point, its absolute mRNA expression level, or mutant phenotypes from experiments about the sensitivity or resistance of disruption mutants under various growth conditions. Features have been used for protein function prediction [15, 16, 22, 39], and for protein function prediction as pattern recognition problems [4, 23, 24]. Drawid and Gerstein [9] developed a general Bayesian approach to predicting protein localization based on a large number of features of individual proteins.

However, no methods are available for predicting protein functions combining all of the different sources of information. In this paper, we extend the MRF-based method to create an integrated approach that includes other protein pairwise relationships such as correlations of gene expression patterns, genetic interactions, and features of individual proteins such as domain information. The model is flexible in that other protein pairwise relationship information such as pairwise protein sequence similarities and features of individual proteins can be easily incorporated. We apply our integrated approach to predict functions of yeast proteins based upon MIPS protein functions and the interaction networks based upon MIPS physical and genetic interactions, gene expression profiles, Tandem Affinity Purification (TAP) protein complex data, and protein domain information. We study the sensitivity and specificity of the integrated approach using different sources of information by the leave-one-out approach. In contrast to using MIPS physical interactions only, the integrated approach combining all of the information increases the sensitivity from 57% to 87% when the specificity is set at 57%-an increase of 30%. It should also be noted that enlarging the interaction network greatly increases the number of proteins whose functions can be predicted.

The paper is organized as follows. In the *Method* section, we first briefly describe the MRF model developed in [6], then the integrated MRF model, and finally the computational methods. In the *Results* section, we apply the integrated approach to predict protein functions using a variety of different information. Finally, we discuss the implications and limitations of our integrated approach.

2. METHOD

To make this paper self-contained, we will briefly describe the MRF model based on protein physical interactions [6]. In this model, given a function of interest, the objective is to predict the probability that an unknown protein has the function, using the protein physical interaction network and the functions of the known proteins.

Suppose a proteome has N proteins $P_1, \cdots, P_N$. Some proteins have already been studied and have known functions, while others have unknown functions. Following the tradition, we will refer to proteins having known functions as *known proteins* and to proteins with unknown functions as *unknown proteins* throughout the paper. Let $P_1, \cdots, P_n$ be the unknown proteins, $P_{n+1}, \cdots, P_{n+m}$ be the known proteins, and $N = n + m$.

Throughout this paper, we will fix a function of interest. Let $X_i = 1$ if the i-th protein has the function and $X_i = 0$ otherwise. Let $X = (X_1, \cdots, X_{n+m})$ be the configuration of the functional labelling, where $X_1 = \lambda_1, \cdots, X_n = \lambda_n$ are unknown and $X_{n+1} = \mu_1, \cdots, X_{n+m} = \mu_m$ are known. We infer the function of the unknown proteins using the protein interaction network obtained from biological experiments.

We first give the prior probability distribution of X based on the interaction network-the *Gibbs distribution* [25]. In the following, X_i will be the random variable and x_i will be its observed value. Without considering the interaction network, the probability of a configuration of X is proportional to

$$\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i} = \left(\frac{\pi}{1-\pi}\right)^{N_1}(1-\pi)^N, \qquad (1)$$

where $N_1 = \sum_{i=1}^{N} x_i$ and π is the probability of a protein having the function of interest.

Next, we consider the protein physical interaction network. The probability of the interactions in the network conditional on the functional labelling is proportional to

$$\exp(\beta_1 N_{10} + \gamma_1 N_{11} + \kappa_1 N_{00}), \qquad (2)$$

where $N_{ll'}$ is the number of (l, l')-interacting pairs in S and

$$\begin{aligned} N_{11} &= \sum_{(i,j)\in S} x_i x_j \\ &= \#\{(1 \leftrightarrow 1) \text{ pairs in S}\}, \\ N_{10} &= \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\ &= \#\{(1 \leftrightarrow 0) \text{ pairs in S}\}, \\ N_{00} &= \sum_{(i,j)\in S} (1-x_i)(1-x_j) \\ &= \#\{(0 \leftrightarrow 0) \text{ pairs in S}\}, \end{aligned} \qquad (3)$$

where S is the set of all the physical interaction pairs under consideration.

Therefore, the total probability of the network based on the functional labelling of the proteins and the interactions is proportional to $\exp(-U(x))$, where

$$\begin{aligned} U(x) &= -\alpha N_1 - \beta_1 N_{10} - \gamma_1 N_{11} - \kappa_1 N_{00} \\ &= -\alpha \sum_{i=1}^{N} x_i - \beta_1 \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\ &\quad - \gamma_1 \sum_{(i,j)\in S} x_i x_j - \kappa_1 \sum_{(i,j)\in S} (1-x_i)(1-x_j), \end{aligned} \qquad (4)$$

where $\alpha = \log(\frac{\pi}{1-\pi})$. Under the above model, one parameter is redundant, and we can set $\kappa_1 = 1$.

Using the above MRF model, we developed a Gibbs sampling scheme to estimate the posterior distribution of (X_1, $\cdots$, X_n) conditional on $X_{n+1} = \mu_1, \cdots, X_{n+m} = \mu_m$. The posterior probability distribution of X_i can be obtained by summing over all of the possible configurations of X_j, $j \neq i$, $1 \leq j \leq n$.

