JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 6, 2003
© Mary Ann Liebert, Inc.
Pp. 947–960

Prediction of Protein Function Using Protein–Protein Interaction Data

MINGHUA DENG, KUI ZHANG, SHIPRA MEHTA, TING CHEN,
and FENGZHU SUN

ABSTRACT

Assigning functions to novel proteins is one of the most important problems in the postgenomic era. Several approaches have been applied to this problem, including the analysis of gene expression patterns, phylogenetic profiles, protein fusions, and protein–protein interactions. In this paper, we develop a novel approach that employs the theory of Markov random fields to infer a protein's functions using protein–protein interaction data and the functional annotations of protein's interaction partners. For each function of interest and protein, we predict the probability that the protein has such function using Bayesian approaches. Unlike other available approaches for protein annotation in which 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 employ our method to predict protein functions based on "biochemical function," "subcellular location," and "cellular role" for yeast proteins defined in the Yeast Proteome Database (YPD, *www.incyte.com*), using the protein–protein interaction data from the Munich Information Center for Protein Sequences (MIPS, *mips.gsf.de*). We show that our approach outperforms other available methods for function prediction based on protein interaction data. The supplementary data is available at *www-hto.usc.edu/~msms/ProteinFunction.*

Key words: protein–protein interaction, pretein function, Markov random field, Bayesian method, Gibbs sampler.

1. INTRODUCTION

With the completion of genome sequencing of several model organisms, the functional annotation of the proteins is of most importance. Up to April 8, 2002, the Yeast Protein Database (YPD) (Costanzo *et al.*, 2001) listed 6,416 proteins in three functional categories—"biochemical function," "subcellular location," and "cellular role"— with almost half of the proteins being unannotated for each category (see Table 1). Throughout this paper, we will use the protein function annotations from YPD. A challenging task that lies ahead is to discover the functional roles of the unannotated proteins. Several research groups have developed methods for protein function prediction. The classical way is to find homologies

Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113.

Table 1. The Numbers of Annotated and Unannotated Proteins for All Proteins and Proteins with at Least One to Six Interaction Partners Based on Three Functional Categories

	# of interaction partners						
	≥ 0	*≥ 1*	*≥ 2*	*≥ 3*	*≥ 4*	*≥ 5*	*≥ 6*
Biochemical function							
Annotated	3353	1257	687	433	284	198	146
Unannotated	3063	618	228	126	76	66	47
Subcellular location							
Annotated	3181	1237	698	455	305	225	168
Unannotated	3235	638	217	104	55	39	25
Cellular role							
Annotated	3894	1469	789	515	345	252	186
Unannotated	2522	406	126	44	15	12	7
Total	6416	1875	915	559	360	264	193

between a protein and other proteins in protein databases using programs such as FASTA (Pearson *et al.*, 1988) and PSI-BLAST (Altschul *et al.*, 1997) and then predict functions based on sequence homologies. Another sequence-based approach is called the “Rosetta stone method," in which two proteins are inferred to interact and thus have similar functions if they are together in another genome (Marcotte *et al.*, 1999a). By comparing a number of sequenced genomes, the phylogenetic pattern (the presence or 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 (Marcotte *et al.*, 1999b) based on phylogenetic patterns.

The development of high-throughput bio-techniques and their applications in many areas of biology has 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 phenotypes, and protein–protein interactions. Clustering analysis of gene expression data can be used to predict functions of unannotated proteins based on the idea that coexpressed genes are more likely to have similar functions (Brown *et al.*, 2000; Eisen *et al.*, 1998; Pavlidis *et al.*, 2001). In addition, functional predictions have been modeled as pattern recognition problems based on sequence homologies and structural information (Kell *et al.*, 2000; King *et al.*, 2001) as well as phenotype data (Clare *et al.*, 2002).

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 chi-square statistics (Hishigaki *et al.*, 2001) and on frequencies of interaction partners having certain functions of interest (Fellenberg *et al.*, 2000; Schwikowski *et al.*, 2000) have been used to assign functions to unannotated proteins. However, these methods lack a systematic mathematical 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 probability that an unannotated protein has certain functions of interest. We employ our approach to predict functions of unannotated proteins based on “biochemical function," “subcellular location," and “cellular role."

