Prediction of protein function using protein-protein interaction data*

Minghua Deng, Kui Zhang, Shipra Mehta, 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
fsun@hto.usc.edu or tingchen@hto.usc.edu

Abstract

Assigning functions to novel proteins is one of the most important problems in the post-genomic era. Several approaches have been applied to this problem, including analyzing gene expression patterns, phylogenetic profiles, protein fusions and protein-protein interactions. We develop a novel approach that applies the theory of Markov random fields to infer a protein's functions using protein-protein interaction data and the functional annotations of its interaction protein partners. For each function of interest and a protein, we predict the probability that the protein has that function using Bayesian approaches. Unlike in other available approaches for protein annotation where a protein has or does not have a function of interest, we give a probability for having the function. This probability indicates how confident we are about the prediction. We apply our method to predict cellular functions (43 categories including a category "others") for yeast proteins defined in the Yeast Proteome Database(YPD), using the protein-protein interaction data from the Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de). We show that our approach outperforms other available methods for function prediction based on protein interaction data.

1. Introduction

With the completion of genome sequencing of several model organisms, the functional annotation of the proteins is of most importance. Up to February 15, 2002, the Yeast Protein Database(YPD) [5] lists 6281 proteins with 3854 being annotated, assigned to some cellular roles, and 2427 being unannotated. A challenging task that lies ahead is to find the functional roles of these unannotated proteins. Several research groups have developed methods for functional annotation. The classical way is to find homologies between a protein and other proteins in protein databases using programs such as FASTA [24] and PSI-BLAST [1], and then predict functions based on sequence homologies. Another sequence-based approach is called the "Rosetta stone method" where two proteins are inferred to interact if they are together in another genome [20]. 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's believed that genes with similar functions are likely to share similar phylogenetic patterns. Using this idea, the functional links between genes can be predicted [21] based on phylogenetic patterns.

The development of high-throughput bio-techniques and their applications in many areas of biology generated a large amount of data that are useful for the study of protein functions. Several attempts have been made to predict protein functions using such data as gene expressions, mutant phenotype, and protein-protein interactions. Clustering analysis of gene expression data can be used to predict functions of unannotated proteins based on the idea that genes with similar functions are likely to be co-expressed [3, 7, 23]. Moreover, functional predictions have been modeled as pattern recognition problems based on sequence homologies and structural information [16, 17] as well as phenotype data [4].

Proteins play an important role in many biological functions within a cell and many cellular processes and biochemical events are ultimately achieved by a group of proteins interacting with one another. Proteins collaborate or interact with one another for a common purpose, and thus it is possible to deduce functions of a protein through the functions of its interaction partners. It should be noted that the interaction partners for a protein may belong to different functional categories. It is this complex network of within-function and cross-function interactions that makes the problem of functional assignments a difficult task. Methods based on χ^2-statistics [12] and on frequencies of interaction partners having certain functions of interest [8, 27] have been applied to assign functions to unannotated proteins. However, these methods lack a systematic mathemat-

*Published in IEEE Computer Society Bioinformatics Conference 2002

ical model. In this paper, we propose a mathematical model for protein-protein interactions, and use Bayesian analysis to assign functions to proteins.

We define a Gibbs distribution for the protein-protein interaction network. With this Gibbs distribution, we develop a Gibbs sampler to estimate the posterior probabilities that an unannotated protein has certain functions of interest.

2. Method

We first describe the basic ideas of our approach. The protein-protein interaction network describes a neighborhood structure among the proteins. If two proteins interact, they are neighbors of each others. For an unannotated protein, the functions of its neighbors can tell us something about the function of the unannotated protein. For a given function, if most of the neighbors of a protein have the function, we are more likely to believe that the protein have the function. We want to associate each unannotated protein with a confidence (probability) or believe about the fact that the protein has the function.

For a given interaction network, how confident are we about the functional annotations of all the proteins? For an interaction pair, we are more likely to believe the interaction if both proteins have the function, followed by both proteins not having the function, and then only one protein having the function. ¿From the annotated proteins, we can also estimate how likely a protein have the function. From the above assumptions, we can assign a believe to each configuration of functional assignment—a believe network. That immediately leads us to the general theory of Markov random field. The problems are how to assign different weights to the parameters and how to estimate the probabilities based on the network.