2.1 The general MRF model

The MRF model and the Bayesian approach described above can be extended to include all protein pairwise relationships and features of individual proteins for protein function prediction.

2.1.1 Using protein complex data to assign prior probabilities

Mass spectrometry has been used to identify protein complexes [13, 19]. Researchers have used a set of proteins as baits to prey other proteins in the same complexes, followed by tandem mass spectrometry experiments to identify each protein in the complexes. Proteins in a complex do not necessarily physically interact with each other, although they are more likely to physically interact than random protein pairs. A direct physical interaction map cannot be established through the protein complex data. It is generally believed that proteins within a complex are more likely to have the same function. For a given function of interest and an unknown protein P_i in a protein complex, we give a prior probability that the protein has the function by

$$\begin{aligned}&\Pr(X_i = 1 \,|\, \text{Complex})\\ =&\frac{\#\{\text{Proteins having the function within the complex}\}}{\#\{\text{Known proteins within the complex}\}}.\end{aligned} \tag{5}$$

A protein may belong to different protein complexes. For example, in TAP protein complex data [13], protein "RPN10" was observed to appear in six protein complexes. For each protein complex, we compute the prior probability that the unknown protein has the function of interest and use the maximum of these prior probabilities as the true prior probability that the protein has the function. The basic idea behind this choice is that proteins in a large complex are more likely to have different functions than proteins in a small complex.

For those proteins that belong to at least one of the identified protein complexes, we can use the above approach to give a prior probability that the protein has the function of interest. For other proteins, we use the fraction of proteins in the entire proteome having the function as the prior. Then, without any information on protein pairwise relationship, the probability of a configuration of X is proportional to

$$P\{\text{labelling}\} \propto \prod_{i=1}^{N} (\pi_i)^{x_i} (1-\pi_i)^{1-x_i}, \tag{6}$$

where π_i is the prior probability that the i-th protein has the function of interest. The main difference between this equation and equation 1 is that π_i can be different for different proteins.

2.1.2 The MRF model including multiple sources of pairwise relationship

It is generally believed that co-expressed genes generally have similar functions. We built a co-expressed network by connecting protein pairs if the correlation coefficient of the expression profiles of the proteins was greater than a certain threshold, say 0.8. Another data source is genetic interactions obtained through mutation analysis. Based on genetic interaction data, we can build a genetic interaction network by connecting proteins if they genetically interact with one another.

Generally speaking, suppose that we have L sources of protein pairwise relationship that may be useful for protein function prediction, and a network that can be built based on each source of data, denoted as $\text{Net}_1, \text{Net}_2, \cdots, \text{Net}_L$, respectively. The entire network we consider is the union of all the networks denoted as S.

Based on the l-th network, similar to equation 2, our belief for the functional labelling of all the proteins is proportional to

$$P\{\text{Net}_l \mid \text{labelling}\} \propto \exp(\beta_l N_{10}^{(l)} + \gamma_1 N_{11}^{(l)} + \kappa_1 N_{00}^{(l)}),$$

where $(N_{10}^{(l)}, N_{11}^{(l)}, N_{00}^{(l)})$ are defined in a manner similar to equation 3, with S replaced by the l-th network.

Multiplying over all the networks, our belief for the functional labelling of all the proteins is proportional to

$$\begin{aligned}P\{\text{networks} \mid \text{labelling}\} &\propto \prod_{l=1}^{L} \exp(\beta_l N_{10}^{(l)} + \gamma_1 N_{11}^{(l)} + \kappa_1 N_{00}^{(l)})\\ &= \exp \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right).\end{aligned} \tag{7}$$

Our total belief for the functional labelling of all the proteins is proportional to the multiplication of equations 6 and 7.

Then an MRF over all the functional labelling is defined by

$$P\{\text{labelling, networks}\} = \exp(-U(x))/Z(\theta), \tag{8}$$

where

$$U(x) = -\sum_{i=1}^{n+m} x_i \alpha_i - \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right), \tag{9}$$

where $\alpha_i = \log(\frac{\pi_i}{1-\pi_i})$ and is given based on protein complex data, θ indicates the vector of parameters, and $Z(\theta)$ is the summation of $\exp(-U(x))$ over all the functional labelling. Under the above model, all the parameters $(\kappa_1, \kappa_2, \cdots, \kappa_L)$ are redundant and are set to 1 in the rest of the paper. In the terminology of MRF, $U(x)$ is called the potential function.

The MRF model defined in equation 8 gives the prior distribution for the functional labelling using information from protein complexes as well as from different sources of pairwise relationship.

2.1.3 Using domain information

The function of a protein is determined by its structure. Therefore, the structure or the amino acid sequence of a protein can be very useful for predicting protein functions. However, it is impossible to directly use the amino acid sequence data for protein function prediction because very large number of parameters are needed. Instead, protein features extracted from sequence data should be used. As a first step toward showing the proof of principles, we simply use the domain information of the proteins: the presence or the absence of a set of domains.