2. METHOD

We will first discuss 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 other. For an unannotated protein, the functions of its neighbors contain information 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 itself has the function. We want to associate each unannotated protein with a confidence (probability) or belief 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 most likely to be confident 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 it is that a protein has the function. From the above assumptions, we can assign a belief to each configuration of functional assignment—a belief network. That immediately leads us to the general theory of Markov random fields. 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, $P_1, \ldots, P_N$, and M functional categories, $F_1, \ldots, F_M$. Some proteins have already been studied and annotated, and others are unannotated. Let $P_1, \ldots, P_n$ be the unannotated proteins and $P_{n+1}, \ldots, P_{n+m}$ be the annotated proteins, with $N = n + m$. Through biological experiments, we also know the interaction status of the protein pairs that form a protein interaction network. Our objective is to assign functions to all of the unannotated proteins based on functions of the annotated proteins and the protein interaction network.

A protein may have several different functions. For example, in YPD (Costanzo *et al.*, 2001), a single protein can have up to eight different cellular roles. 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 $X = (X_1, \ldots, X_{n+m})$ be the configuration of the functional labeling, where $X_1 = \lambda_1, \ldots, X_n = \lambda_n$ are unknown and $X_{n+1} = \mu_1, \ldots, X_{n+m} = \mu_m$ 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 (Ito *et al.*, 2000; Ito *et al.*, 2001; Uetz *et al.*, 2000) and the mass spectrometric analysis of protein complexes (Gavin *et al.*, 2002; Ho *et al.*, 2002). However, the interaction data from these high-throughput experiments have high false positive rates and are not highly reliable (Mrowka *et al.*, 2001; Deane *et al.*, 2002; Deng *et al.*, 2003). The Munich Information Center for Protein Sequences' (MIPS) (Mewes *et al.*, 2002) physical interaction data include interactions collected from small-scale experiments and the core data of Ito *et al.* (2000, 2001), and they are believed to be highly reliable. Therefore, we use the protein physical interaction data in MIPS in this study.

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 are $O_{ij} = o_{ij}, \quad i, j = 1, \ldots, N$, 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 consider only 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. We then have

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

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, $i = 1, 2, \ldots, N$,
- 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, $j = 1, 2, \ldots, M$, and
- π_j: the fraction of all proteins having function F_j.

2.1. Available methods

Several investigators have developed methods to infer protein functions based on protein interaction networks. Schwikowski *et al.* (2000) proposed to infer the functions of an unannotated protein based on the frequencies of its neighbors having certain functions. They assigned 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 method 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.* (2001) developed another method for inferring protein functions based on chi-square 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, the authors assigned an unannotated protein with k functions having the top k chi-square 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 chi-square statistics is questionable.

The above approaches have been extended to n-neighbors, where two proteins are n-neighbors of each other if they are separated by at most $n - 1$ proteins through interactions (Schwikowski *et al.*, 2000; Hishigaki *et al.*, 2001). Both methods treated all of the n-neighbors equally in their analysis. To infer the functions of protein P_i, it must be recognized that proteins far away from P_i contribute less information than those that are close neighbors. In other words, less weight should be placed on proteins far away from protein P_i than on the close neighbors. However, it is not clear how to choose the correct weight in the above two approaches.

2.2. The new approach based on Markov random fields

Here we develop a novel approach to inferring the function of unannotated proteins based on the theory of Markov random fields (MRF) (Li, 1995). This approach overcomes all of 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 neighbors, with less weight being placed on far away neighbors than close neighbors. We calculate the probability that an unannotated protein has a function of interest, and this probability indicates how confident we are about the assignment.

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 $X = (X_1, X_2, \ldots, X_N)$ be the functional annotation for all of the proteins. We first give the prior probability distribution of X based on the interaction network—the Gibbs distribution (Li, 1995). In the following, X_i will be the random variable and x_i will be its observed value. Conditional on the functions of the annotated proteins, we calculate the posterior probability of the functions 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

(Schwikowski *et al.*, 2000). Therefore, the probability of the network conditional on the functional labeling 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

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

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

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

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

In the terminology of MRF, $U(x)$ is referred as the *potentialfunction.* 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 that 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 of the proteins in the protein interaction network. The data we have are the functional labeling of the annotated proteins, $(X_{n+1} = \mu_1, \ldots, X_{n+m} = \mu_m)$. The objective of the study is to find the posterior distribution of $(X_1, \ldots, X_n)$ given the data using a Bayesian approach:

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

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

To achieve this objective, we use a Gibbs sampler (Liu, 2001), 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 $X_{[-i]} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+m})$, $M_0^{(i)} = \#\{j \in \text{Nei(i)} : X_j = 0\}$, and $M_1^{(i)} = \#\{j \in \text{Nei(i)} : X_j = 1\}$. Quantities $M_0^{(i)}$ 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 of the functions of the interaction partners of a protein are given, the equation 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 of 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 generates samples for the functional labeling of all of the unannotated proteins. This is the Gibbs sampler strategy, which 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 for estimating the parameters based on the functions of the annotated proteins. Consider the subnetwork of all of the annotated proteins. In other words,

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

We estimate the parameters based on this subnetwork.

It is difficult to use the maximum likelihood estimation (MLE) method directly because the partition function $Z(\theta)$ in Equation 2 is also a function of the parameters. Here, we use the quasi-likelihood approach that has been used in image analysis (Li, 1995). From Equation 3, we have

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

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 consists of estimating the parameters based on a standard linear logistic model that treats the observations as independent. It is known that the functional labels of the proteins in the network are not independent, and thus the quasi-likelihood approach is not an MLE approach. In image analysis, it has been shown that the quasi-likelihood approach gives reasonably good results in practice (Li, 1995).

2.5. Bayesian analysis

For a function of interest, we first 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. Second, 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 $X_i = \lambda_i, i = 1, \ldots, n$ with probability π.
2. For each protein P_i, using Equation 3, update the value of X_i.
3. Repeat the second step T times until all of 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 stabilized, and the simulation results in the burn-in period are discarded in order to reduce or eliminate the effect of initial values. After the burn-in period, we approximate the probability that an unannotated protein has the function by averaging the simulation results in steps of the lag period in order 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 steps, respectively. The total number of simulations is 2,000 steps. We repeat this process for every functional category, and the probability that an unannotated protein has a potential function is estimated.

3. RESULTS

We employ our approach to infer the functions of unannotated proteins in Yeast, using the functional annotations from YPD. In YPD, proteins are assigned functions based on three criteria: biochemical function, subcellular location, and cellular role. Up to April 8, 2002, YPD included 6,416 proteins. In this paper, we consider functional annotation based on these three functional categories. The numbers of annotated and unannotated proteins based on different functional categories for all of the proteins 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 2,439 interaction pairs (excluding 120 pairs of self-interactions) involving 1,877 proteins. The average number of interaction partners per protein is about 2.6.

3.1. *Functional annotation based on YPD function categories*

We apply our Bayesian method to predict protein functions based on the three YPD function categories. The parameters can be estimated with the quasi-likelihood approach described above, using the interaction network consisting of only the annotated proteins. The computation is done using S-Plus (Venables *et al.*, 1996). Note that $\alpha = \log(\pi/(1 - \pi))$ with π being the fraction of proteins having the function of interest. Generally, π is small, and thus α should be negative. The quantity $\beta - 1$ is the contribution of an interaction partner not having the function to the log-odds of the protein of interest having the function. Thus, $\beta - 1$ should be negative. The quantity $\gamma - \beta$ is the contribution of an interaction partner having the function to the log-odds of the protein of interest having the function. Thus, $\gamma - \beta$ should be positive. For the three functional categories, the above observations hold for 79%, 79%, and 93% of the functions based on biochemical function, subcellular location, and cellular role, respectively (see tables in supplementary materials). The condition is violated where either a small number of proteins have the function of interest or some classes have more interclass interactions than intraclass interactions. For example, based on cellular role, all of the other function classes satisfy the above conditions except for classes 4 (cell adhesion), 20 (mitochondrial transcription), and 40 (septation). 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 the functional classes in which the above conditions are not satisfied.

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. For each annotated protein with at least one annotated interaction partner, we assume it to be unannotated and 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 all such proteins $P_1, \ldots, P_K$. Let n_i be the number of functions for protein P_i in YPD, m_i the number of *predicted* functions for protein P_i, and k_i 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

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

The corresponding values of K for biochemical function, subcellular location, and cellular role are 1,128, 1,133, and 1,398, respectively. Figure 1 shows the relationship between the specificity and sensitivity of our approach using different thresholds for posterior probabilities. With the threshold equal to 0.13, 0.25, and 0.17 for biochemical function, subcellular location, and cellular role, respectively, the corresponding specificity and sensitivity are roughly the same and equal to 45%, 64%, and 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 in the assignment. On the other hand, if a protein does not have a function based on YPD, the protein may have the function, but that 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 neighboring counting method (Schwikowski *et al.*, 2000) and the chi-square method (Hishigaki *et al.*, 2001) 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 chi-square method, and the neighboring counting method. The figure indicates that for any given specificity, the sensitivity of the Bayesian method is higher than the sensitivities of the neighboring counting method and the chi-square method for all of the three functional categories. Our new approach outperforms the other two approaches for functional annotation.