Suppose a genome has N proteins $P_1, \cdots, P_N$ and M functional categories $F_1, \cdots, F_M$. Some proteins have already been studied and annotated and others are unannotated. Let $P_1, \cdots, P_n$ be the unannotated proteins and $P_{n+1}, \cdots, P_{n+m}$ be the annotated proteins, $N = n + m$. Through biological experiments, we also know the interaction status of the protein pairs which form a protein interaction network. Our objective is to assign functions to all the unannotated proteins based on functions of the annotated proteins and the protein interaction network.

A protein may have several different functions reflected in YPD [5]. In YPD, a single protein can have up to seven different functions. For interacting protein pairs with multiple functions, we do not know which combinations of the functions contribute to the interaction. To simplify the problem, we study each functional category separately. For a function of interest, let $X_i = 1$ if the i-th protein has the function and 0 otherwise. Let $X = (X_1, \cdots, X_{n+m})$ be the configuration of the functional labelings, 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 annotated. We infer the function of the unannotated proteins using the protein interaction network.

Several protein-protein interaction databases for yeast are available including data based on the yeast two-hybrid systems [14, 15, 29], the mass spectrometric analysis of protein complexes [9, 13], and physical interactions (MIPS) [22]. We can also use the predicted protein-protein interactions based on an analytical approach [6]. In this paper, we use the protein interaction data in MIPS.

Let O_{ij} be the variable for the observed interaction result for proteins P_i and P_j: $O_{ij} = 1$ if the interaction is observed and $O_{ij} = 0$ otherwise. Then the data we used is $O_{ij} = o_{ij}, \quad i, j = 1, \cdots, N$, where

$$o_{ij} = \begin{cases} 1 & \text{if } P_i \text{ and } P_j \text{ are observed to interact} \\ 0 & \text{otherwise.} \end{cases}$$

We only consider the interacting pairs. All the proteins together with the interaction information form a network, with proteins as nodes and interactions between proteins as edges. Let S be the collection of all the interacting pairs

$$S = \{P_i <-> P_j : o_{ij} = 1, \quad i, j = 1, \cdots, N\}$$

For each protein P_i, we define its neighbor, Nei(i), as the set of proteins directly interacting with P_i. Let π_j be the fraction of all proteins having function F_j. In summary, we have the following notations:

- P_i: the i-th protein, $i = 1, 2, \cdots, N$,
- Nei(i): neighbors of protein P_i, that is, the set of proteins interacting with protein P_i,
- F_j: the j-th function category, $j = 1, 2, \cdots, M$, and
- π_j: the fraction of all proteins having function F_j.

2.1. Available Methods

Several investigators developed methods to infer protein functions based on protein interaction network. Schwikowski et al. [27] proposed to infer the functions of an unannotated protein based on the frequencies of its neighbors having certain functions. They assign k functions to the unannotated protein with the k largest frequencies in its neighbors. This approach will be referred as the *neighboring counting method*. This approach does not consider the frequency of the proteins having a function among all the proteins. If a function is more common than other functions among all the proteins, the probability that an unannotated protein has this function should be higher than the

probability that it has other functions even if the protein does not have interaction partners.

Hishigaki et al. [12] developed another method to infer protein functions based on χ^2-statistics. For a protein P_i, let $n_i(j)$ be the number of proteins interacting with P_i and having function F_j. Let $e_i(j) = \#\text{Nei(i)} \times \pi_j$ be the expected number of proteins in Nei(i) having function F_j, where $\#$Nei(i) is the number of proteins in Nei(i). Define

$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)}.$$

For a fixed k, they assign an unannotated protein with k functions having the top k χ^2-statistics. Although this approach takes the frequency of the proteins having a function into consideration, $n_i(j)$ is generally small and the applicability of the χ^2-statistics is questionable.

The above approaches have been extended to l-neighbors, where two proteins are l-neighbors of each other if they are separated by at most $l-1$ proteins through interactions [27, 12]. Both methods treated all the l-neighbors equally in their analysis. To infer the functions of protein P_i, it is obvious that proteins far away from P_i contribute less information than those close neighbors. Less weight should be placed on proteins far away from protein P_i than the close neighbors. However it is not clear how to choose the correct weight in the above two approaches.

Here we develop a novel approach to infer the function of unannotated proteins based on the theory of Markov random fields (MRF) [18]. This approach overcomes all the above problems by considering the entire interaction network. Our approach considers the frequency of proteins having the function of interest as well as all the neighbors with less weight placed on far away neighbors than close neighbors. We calculate the probability that an unannotated protein has a function of interest. This probability indicates how confident we are about the assignment.