Several investigators have shown that protein domains are an informative feature for protein function prediction [17, 37]. For a given domain set $D_1, D_2, \cdots, D_M$, the domain structure of each protein P_i, $d_i = (d_{i1}, d_{i2}, \cdots, d_{iM})$ can be defined, where $d_{im} = 1$ if the i-th protein P_i contains domain D_m and $d_{im} = 0$ otherwise. Let p_{1m} (p_{0m}) be the conditional probability of $d_m = 1$ given that a protein has (does not have) the function of interest. For simplicity, we assume that all the domains independently contribute to the functions of proteins.

For a given domain structure $d = (d_1, d_2, \cdots, d_M)$, we let

$$P_1(d) = \prod_{m=1}^{M} p_{1m}^{d_m}(1-p_{1m})^{1-d_m},$$
$$P_0(d) = \prod_{m=1}^{M} p_{0m}^{d_m}(1-p_{0m})^{1-d_m}.$$

Then we are able to calculate the probability of the domain features of all the proteins given the functional labelling.

$$P\{\text{domain features} \mid \text{labelling}\} = \prod_{i:X_i=1} P_1(d_i) \times \prod_{i:X_i=0} P_0(d_i). \tag{10}$$

Multiplying equations (8, 10), we have the following probability model

$$P\{\text{labelling, networks, domain features}\} = P\{\text{labelling, networks}\} \times P\{\text{domain features} \mid \text{labelling}\}.$$

The problem is how to estimate the posterior distribution of the functions of the unknown proteins given the features of all the proteins, the different sources of protein pairwise relationship, and the annotations of the known proteins.

2.2 Computational Issues

To implement the above procedure for protein function prediction, we need to estimate the parameters involved in the model. A maximum likelihood estimation procedure for estimating the parameters is impractical due to the high dependency among the functions of the proteins introduced by the interaction networks. In this paper, we consider the following estimation procedures. We estimate π_i in equation 5 by the functions of the known proteins. Similarly, we estimate p_{1m} and p_{0m} using the domain features of known proteins as follows.

$$p_{1m} = \frac{\#\text{proteins having the function and containing domain } D_m}{\#\text{proteins having the function}},$$
$$p_{0m} = \frac{\#\text{ proteins not having the function and containing domain } D_m}{\#\text{ proteins not having the function}},$$

We use a pseudo-likelihood approach to estimate β_l, γ_l, $1 \le l \le L$ [25]. Based on the above general model, we have

$$\Pr(X_i = 1 \mid D, X_{[-i]}, \theta) = \frac{e^{\alpha_i + \sum_{l=1}^{L}(\beta_l - 1)M_0^{(i)}(l) + (\gamma_l - \beta_l)M_1^{(i)}(l)}}{1 + e^{\alpha_i + \sum_{l=1}^{L}(\beta_l - 1)M_0^{(i)}(l) + (\gamma_l - \beta_l)M_1^{(i)}(l)}}, \tag{11}$$

or, equivalently,

$$\log \frac{Pr(X_i = 1 \mid D, X_{[-i]}, \theta)}{1 - Pr(X_i = 1 \mid D, X_{[-i]}, \theta)} = \alpha_i + \sum_{l=1}^{L}(\beta_l - 1)M_0^{(i)}(l) + (\gamma_l - \beta_l)M_1^{(i)}(l), \tag{12}$$

where D is the domain information for all the proteins, $X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{n+m})$, $\alpha_i = \log \frac{\pi_i P_1(d_i)}{(1-\pi_i)P_0(d_i)}$, $M_0^{(i)}(l)$ and $M_1^{(i)}(l)$ are the numbers of neighbors of protein P_i labelled with 0 and 1 according to the l-th network, respectively.

We used the network consisting of the known proteins to estimate those parameters by an S-plus routine [41] using equation 12.

Once all the parameters have been defined, we use a Gibbs sampler to estimate the posterior probability distribution of $(X_1, \cdots, \cdots, X_n)$. The algorithm can be described as follows:

1. Randomly set the value of missing data $X_i = \lambda_i, i = 1, \cdots, n$ with probability π_i.
2. For each protein P_i, update the value of X_i using Equation 11.
3. Repeat step 2 T times until all the posterior probabilities $\Pr(X_i \mid D,\ X_{[-i]}, \theta)$ are stabilized.

In Gibbs sampling, the "burn-in period" and the "lag period" need to be specified [26]. The burn-in period is the time it takes the Markovian process to become stabilized, and the simulation results in the burn-in period are discarded to reduce or eliminate the effect of initial values. After the burn-in period, the probability that an unknown protein has a particular function is estimated by averaging the simulation results in steps of the lag-period to reduce or eliminate the dependence of the Markovian process. In this study, the burn-in period and the lag period are 100 and 10, respectively. The total number of simulations is 2000.