We further analyze the prediction results of the Bayesian method by applying the leave-one-out measure to proteins having at least one interaction partner, at least two interaction partners, and so on. The corresponding relationships for specificities and sensitivities are shown in Fig. 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 are our predictions.

3.3. *Novel predictions*

The Bayesian method is a global approach to estimating the posterior probabilities of protein functions. Not only do we use the annotation of direct interaction partners, but we also use information from indirect interaction partners. For example, consider the interaction network shown in Fig. 4. Using direct interaction partners as in the neighbor counting method and the chi-square method, it is impossible to infer the functions for protein YDR084C because 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, vesicular transport—we can predict that YDR084C has the vesicular transport function with probability 0.8496. The situation is the same for protein YGL198W. Protein YGL161C has four annotated interaction partners with the vesicular 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 vesicular transport function, respectively. The estimated probabilities for both proteins are 0.9956, and these estimated probabilities indicate how confident we are about the assignment.

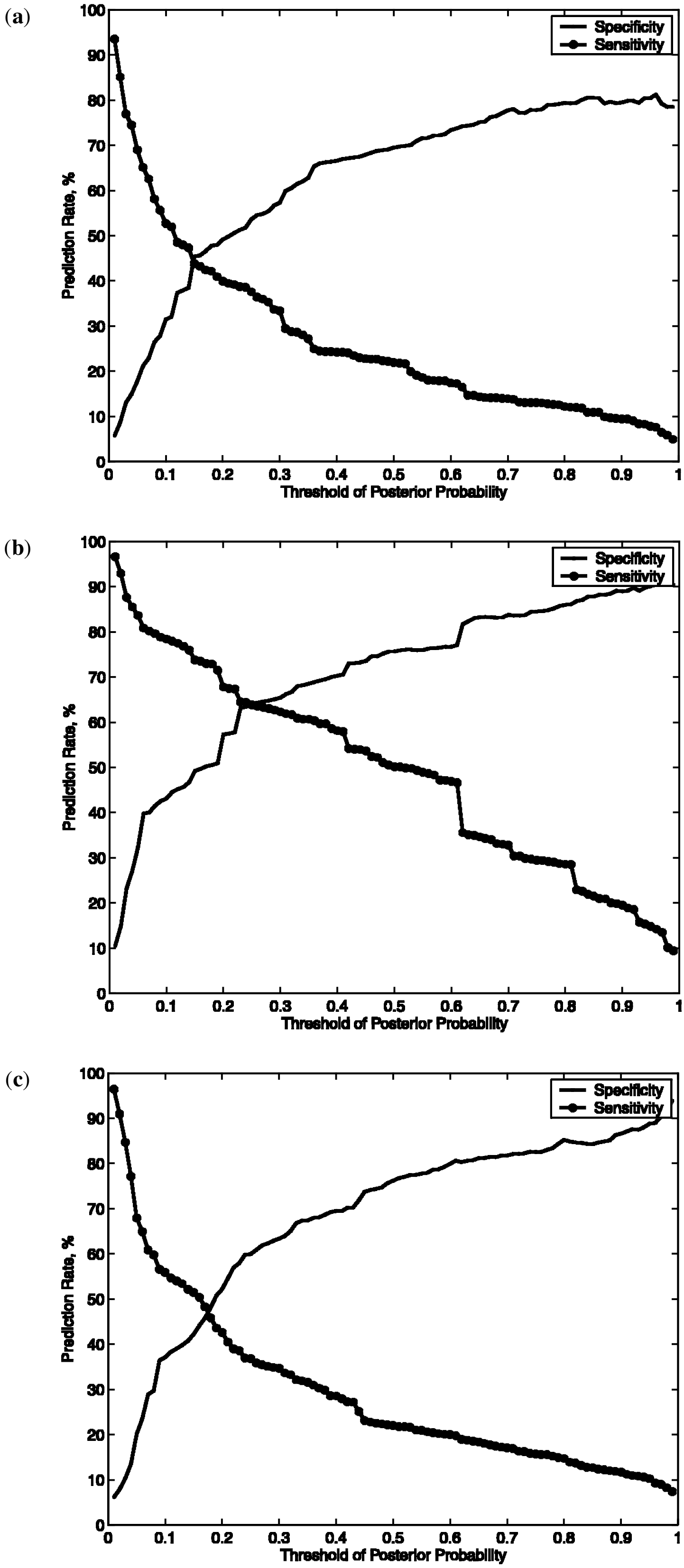

FIG. 1. Specificity and sensitivity of Bayesian predictions for different thresholds based on (**a**) biochemical function, (**b**) subcellular location, and (**c**) cellular role.

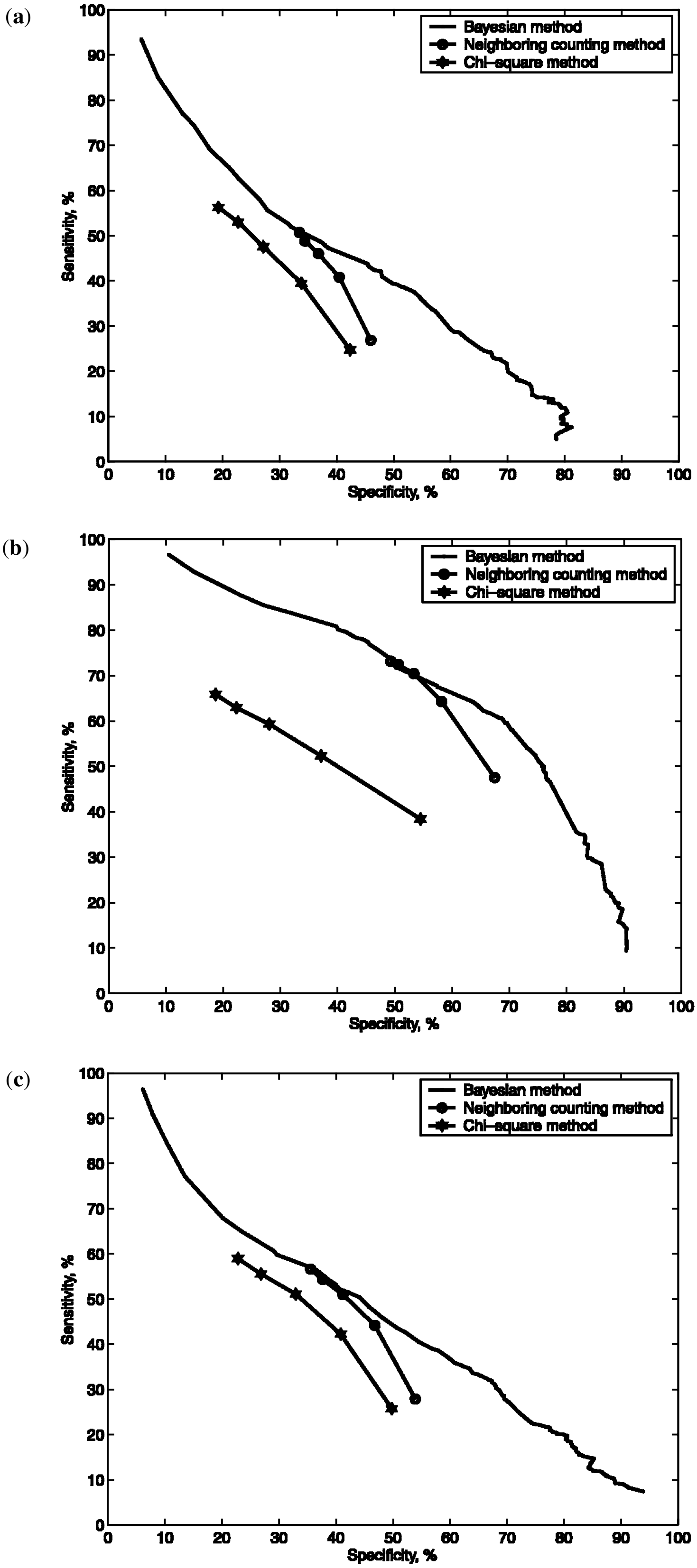

FIG. 2. Sensitivity and specificity of predictions for the neighboring counting, chi-square, and the Bayesian methods based on (**a**) biochemical function, (**b**) subcellular location, and (**c**) cellular role.