2.2. The new approach based on Markov random fields

Considering a function of interest, we want to assign this function to unannotated proteins. Let $x_i = 1$ if the i-th protein has the function and 0 otherwise. Let $X = (X_1, X_2, \cdots, X_N)$ be the functional annotation for all the proteins. We first give the prior probability distribution of X based on the interaction network, the *Gibbs distribution* [18]. In the following X_i will be the random variable and x_i will be its observed value. Conditional on the functional annotations of the annotated proteins, we calculate the posterior probability of the functional annotations of the unannotated proteins.

Let π be the probability of a protein having the function of interest. Without considering the interaction network, the probability of a configuration of X is proportional to

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

where $N_1 = \sum_{i=1}^{N} x_i$.

Next let us consider the interaction network. Studies have shown that the probability that a pair of interacting proteins have the same function is higher than the probability that they have different functions [27]. Therefore the probability of the network conditional on the functional labelings is proportional to

$$\exp(\beta N_{01} + \gamma N_{11} + N_{00}),$$

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

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

Therefore, the total probability of the functional labeling is proportional to $\exp(-U(x))$, where

$$\begin{aligned}
U(X) &= -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00} \\
&= -\alpha \sum_{i=1}^{N} x_i - \beta \sum_{(i,j)\in S} x_i x_j \\
&\quad - \gamma \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\
&\quad - \sum_{(i,j)\in S} (1-x_i)(1-x_j).
\end{aligned} \tag{1}$$

where $\alpha = \log(\frac{\pi}{1-\pi})$.

In the terminology of MRF, $U(X)$ is referred as the *potential function*. This potential function defines a global *Gibbs distribution* of the entire network,

$$\Pr(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(x)), \tag{2}$$

where $\theta = (\alpha, \beta, \gamma)$ are parameters and $Z(\theta)$ is a normalized constant which is calculated by summing over all the configurations,

$$Z(\theta) = \sum_{x} \exp(-U(x)).$$

$Z(\theta)$ is called the partition function in the general theory of MRF.

The Gibbs distribution defined in Equation 2 gives the prior distribution of the functional labeling for all the proteins in the protein interaction network. The data we have are the functional labeling of the annotated proteins, $(X_{n+1} = \mu_1, \cdots, X_{n+m} = \mu_m)$. The objective of the study is to find the posterior distribution of $(X_1, \cdots, X_n)$ given the data using Bayesian approach,

$$\Pr(X_1, \cdots, X_n \mid X_{n+1} = \mu_1, \cdots, X_{n+m} = \mu_m).$$

The posterior probability distribution of X_i can be obtained from the above equation by summing over all the possible configurations of $X_j,\ j \neq i,\ 1 \leq j \leq n$.

To achieve this objective, we use *Gibbs sampler* [19], a computational technique generally used in Bayesian statistics.

2.3. The Gibbs Sampler

To introduce the Gibbs sampler, we note that

$$\begin{aligned} &\Pr(X_i = 1 \mid X_{[-i]}, \theta) \\ =&\frac{\Pr((X_i = 1, X_{[-i]}) \mid \theta)}{\Pr((X_i = 1, X_{[-i]}) \mid \theta) + \Pr((X_i = 0, X_{[-i]}) \mid \theta)} \\ =&\frac{e^{\alpha+(\beta-1)M_0^{(i)}+(\gamma-\beta)M_1^{(i)}}}{1 + e^{\alpha+(\beta-1)M_0^{(i)}+(\gamma-\beta)M_1^{(i)}}}, \end{aligned} \tag{3}$$

where $X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{n+m})$, and $M_0^{(i)} = \#\{j \in \text{Nei(i)} : X_j = 0\}$, $M_1^{(i)} = \#\{j \in \text{Nei(i)} : X_j = 1\}$. $M_0^{(i)}$ and $M_1^{(i)}$ are the numbers of interaction partners of protein P_i labeled with 0 and 1, respectively. Equation 3 can be derived from Equation 2.

Equation 3 defines the local dependency of the network. When all the functions of the interaction partners of a protein are given, it can be used to derive the probability that the protein has the function, which is the basis of the Gibbs Sampler.