3. RESULTS

3.1 Sources of data

We applied the above integrated approach to predict functions of unknown proteins in Yeast. First, we used 6278 genes from the SGD database [10]. Second, we used the Protein Families Database of Alignments and HMMs (Pfam domain) to define the domain structures of all the proteins. The SwissPfam (Ver7.5) defines the mapping between proteins' SWISS-PROT/TrEMBL accession numbers and Pfam domains. The final mapping between SGD proteins and their Pfam domains was built by their SWISS-PROT/TrEMBL accession numbers. Third, we used the functional classification catalogue based on the Munich Information Center for Protein Sequences (MIPS) to define functions [30]. The MIPS functional classification catalogue is hierarchical, and, for simplicity, we used only the level-one classification. There are 18 level-one functional classes, including "classification not yet clear-cut" and "unclassified proteins" which were merged as a single class "unknown". The following functional classes contained small numbers of proteins

and were thus merged into one class: "cellular communication/signal transduction mechanism" (59), "protein activity regulation" (13), "protein with binding function or cofactor requirement (structural or catalytic)" (4), and "transposable elements, viral and plasmid proteins" (116). We thus had 13 known functional classes and one "unknown" in our analysis. Fourth, we used three sources of protein pairwise relationship, including MIPS physical interactions, TAP protein complexes, and the cell cycle gene expression data of [38]. The MIPS physical interaction data contain 2,448 interaction pairs (excluding 120 pairs of self-interactions) involving 1,877 proteins extracted from the literatures It is generally believed that this data set is more reliable than other pairwise protein-protein interaction data [7, 29, 31]. The TAP protein complex data contain 232 complexes involving 1,088 known and 237 unknown proteins with respect to MIPS function classification [13]. The cell cycle gene expression data [38] contains expressions of 6,086 genes with 77 data points (2 cln3, 2 clb, 18 alpha, 24 cdc15, 17 cdc28 and 14 elut).

We studied the sensitivity and the specificity of the integrated approach using different combinations of protein pairwise relationship and domain information of individual proteins using the leave-one-out approach. The sensitivity (or specificity) is defined as the fraction of overlaps between the known functions and predicted functions over all of the known (or predicted) functions.

3.2 Combining different sources of pairwise relationship

We used the MIPS physical interaction data as the basis for comparison because it contains the largest number of interactions. We then added the MIPS genetic interaction data and the network defined by highly co-expressed protein pairs (correlation coefficient ≥ 0.8) to the MIPS physical interaction data. Figure 1 shows the relationship between sensitivity and specificity of the integrated approach using (1) physical and genetic interactions, (2) physical interactions and gene expressions, and (3) physical interactions, genetic interactions, and gene expressions. In contrast to using physical interactions only, the genetic interactions can substantially increase the performance of the method (Figure 1a). It should be noted, however, that although the prediction method based on MIPS genetic interaction data seems to outperform that using the combined physical and genetic interactions in some cases, the number of proteins that can be predicted based on genetic interactions alone is much smaller than that of the combined physical and genetic data (See Table 1). Figure 1b shows that adding gene expression data to the physical interaction data does not significantly increase the performance of the method based only on physical interactions. Figure 1c shows that the performance of the integrated approach that combines the three sources of pairwise relationship is similar to that combining MIPS physical and genetic interactions. The above observations also hold when we use different thresholds for the correlation coefficients of gene expression profiles for defining the network (data not shown).

3.3 Combining physical interactions with protein complexes and domain information

We added the protein complex data and the domain information onto the MIPS physical interaction data in the integrated approach. Figure 2 shows the relationship between the sensitivity and specificity of the integrated approach using (1) the MIPS physical interactions and the TAP protein complexes, (2) the MIPS physical interactions and the domain information, and (3) the MIPS physical interactions, the TAP protein complexes, and the domain information. There were 1,008 known and 237 unknown proteins in the TAP complex data, and the prior probabilities for those proteins could be estimated as described in the method section. For other proteins, we used the fraction of proteins having the function among all of the proteins as the prior. For the prediction by domain only, we simply computed the posterior probability $\Pr(X = 1 \mid D)$ without other information as predictions. Figure 2 shows that both the protein complex data and the domain information can significantly improve the performance of the methods.

3.4 Combing all the information for protein function prediction

Finally, we combined all of the information, including the MIPS physical interactions, the MIPS genetic interactions, the gene expressions, the TAP protein complex data, and the Pfam domain information. Figure 3 shows that for a given specificity the sensitivity of the integrated approach increases rapidly as more information is added. For example, when only the MIPS physical interaction data were used, the sensitivity and specificity were roughly the same when the specificity was set at 57%. At this specificity, the sensitivity of the integrated approach incorporating all of the information reached 87%. When all the information was used, the sensitivity and specificity were roughly the same at 76%. Figure 3b shows that for a given specificity the sensitivity increases as the number of interaction partners increases.

4. DISCUSSION

We developed an integrated probabilistic model to combine the protein physical interactions, the genetic interactions, the highly correlated gene expression network, the protein complex data, and the domain structures of individual proteins to predict protein functions. We estimated the posterior probability that the protein has the function of interest given all of the available information. The posterior probability indicates how confident we are about assigning the function to the protein. The distinction of the Bayesian approach we develop here is that it is a global approach taking into consideration all of the interaction network and the functions of known proteins.

We applied our integrated approach to predict functions of yeast proteins based upon MIPS protein function classifications and upon the interaction networks based on MIPS physical and genetic interactions, the gene expression profiles, the TAP protein complex data, and the protein domain information from the Pfam database. We studied the sensitivity and specificity of the integrated approach using different sources of information by the leave-one-out approach. In contrast to using MIPS physical interactions only, the integrated approach combining all of the information increases the sensitivity and specificity significantly, and at the same time, it uses a much larger interaction network, greatly increasing the number of proteins whose functions can be predicted. It should be noted that the probability model is flexible enough to be able to incorporate other information.