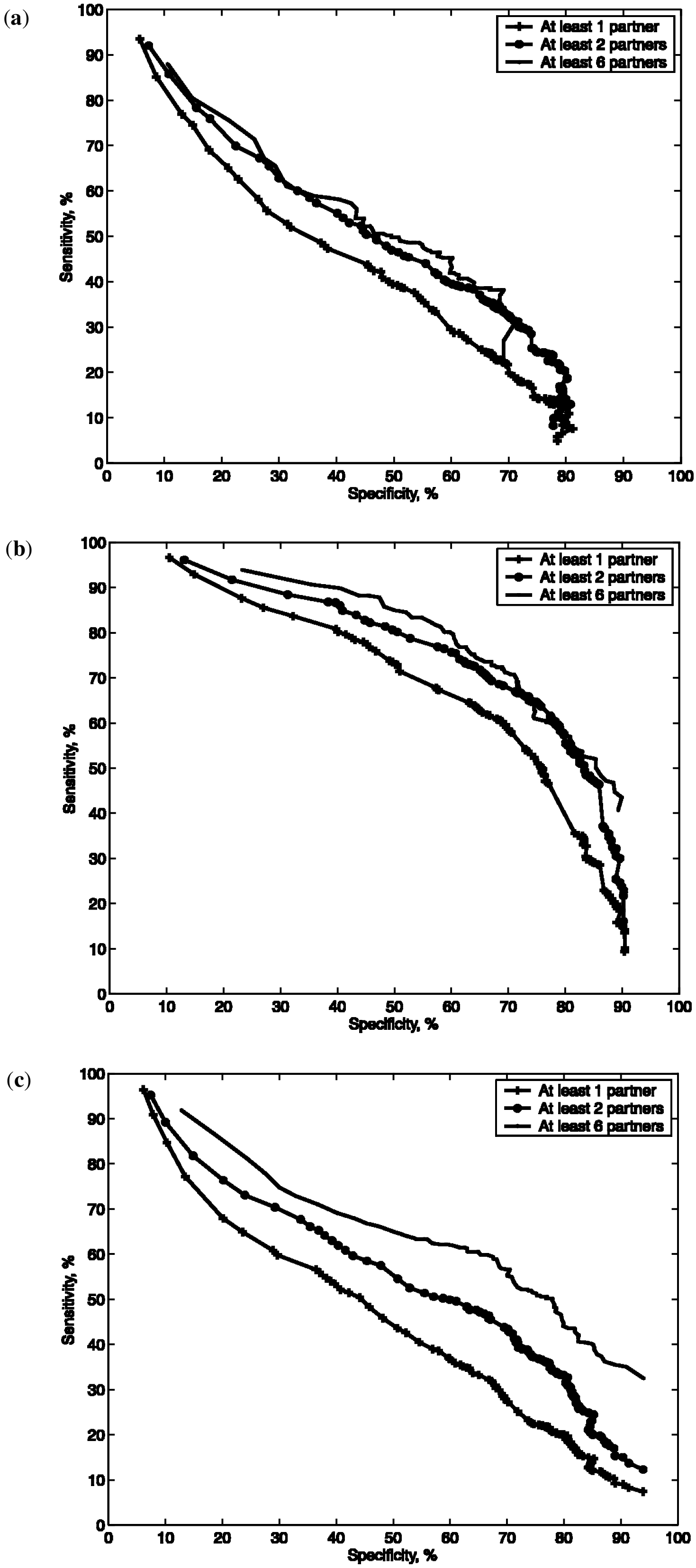

FIG. 3. The relationship between specificity and sensitivity for proteins with at least one, two, and six interaction partners using the Bayesian method based on (**a**) biochemical function, (**b**) subcellular location, and (**c**) cellular role. The corresponding numbers of proteins with one to six interaction partners are given in Table 1.