Assume that the parameters $\theta = (\alpha, \beta, \gamma)$ are given. For a given protein P_i, conditional on the functional labeling of all the other proteins, we can use the conditional probability $\Pr(X_i \mid X_{[-i]}, \theta)$ in Equation 3 to generate samples to update the functional labeling of protein P_i. Repeating this procedure many times will generate samples for the functional labeling of all the unannotated proteins. This is the Gibbs sampler strategy and is used as a core algorithm in this paper.

2.4. Parameter Estimation

In practice, we do not know the parameters $\theta = (\alpha, \beta, \gamma)$. Here we propose a method to estimate the parameters based on the functions of the annotated proteins. Consider the subnetwork of all the annotated proteins. i.e,

$$S' = \{P_i <-> P_j : o_{ij} = 1, \quad i, j = n+1, \cdots, n+m\}.$$

We estimate the parameters based on this subnetwork.

It's difficult to use the maximum likelihood estimation (MLE) directly since the partition function $Z(\theta)$ in Equation 2 is also a function of parameters. Here we use the quasi-likelihood approach that has been used in image analysis [18]. From Equation 3, we have

$$\begin{aligned} &\log \frac{\Pr(X_i = 1 | X_{[-i]}, \theta)}{1.0 - \Pr(X_i = 1 | X_{[-i]}, \theta)} \\ =&\alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)}, \end{aligned} \tag{4}$$

where M_0^i and $M_1^{(i)}$ are the numbers of interaction partners for protein P_i labeled with 0 or 1, respectively.

The quasi-likelihood estimation method is to estimate the parameters based on standard linear logistic model treating the observations as independent. It is known that the functional labeling of the proteins in the network are not independent and thus the quasi-likelihood approach is not a MLE approach. In image analysis, it has been shown that the quasi-likelihood approach gives reasonably good results in practice [18].

2.5. Bayesian analysis

For a function of interest, first we estimate the probability, π, that a protein has the function (without the information on interaction network) by the fraction of all the proteins having that function.

Secondly, we estimate the parameters $\theta = (\alpha, \beta, \gamma)$ using the quasi-likelihood approach based on linear logistic regression that is outlined above.

With the above parameters, we have the following algorithm.

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

In Gibbs sampling, we need to specify the "burn-in-period" and the "lag-period". The burn-in-period is the time we wait until the Markovian process is stablized and the simulation results in the burn-in period are discarded to reduce or eliminate the effect of intial values. After the burn-in-period, we approximate the probability that an unannotated protein has the function by averaging the simulations

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. We repeat this process for every functional category and the probability that an unannotated protein has the function is estimated.

3. Results

We apply our approach to infer the functions of unannotated proteins in Yeast. We use the functional annotations from YPD. In YPD, proteins are assigned functions based on three criteria: "cellular role", "subcellular localization", and "biochemical function". Up to February 15, 2002, YPD includes 6281 proteins. In this paper, we will consider functional annotation based on cellular role. The results based on "subcellular localization" and "biochemical function" will be presented in the full paper. There are 43 functional categories based on cellular role including a category termed "others" including all the proteins whose function does not belong to the other 42 functional categories. The numbers of annotated and unannotated proteins based on cellular roles for all the proteinss and proteins with at least one to six interaction partners are given in Table 1.

For protein interactions, we use the MIPS physical interaction data consisting of 2442 interaction pairs involving 1877 proteins. The average number of interaction partners per protein is about 2.6. For unannotated proteins without interaction partners, the probability of having a function of interest equals the fraction of proteins having that function among all the proteins.

3.1. Functional annotation based on cellular role

We apply our Bayesian method to predict protein functions based on cellular roles. The parameters can be estimated by the quasi-likelihood approach described above, using the interaction network consisting of only the annotated proteins. The computation is done using SPLUS [30]. The results are given in Table 2. Note that $\alpha = \log(\pi/(1-\pi))$ with π being the fraction of proteins having the function of interest. π is generally small and thus α should be negative. $\beta-1$ is the contribution of an interaction partner not having the function to the log-odds of having the function for the protein of interest. Thus, $\beta - 1$ should be negative. $\gamma - \beta$ is the contribution of an interaction partner having the function to the log-odds of having the function for the protein of interest. Thus, $\gamma - \beta$ should be positive. Except for functional categories 4 ("Cell adhesion"), 20 ("Mitochondrial transcription"), and 40 ("Septation"), all the other functional categories satisfy the above conditions. We check the three exceptional cases and find that the numbers of proteins having the corresponding functions are very small, 4, 4 and 1 for "Cell adhesion", "Mitochondrial transcription", and "Septation", respectively. Therefore the estimated parameters are not accurate. In the following, we will ignore those three functional categories.