There are several limitations to our approach. Both the

interaction network and the functional annotations of the proteins are incomplete. The actual number of interacting protein pairs might be much higher than what have been obtained in MIPS.

Our method treats each function independently and separately. Generally, that fact that a protein has one function does not prevent it from having other functions. Therefore, our model determines each function for each protein without a bias. However, there are correlations between functions. The fact that a protein has a function A may increase the chance of it having function B because functions A and B are highly correlated. Incorporating these information into a generalized model remains a challenging task. Our model assumes that known proteins have complete functional annotations, and it predicts functions for unknown proteins using this information. In reality, we know that these known proteins may have other functions that have not been determined. As biologists continue to experimentally determine the functions of proteins, the functional classifications of proteins will be more and more complete.

Electronic-Database Information

URLs for data presented herein are as follows:

SGD database, http://genome-www.stanford.edu/Saccharomyces/
SwissPfam, ftp://ftp.genetics.wustl.edu/pub/pfam/
MIPS, http://mips.gsf.de/

5. REFERENCES

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **25:** 3389 − 3402. 1997.

[2] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller. L., Eddy, S.R., Griffiths-Jones S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. The Pfam Protein Families Database. *Nucleic Acids Research* **30:** 276 − 280. 2002.

[3] Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. Jr., and Haussler, D. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines *Proc. Natl. Acad. Sci. USA* **97:** 262 − 267. 2000.

[4] Clare, A. and King, R.D. Machine Learning of Functional Class from Phenotype Data. *Bioinformatics* **18:** 160 − 166. 2002.

[5] Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., Lengieza, C., Lew-Smith, J.E., Tillberg, M., and Garrels, J.I. YPD^{TM}, PombePD^{TM}, and WormPD^{TM}: Model Organism Volumes of the BioKnowledge Library, an Integrated Resource for Protein Information. *Nucleic Acids Res.* **29:** 75 − 79. 2001.

[6] Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. Prediction of Protein Function Using Protein-protein Interaction Data. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002):* 197 − 206. 2002.

[7] Deng, M., Chen, T., and Sun, F. Assessment of the Reliability of Protein-protein Interactions and Protein Function Prediction. To appear in *Pacific Symposium of Biocomputing (PSB2003)* . 2002.

[8] Devos,D., and Valencia, A. Pratical Limits of Function Prediction. *Proteins: Structure, Function, and Genetics* **41:** 98 − 107. 2000.

[9] Drawid, A. and Gerstein, M. A Bayesian System Integrating Expression Data with Squence Patterns for Localizing Proteins: Comprehensive Application to the Yeast Genome. *J. Mol. Bio.* **301:** 1059 − 1075. 2000.

[10] Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., and Sherlock, G. et al. Saccharomyces Genome Database (SGD) Provides Secondary Gene Annotation Using the Gene Ontology (GO). *Nucleic Acids Res.* **30:** 69 − 72. 2002.

[11] Eisen, M.B., Spellman, P.T., Brown, P.O., and Bostein D. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Natl. Acad. Sci. USA* **95:** 14863 − 14868. 1998.

[12] Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.W., and Hani J. Integrative Analysis of Protein Ineraction Data. In *Proc. of the Eighth Int. Conf. on Intelligent System for Molecular Biology (ISMB2000)*: 152 − 161. 2000.

[13] Gavin, A., Böche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., and Cruciat, C. et al. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature* **415:** 141 − 147. 2002.

[14] Greenbaum, D., Luscombe, N.M.,Jansen, R., Qian, J., and Gerstein, M. Interrelating Different Types of Genomic Data, from Proteome to Secretome: Coming in on Function. *Genome Research* **11:** 1463 − 1468. 2001.

[15] Gupta, R. and Brunak, S. Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function. *Pacific Symposium of Biocomputing (PSB2002):* 310 − 322. 2002.

[16] Hegyi,H. and Gerstein, M. (1999). The Relationship Between Protein Structure and Function: a Comprehensive Survey with Application to Yeast Genome. *J. Mol. Bio.* **288:** 147 − 164. 1999.

[17] Hegyi,H. and Gerstein, M. Annotation Transfer for Genomics: Measuring Fuctinal Divergence in Multi-domain Proteins. *Genome Research* **11:** 1632 − 1640. 2001.

[18] Hishigaki H., Nakai K., Ono T., Tanigami A., and Takagi T. Assessment of Prediction Accuracy of Protein Function from Protein-protein Interaction Data. *Yeast* **18:** 523 − 531. 2001.

[19] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., and Boutilier, K. et al. Systematic Identification of Protein Complexes in *Saccharomyces Cerevisiae* by Mass Spectrometry. *Nature* **415:** 180 − 183.2002.

[20] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto,K., Kuhara, S., Sakaki, Y. Toward a Protein-protein Interaction Map of the

Budding Yeast: a Comprehensive System to Examine Two-hybrid Interactions in All Possible Combinations Between the Yeast Proteins. *Proc. Natl. Acad. Sci. USA* **97:** 1143 − 1147. 2000.