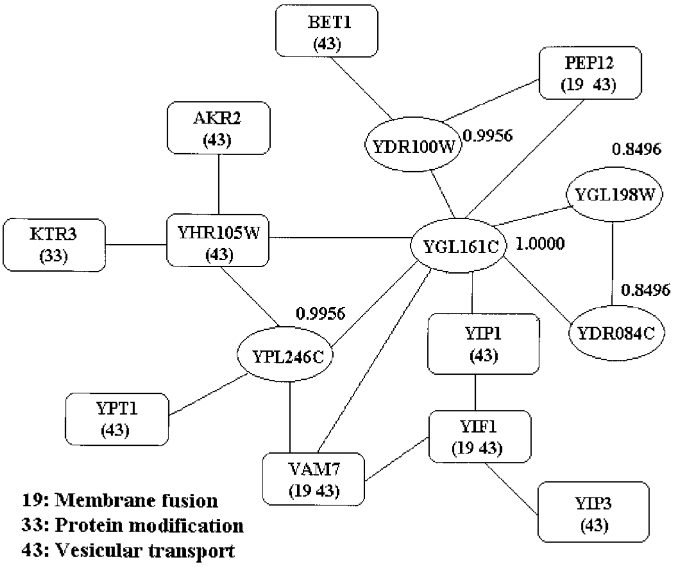

FIG. 4. An example of a protein–protein interaction subnetwork. Proteins in rectangles are annotated, and the numbers in parentheses are the functional categories of the proteins. The proteins in circles are unannotated, and the values beside the circles are the posterior probabilities that the unannotated proteins belong to functional category 43, vesicular transport.

4. DISCUSSION

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 that 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 develop here is that it is a global approach that takes all of 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 chi-square method and the neighboring 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 of 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 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. For a conservative estimate, if we assume that each protein interacts on average with five other proteins, we would expect about $6,000 \times 5/2 = 15,000$ interactions, much higher than the 2,439 interactions in MIPS. With the advance of other high-throughput technologies for detecting protein–protein interactions, our understanding of the protein interaction network will become more complete.

Our method treats each function independently and separately. Generally, the 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. If a protein has function A, that may increase the chance of its having function B because functions A and B are highly correlated—for example, cellular role “RNA processing/modification" and cellular role “RNA splicing." Incorporating that 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 this information. In reality, we know that these annotated proteins may have other functions that have not been determined. As biologists continue to experimentally determine the functions of proteins, the functional annotations will be more and more complete.

Despite the limitations, we show that the results from our approach are reasonably good. The probabilities of protein functions in Fig. 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 neighboring counting method and the chi-square method.

REFERENCES

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. *Nucl. Acids Res.* 25, 3389–3402.

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 microarray gene expression data by using support vector machines *Proc. Natl. Acad. Sci. USA* 97, 262–267.

Clare, A., and King, R.D. 2002. Machine learning of functional class from phenotype data. *Bioinformatics* 18, 160–166.

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. 2001. YPD^{TM}, PombePD^{TM}, and WormPD^{TM}: Model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucl. Acids Res.* 29, 75–79.

Deane, C.M., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions: Two methods for assessment of the reliability of high-throughput observation. *Molecular and Cellular Proteomics* 1, 349–356.

Deng, M.H., Sun, F.Z., and Chen, T. 2003. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pacific Symposium of Biocomputing (PSB2003)*, 140–151.

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.

Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.W., and Hani J. 2000. Integrative analysis of protein ineraction data. *Proc. 8th Int. Conf. on Intelligent System for Molecular Biology (ISMB 2000)*, 152–161.

Gavin, A., Böche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C., *et al.* 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147.

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.

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.

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.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and 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.

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.

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.

Li, S.Z. 1995. *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, Tokyo.

Liu, J.S. 2001. *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.

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.

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.

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. *Nucl. Acids Res.* 30, 31–34.

Mrowka, M., Patzak, A., and Herzel, H. 2001. Is there a bias in proteome research? *Genome Res.* 11, 1971–1973.

Pavlidis, P., and Weston, J. 2001. Gene functional classification from heterogeneous data. *Proc. 5th Int. Conf. on Computational Molecular Biology (RECOMB2001)*, 249–255.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.

Schwikowski, B., Uetz, P., and Fields, S. 2000. A network of protein–protein interactions in yeast. *Nature Biotechnol.* 18, 1257–1261.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., and Pochart, P., *et al.* 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627.

Venables, W.N., and Ripley, B.D. 1996. *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.

Address correspondence to:
Ting Chen
Department of Biological Sciences
University of Southern California
1042 West 36th Place
Los Angeles, CA 90089-1113

E-mail: tingchen@hto.usc.edu