Although the main objective is to estimate the posterior probability that a protein has a function of interest, we can also assign functions to an unannotated protein if the posterior probability is above a certain threshold.

The accuracy of the predictions is measured by the leave-one-out method. The method randomly selects an annotated protein and assumes it as unannotated. Then we predict its functions by the above methods. We then compare the predictions with the annotations of the protein. We repeat the leave-one-out experiment for K proteins, $P_i, \cdots, P_K$. Let n_i be the number of functions for protein P_i in YPD, m_i be the number of *predicted* functions for protein P_i, and k_i be the overlap between the set of observed functions and the set of predicted functions. The specificity (SP) and the sensitivity (SN) can be defined as

$$\begin{aligned} SP &= \frac{\sum_i^K k_i}{\sum_i^K m_i} \\ SN &= \frac{\sum_i^K k_i}{\sum_i^K n_i} \end{aligned} \tag{5}$$

Figure 1 shows the relationship between specificity and sensitivity of our approach using different thresholds for posterior probabilities. With the threshold equal to 0.17, the specificity and the sensitivity are roughly the same and equal 47.0%. It should be noted that the functional annotations for the annotated proteins are not complete. If a protein has a function based on YPD, we have high confidence for the assignment. On the other hand, if a protein does not have a function based on YPD, the protein may have the function but has not been experimentally verified. Thus we might wish to lower the specificity to increase sensitivity by lowering the threshold.

3.2. Comparison with other methods

For comparison, we implement the neighborhood-counting method [27] and the χ^2 method [12] for functional annotation. We choose the top 1, 2, 3, 4 and 5 functions, respectively, and assign these functions to each unannotated protein. Figure 2 shows the relationship between sensitivity and specificity for the three different methods discussed above: the Bayesian method, the χ^2 method, and the neighborhood-counting method. The figure indicates that for any given specificity, the sensitivity of the the Bayesian method is higher than the sensitivities of the neighborhood-counting method and the χ^2 method. Our new approach outperforms the other two approaches for functional annotation.

	All proteins	≥ 1 partner	≥ 2 partners	≥ 3 partners	≥ 4 partners	≥ 5 partners	≥ 6 partners
Annotated	3854	1455	785	514	346	252	186
Unannotated	2427	422	131	45	15	12	7
Total	6281	1877	916	559	361	264	193

Table 1. The numbers of annotated and unannotated proteins for all the proteins and proteins with at least one to six interaction partners.

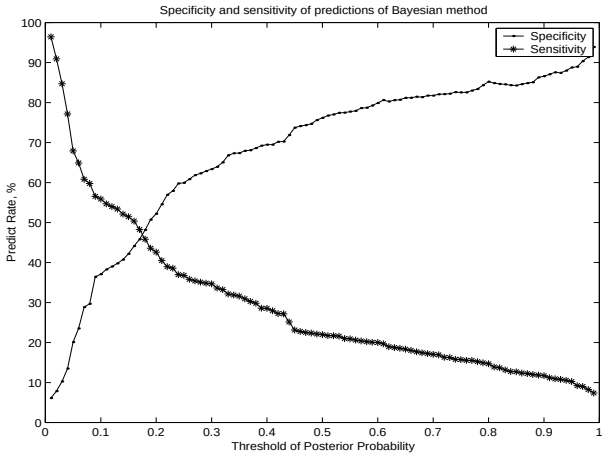

Figure 1. Specificity and sensitivity of Bayesian predictions for different thresholds.

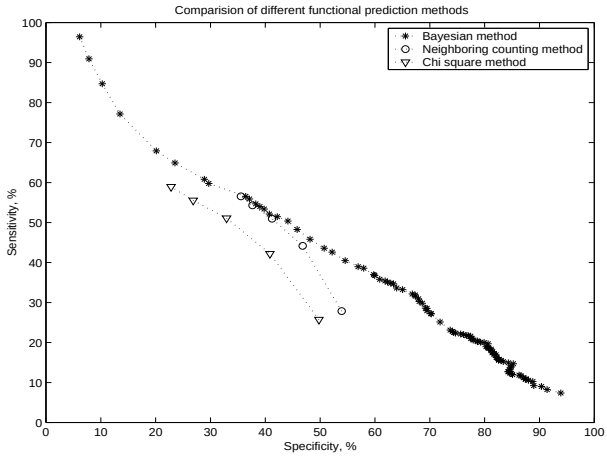

Figure 2. Sensitivity and specificity of predictions for three different methods.