[21] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. A Comprehensive Two Hybrid Analysis to Explore the Yeast Protein Interactome. *Proc. Natl. Acad. Sci. USA* **98:** 4569 − 4574. 2001.

[22] Jensen, L.J., Gupta, R., Blom,N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H.H, Rapacki,K., and Workman, C. et. al. Prediction of Human Protein Function from Post-translational Modifications and Localization Features. *J. Mol. Bio.* **319:** 1257 − 1265. 2002.

[23] Kell, D.B. and King, R.D. On the optimization of Classes for the Assignment of Unidentified Reading Frames in Functional Genomics Programmes: the Need for Machine Learning. *Trends Biotechnol.* **18:** 93 − 98. 2000.

[24] King, R.D. , Karwath, A., Clare, A., and Dehaspe, L. The Utility of Different Representations of Protein Sequence for Predicting Functional Class. *Bioinformatics* **17:** 445 − 454. 2001.

[25] Li, S.Z. (1995). Markov Random Field Modeling in Computer Vision. Springer-Verlag: Tokyo.

[26] Liu, J.S. (2001). Monte Carlo Strategies in Scientific Computing. Springer-Verlag: New York.

[27] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice,D.W., Yeates, T.O., and Eisenberg, D. Detecting Protein Function and Protein-protein Interactions from Genome Sequences. *Science* **285:** 751 − 753. 1999.

[28] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999). A Combined Algorithm for Genome-wide Prediction of Protein Function. *Nature* **402:** 83 − 86. 1999.

[29] C.V. Mering, R. Krause, M. Snel, S.G. Oliver, S. Fields, and P. Bork. Comparative Assessment of Large Scale Data Sets of Protein-protein Interactions. *Nature* **417:** 399 − 403. 2002.

[30] H.W. Mewes, D. Frishman, U. Guldener,G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a Database for Genomes and Protein Sequences. *Nucleic Acids Research* **30:** 31 − 34. 2002.

[31] R. Mrowka, A. Patzak, and H. Herzel. (2001) Is There a Bias in Proteome Research? *Genome Research* **11:** 1971 − 1973. 2001.

[32] Oliver, S. Guilt-by-association Goes Global. *Nature* **403:** 601 − 603. 2000.

[33] Pavlidis, P. and Weston,J. Gene Functional Classification from Heterogeneous Data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001):* 249 − 255. 2001.

[34] Pearson,W.R. and Lipman,D.J. Improved Tools for Biological Sequence Comparison. *Proc. Natl Acad. Sci. USA* **85:** 2444 − 2448. 1988.

[35] Pellegrini, M., Marcotte, E.M., Thompson,M.J., Eisenberg, D., and Yeates, T.O. Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles. *Proc. Natl. Acad. Sci. USA* **96:** 4285 − 4288. 1999.

[36] Schwikowski, B., Uetz, P., and Fields, S. A Network of Protein-protein Interactions in Yeast. *Nature Biotechnology* **18:** 1257 − 1261. 2000.

[37] Schug,J., Diskin, S., Mazzarelli, J., Brunk, B.P. ,and Stoeckert, C.J,Jr. Prediction Gene Ontology Functions from Prodom and CDD Protein Domains. *Genome Research* **12:** 648 − 655. 2002.

[38] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9:** 3273 − 3297.1998.

[39] Stawiki, E.W., Mandel-Gutfreund, Y., Lowenthal, A.C. ,and Gregoret, L.M. Progress in Predicting Protein Function from Structure: Unique features of O-Glycosidases. *Pacific Symposium of Biocomputing (PSB2002):* 637 − 648. 2002.

[40] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., and Pochart, et al. A Comprehensive Analysis of Protein-protein Interactions in *Saccharomyces Cerevisiae*. *Nature* **403**: 623 − 627. 2000.

[41] Venables, W.N. and Ripley, B.D. Modern Applied Statistics with S−Plus. Springer-Verlag; New York. 1996.

[42] Wu, L., Hughes, T.R., Davierwala A.P., Robinson, M.D., Stoughton,R., and Altschuler S.J. Large-scale Prediction of *Saccharomyces Cerevisiae* Gene Function using Overlapping Transcriptional Clusters. *Nature Genetics* **31:** 255 − 265. 2002.

[43] Zhou, X., Kao, M., and Wong, W. Transitive Functional Annotation by Shortest-path Analysis of Gene Expression Data. *Proc. Natl. Acad. Sci. USA early edition.* 2002.

Data	Phy	Gen	Exp	Phy+Gen	Phy+Exp	Phy+Exp+Gen
Known	1429	823	373	1736	1655	1931
Unknown	455	17	210	463	622	630
Total	1884	840	583	2199	2277	2561

Table 1: The numbers of proteins having at least one partners in the different networks. Phy: MIPS physical interaction, Gen: MIPS genetic interaction, Exp: highly co-expressed protein pairs."

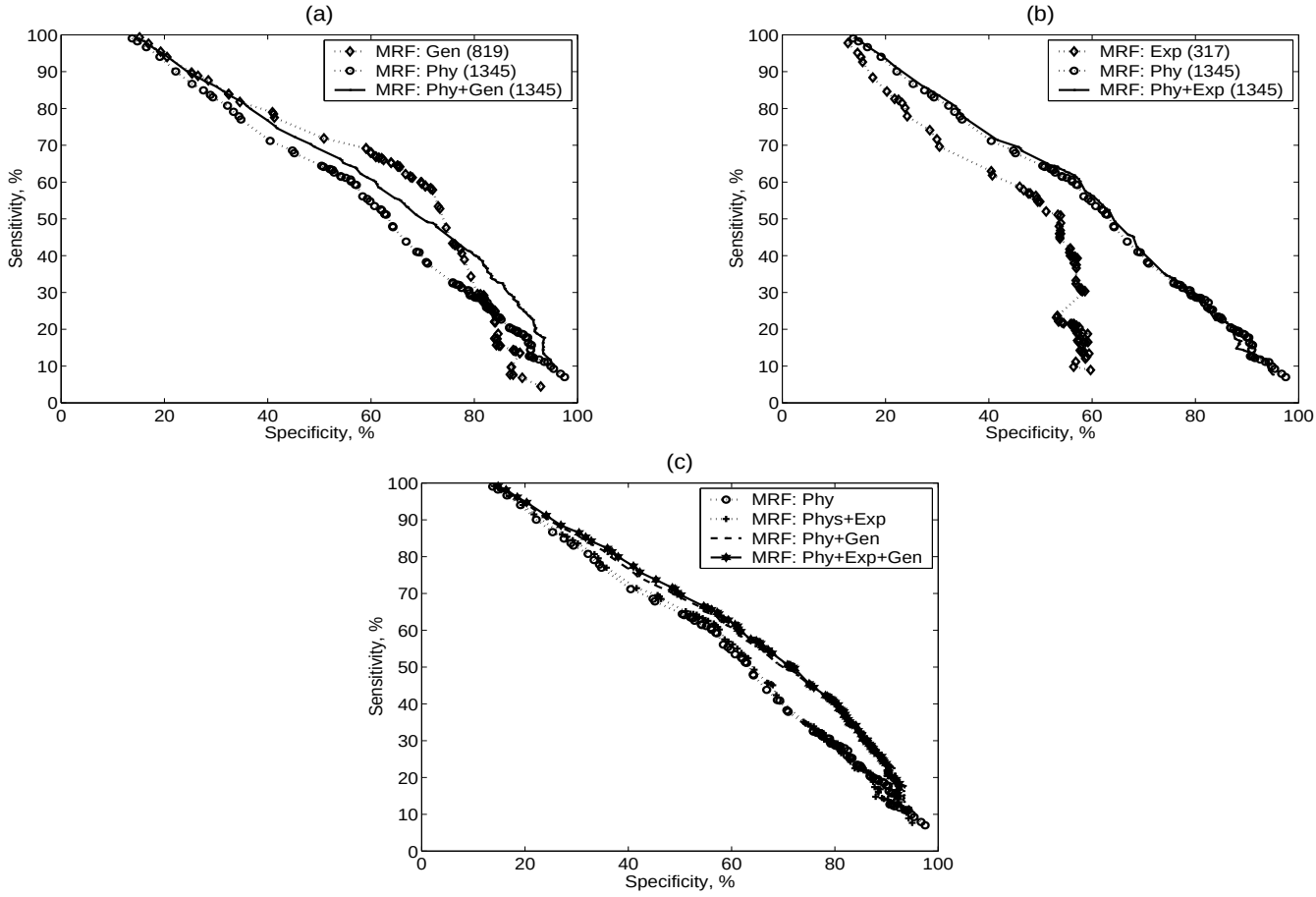

Figure 1: The relationship between sensitivity and specificity for the integrated approach by combining a) MIPS physical and genetic interactions, b) MIPS physical interactions and gene expressions, and c) MIPS physical interactions, genetic interactions and gene expressions. In a) and b), the numbers in the bracket are the numbers of proteins used in calculating the sensitivity and specificity.