We further analyze the prediction results of the Bayesian method by applying the leave-one-out measure on proteins having at least one interaction partner, at least two interaction partners, and so on. The corresponding relationship for specificities and sensitivities are shown in Figure 3. As expected, for a given specificity, the sensitivity increases with the number of interaction partners. The more interaction partners a protein has, the more accurate our prediction is.

The Bayesian method is a global approach to estimate the posterior probabilities of protein functions. Not only do we use the annotation of direct interaction partners, this approach also use information from indirect interaction partners. For example, consider the following interaction network shown in Figure 4. Table 3 lists the functions of the annotated proteins in this network. Using direct interaction partners, it is impossible to infer the functions for protein YDR084C since its two direct interaction partners YGL161C and YGL198W are both unannotated. However, from the indirect interaction partners, specifically, the partners of YGL161C, which share the same function 43, "versicular transport", we can predict that YDR084C has the "versicular transport" function with probability 0.8496. The situation is the same for protein YGL198W. Protein YGL161C has four annotated interaction partners with the "versicular transport" function and four unannotated interaction partners. The estimated probability that protein YGL161C has the function is approximately 1. Proteins YDR100D and YPL246C have two and three interaction partners with the "versicular transport" function, respectively. The estimated probabilities for both proteins are 0.9956. These estimated probabilities indicate how confident we are about the assignment.

4. Discussions

We develop a novel approach for function prediction of unannotated proteins based on the protein-protein interaction network and the functional annotations of annotated proteins. Unlike other available function predication methods where they predict whether a protein has a function or not, we estimate the posterior probability that the protein has the function of interest. The posterior probability indicates how confident we are about assigning the function to the protein. The distinction of the Bayesian approach we

Function	α	$\beta-1$	$\gamma-\beta$	Function	α	$\beta-1$	$\gamma-\beta$
1	-3.9879	-0.3172	2.7341	2	-2.6456	-0.4714	1.8360
3	-2.8112	-0.1814	1.1745	4	-8.3204	0.2217	-3.4507
5	-2.5080	-0.1144	0.9728	6	-3.6809	-0.2022	1.9735
7	-2.5806	-0.1035	1.0481	8	-3.0467	-0.2827	1.6667
9	-3.7773	-0.0297	1.2879	10	-2.2585	-0.1909	0.8392
11	-4.0458	-0.1524	1.6594	12	-3.0164	-0.3258	2.1116
13	-4.0479	-0.0892	2.4368	14	-3.7228	-0.0231	1.1739
15	-2.7456	-0.3547	1.6954	16	-3.7361	-0.4455	3.2861
17	-3.0650	-0.1330	1.1784	18	-2.8717	-0.1497	1.3777
19	-4.1841	-0.4124	2.2684	20	-4.5592	-1.9361	-2.7715
21	-3.3293	-0.1135	1.8997	22	-3.9139	-0.2314	2.5797
23	-3.6166	-0.5120	3.1767	24	-4.5016	-0.1784	2.3734
25	-3.1298	-0.2882	1.4582	26	-5.5494	-0.1173	4.5037
27	-5.1278	-0.1519	3.9724	28	-1.7856	-0.4585	1.4175
29	-4.3443	-0.3402	2.8155	30	-4.8546	-0.0992	2.9012
31	-2.8442	-0.2882	1.6762	32	-2.4807	-0.9796	1.9139
33	-3.0611	-0.1834	1.7934	34	-2.5008	-0.5635	2.1662
35	-2.7185	-0.9655	2.4446	36	-2.7782	-0.2036	1.4811
37	-3.5689	-0.0903	1.3102	38	-3.8578	-0.1769	1.0905
39	-3.5124	-0.2164	2.1625	40	11.2029	0.0000	–
41	-3.3061	-0.1664	1.6291	42	-1.9998	-0.7523	1.6539
43	-2.7470	-0.6196	2.7236				

Table 2. The estimated parameters $\alpha, \beta-1,$ and $\gamma-\beta$ using linear logistic regression for the 43 functional categories based on cellular role.

develop here is that it is a global approach taking all the interaction network and the functions of annotated proteins into consideration.