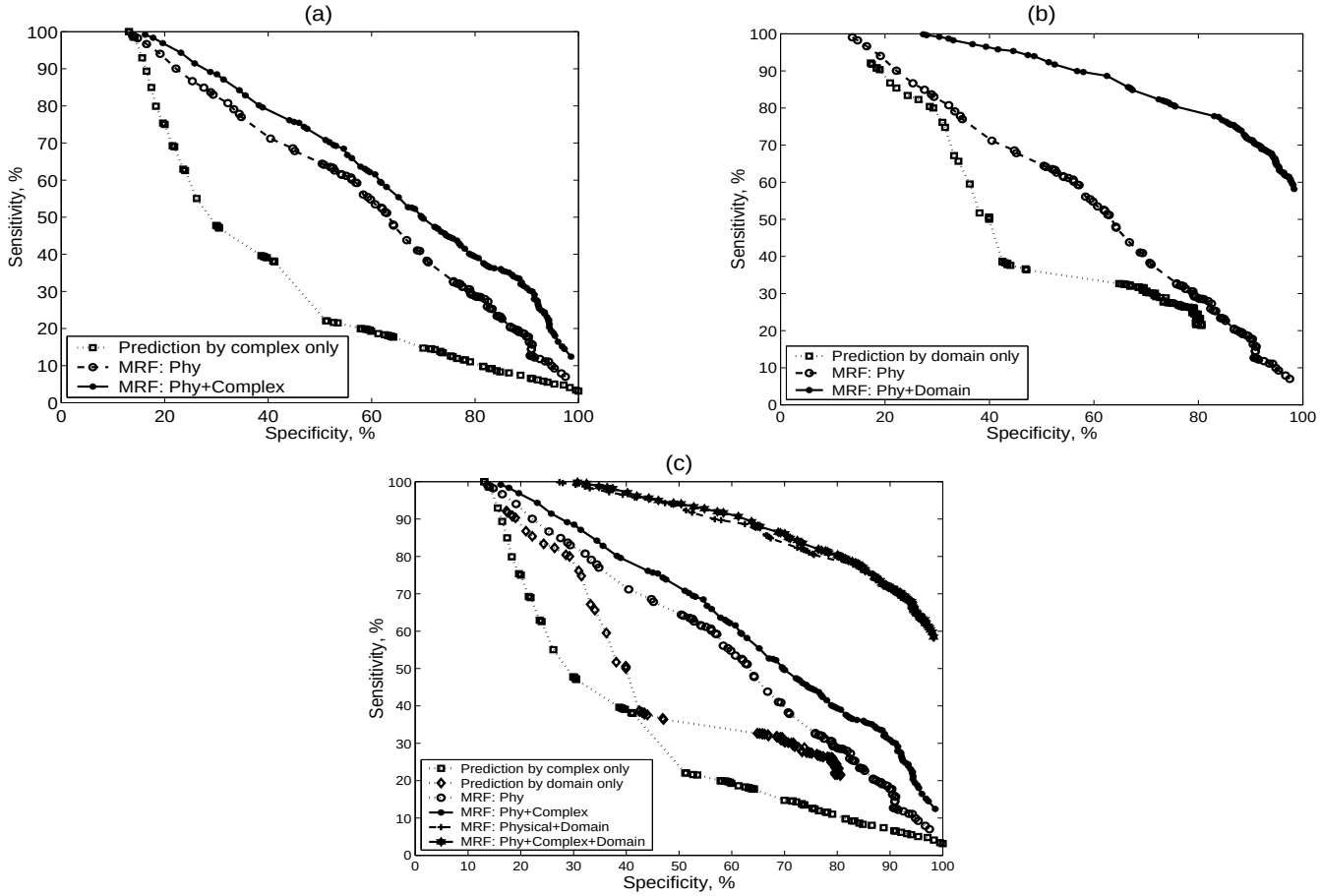

Figure 2: The relationship between sensitivity and specificity of the integrated approach by combining a) MIPS physical interactions and TAP protein complexes, b) MIPS physical interactions and domains, and c) MIPS physical interactions, TAP protein complexes and domains. In a) and b), the numbers in the bracket are the numbers of proteins used in calculating the sensitivity and specificity.

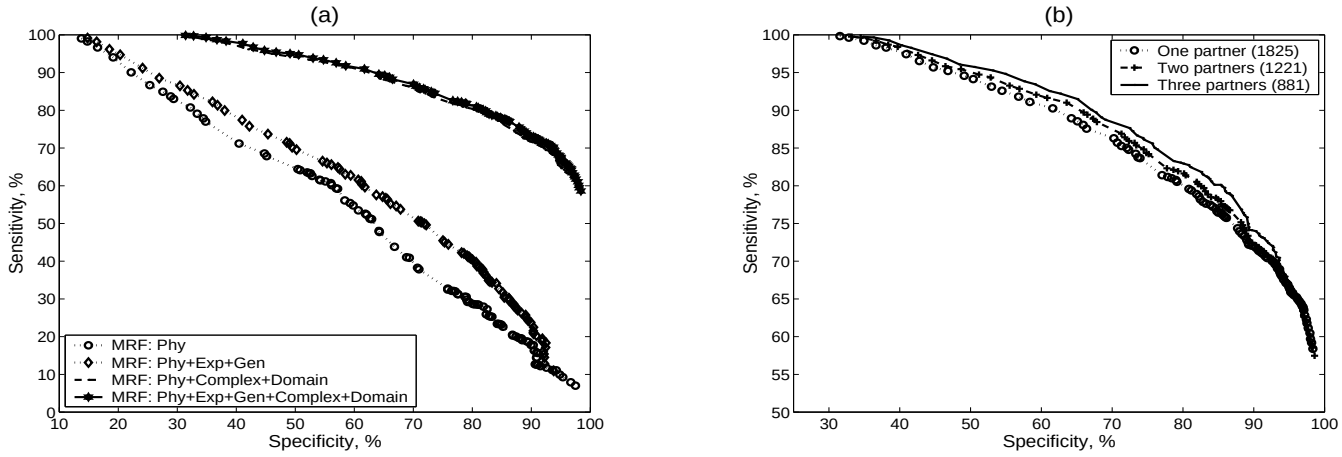

Figure 3: Prediction of the integrated approach by combining MIPS physical interactions, TAP protein complexes, Pfam domain, MIPS genetic Interaction and highly co-expressed gene pairs. a) The relationship between sensitivity and specificity. b) The specificity and sensitivity of those proteins with different interaction partners, the numbers in the bracket are the numbers of proteins used in calculating the sensitivity and specificity.