We apply our approach to the interaction network of yeast proteins in MIPS and the protein function annotations based on YPD. We study the sensitivity and specificity of our method by the leave-one-out approach and compare the results with the χ^2 method and the neighbor-counting method. We show that, for a given specificity, the sensitivity of our new approach is higher than the sensitivities of the other two approaches. Because not all the functions have been identified even for the annotated proteins, we may wish to sacrifice specificity to increase sensitivity. We also apply our approach to proteins with at least two or more interaction partners. As expected, for any given specificity, the sensitivity increases with the number of interaction partners.

There are several limitations of 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 obtained in MIPS. For a conservative estimate, if we assume that each protein interacts with on average five other proteins, we would expect about $6,000 \times 5/2 = 15,000$ interactions, much higher than the 2442 interactions in MIPS. With the advance of other high-throughput technologies for detecting protein-protein interactions, our understanding of the protein interaction network will be more complete.

Our method treats each function independently and separately because it is known that a protein may be involved in multiple functions. Generally, a protein having 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. A protein having function A may increase the chance of it having function B because functions A and B are highly correlated, for example, functional category 36 ("RNA processing/modification") and functional category 37 ("RNA splicing"). How to incorporate these information into a generalized model remains a challenging task. Our model assumes that annotated proteins have complete functional annotations, and predicts functions for unannotated proteins using these information. In reality, we know that these annotated proteins may have other functions that have not been determined. As biologists continue experimentally determining the functions of proteins, the functional annotations will be more and more complete.

Despite the limitations, we show that the results from our

VAM7	Subunit of the vacuolar SNARE complex involved in morphogenesis of the vacuole; homologous to SNAP-25.
YHR105W	Protein that may play a role in vesicular transport, has similarity to Grd19p, bacterial helix-turn-helix regulator protein of the argR group, and human SNX1.
YIP1	Protein involved in vesicular transport, interacts with transport GTPases Ypt1p and Ypt31p at the Golgi membrane.
BET1	Synaptobrevin (v-SNARE) homolog present on ER vesicles recycling from Golgi.
PEP12	Syntaxin homolog (t-SNARE) involved in Golgi to vacuole transport.
AKR2	Protein involved in constitutive endocytosis of Ste3p.
YIF1	Component of COPII vesicles, has similarity to NADH dehydrogenases.
KTR3	Alpha-1,2-mannosyltransferase of the KRE2 family.

Table 3. Functions of the annotated proteins in Figure 4.

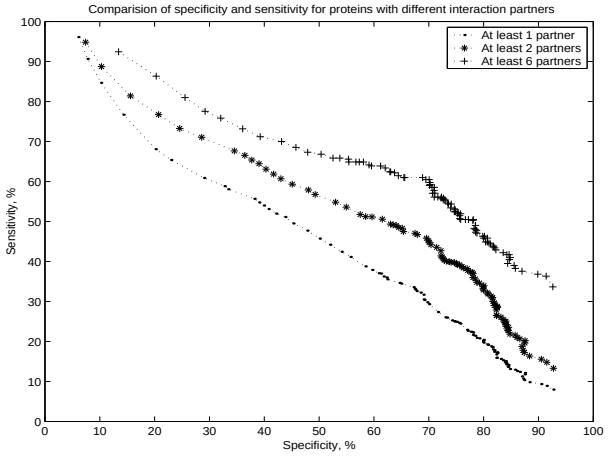

Figure 3. The relationship between sensitivity and sensitivity for proteins with at least one, two, and six interaction partners using the Bayesian method. The corresponding numbers of proteins with one to six interaction partners are given in table 1.

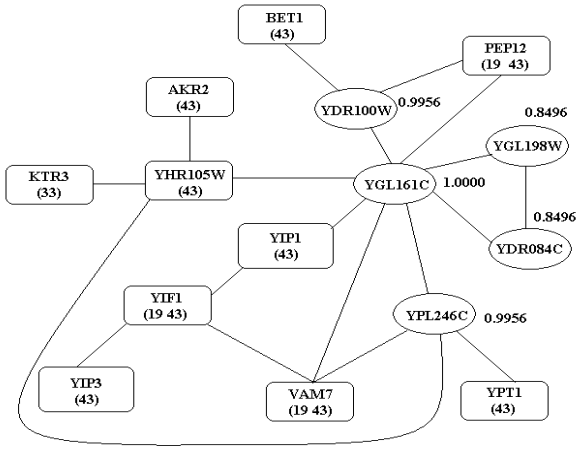

Figure 4. An example of a protein-protein interaction subnetwork. Proteins in rectangle are annotated; the numbers in the parentheses are the functional categories of the proteins. The proteins in circle are unannotated; values beside the circles are the posterior probabilities that the unannotated proteins belong to functional category 43, "versicular transport".

approach are reasonably good. The probabilities of protein functions in Figure 4 show a very important and desirable feature of our model: the impact of a protein's function on unannotated proteins decreases as these proteins are farther away from the protein in the interaction network. This feature could not be obtained in local approaches such as the neighborhood-counting method and the χ^2 method.

References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** $3389 - 3402$.

[2] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., Hogue, C.W.. 2001. BIND–The Biomolecular interaction network database. Nucleic Acids Research **29:** $242 - 245$.

[3] Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. Jr., and Haussler, D.. 2000. Knowledge-based analysis of microar-

ray gene expression data by using support vector machines Proc. Natl. Acad. Sci. USA **97:** $262-267$.

[4] Clare, A. and King, R.D.. 2002. Machine learning of functional class from phenotype data. Bioinformatics **18:** $160-166$.

[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., Garrels, J.I. 2001. YPD^{TM}, PombePD^{TM}, and WormPD^{TM}: model organism volumes of the BioKnowledge library, an integrated resource for protein information. Nucleic Acids Research **29:** $75-79$.

[6] Deng, M., Mehta, S., Sun, F. and Chen, T.. 2002. Inferring domain-domain interaction from protein-protein interaction. Will appeared in *Proceedings of the Sixth International Conference on Computational Molecular Biology (RECOMB2002)*.

[7] Eisen, M.B., Spellman, P.T., Brown, P.O. and Bostein D.. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**,$14863-14868$.

[8] Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.W. and Hani J.. 2000. In *Proc. of the Eighth Int. Conf. on Intelligent System for Molecular Biology (ISMB2000)*: $152-161$.

[9] Gavin, A., B*ö*che, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415:** $141-147$

[10] Hazbun, T.R. and Fields, S.. 2001. Networking proteins in yeast. Proc. Natl. Acad. Sci. USA **98**: $4277-4278$.

[11] Gerstein, M., Lan, N. and Jansen, R.. 2002. Integrating interactomes. Science **295**:$284-287$.

[12] Hishigaki H., Nakai K., Ono T., Tanigami A. and Takagi T.. 2001. Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast **18**: $523-531$.

[13] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al.. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415:** $180-183$.

[14] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto,K., Kuhara, S., Sakaki, Y.. 2000. 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**: $1143-1147$.

[15] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.. 2001. A Comprehensive two hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**: $4569-4574$.

[16] Kell, D.B. and King, R.D.. 2000. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. Trends Biotechnol., **18:** $93-98$.

[17] King, R.D., Karwath, A., Clare, A. and Dehaspe, L.. 2001. The utility of different representations of protein sequence for predicting functional class. Bioinformatics **17:** $445-454$.

[18] Li, S.Z.. 1995. Markov random field modeling in Computer vision. Springer-Verlag: Tokyo.

[19] Liu, J.S.. 1995. Monte Carlo strategies in scientific computing. Springer-Verlag: New York.

[20] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice,D.W., Yeates, T.O. and Eisenberg, D.. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285**: $751-753$.

[21] 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**: $83-86$.

[22] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Research **30**: $31-34$.

[23] Pavlidis, P. and Weston, J.. 2001. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001):* $249-255$.

[24] Pearson,W.R. and Lipman,D.J.. 1988. Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA, **85**, 24442448.

[25] Pellegrini, M., Marcotte, E.M., Thompson,M.J., Eisenberg, D. and Yeates, T.O.. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96**: $4285-4288$.

[26] Saito, R., Suzuki, H. and Hayashizaki, Y.. 2002. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. Nucleic Acids Research **30:** $1163 - 1168.$

[27] Schwikowski, B., Uetz, P. and Fields, S.. 2000. A network of protein-protein interactions in yeast. Nature Biotechnology **18**:$1257 - 1261.$

[28] Tong A.H.Y., Drees, B., Nardelli, G., Bader G.D., Brannetti, B., Castagnoli, L. Evangelista, M., Paoluzi, S., Quondam, M., Zucconim A, et al.. 2002. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295**: $321 - 324.$

[29] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, et al. 2000. A Comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature **403**: $623 - 627.$

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

[31] Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S., Eisenberg, D. 2002. DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nucleic Acids Research **30:** $303 - 305.$