Mapping gene ontology to proteins based on protein–protein interaction data

Minghua Deng, Zhidong Tu, Fengzhu Sun and Ting Chen**

Department of Biological Sciences, Molecular and Computational Biology Program, University of Southern California, 1042 West 36th Place, Los Angeles, CA 90089-1113, USA

ABSTRACT

Motivation: Gene Ontology (GO) consortium provides structural description of protein function that is used as a common language for gene annotation in many organisms. Large-scale techniques have generated many valuable protein–protein interaction datasets that are useful for the study of protein function. Combining both GO and protein–protein interaction data allows the prediction of function for unknown proteins.

Result: We apply a Markov random field method to the prediction of yeast protein function based on multiple protein–protein interaction datasets. We assign function to unknown proteins with a probability representing the confidence of this prediction. The functions are based on three general categories of cellular component, molecular function and biological process defined in GO. The yeast proteins are defined in the *Saccharomyces* Genome Database (SGD). The protein–protein interaction datasets are obtained from the Munich Information Center for Protein Sequences (MIPS), including physical interactions and genetic interactions. The efficiency of our prediction is measured by applying the leave-one-out validation procedure to a functional path matching scheme, which compares the prediction with the GO description of a protein's function from the abstract level to the detailed level along the GO structure. For biological process, the leave-one-out validation procedure shows 52% precision and recall of our method, much better than that of the simple guilty-by-association methods.

Supplementary material: http://www.cmb.usc.edu/~msms/gomapping

Contact: fsun@hto.usc.edu

1 INTRODUCTION

With the completion of genome sequencing of several model organisms, the functional annotation of the proteins is of most importance. Up to December 4, 2002, the *Saccharomyces* Genome Database (SGD, http://genome-www.stanford.edu/Saccharomyces/) (Dwight *et al.*, 2002)

lists 6264 open reading frames (ORF) with three Gene Ontology (GO, http://www.geneontology.org/) categories of cellular component, molecular function and biological process, while at least one-third of the proteins are unknown for each category. A major challenge is to assign function to those proteins whose biological function is yet to be understood.

Several methods have been developed to assign function to an unknown protein. The classical way is to find homologous proteins in databases such as SWISSPROT, using programs such as FASTA (Pearson and Lipman, 1988) and PSI-BLAST (Altschul *et al.*, 1997), and then to predict function of this unknown protein based on function of the homologous proteins. However, not all unknown proteins have homologous proteins in databases. Accordingly, several non-homology-based methods have recently been introduced to assign putative functions to unknown proteins, e.g. the chromosomal proximity method (Overbeek *et al.*, 1999), the Rosetta stone method (Marcotte *et al.*, 1999a; Enright *et al.*, 1999), the phylogenetic method (Pellegrini *et al.*, 1999) and the combined method (Marcotte *et al.*, 1999b; Zheng *et al.*, 2002; Pavlidis and Weston, 2001).

The development of high-throughput bio-techniques and their applications in many areas of biology have generated a large amount of data that are useful for the study of protein function. Several attempts have been made to predict protein function using such data as gene expressions, mutant phenotype and protein–protein interactions. Clustering analysis of gene expression data can be used to predict function of unknown proteins based on the idea that genes with similar function are likely to be co-expressed (Brown *et al.*, 2000; Eisen *et al.*, 1998). Several methods have been developed to predict protein function based on simple guilty-by-association rules, such as the neighbor-counting method (Fellenberg *et al.*, 2000; Schwikowski *et al.*, 2000) and the Chi-square method (Hishigaki *et al.*, 2001). We have developed a Markov random field (MRF) model (Deng *et al.*, 2002) to combine a physical interaction network and the protein function defined in Yeast Proteome Database (YPD, http://www.incyte.com/proteome/) (Costanzo *et al.*, 2001) for function prediction. Recently, Vazquez *et al.* (2003) proposed

---

*To whom correspondence should be addressed.

a combinatorial method to assign protein functions based on physical interaction network by minimizing the number of protein interactions among different functional categories. Our approach is significantly different from that of Vazquez *et al.* (2003) in two important aspects. (1) Vazquez *et al.* (2003) used only the interaction network and did not consider the fraction of proteins having the function of interest in the known proteins. We considered the fraction of proteins having the function of interest in our model. (2) They gave an equal weight to intra-function class interactions. In reality, different intra-function class interactions do not contribute the same to protein function prediction. We estimated the contributions of inter- and intra-function classes. Letovsky and Kasif (2003) proposed a model to assign functions to proteins based on a probabilistic analysis of graph neighborhoods in a protein–protein interaction network, which is fundamentally a MRF model, and the belief propagation algorithm was used to assign function probabilities for proteins in the network. Our approach differs from Letovsky and Kasif (2003) in that we apply the MRF method to the function categories in GO and multiple protein–protein interaction networks (Deng *et al.*, 2003b). The efficiency of our prediction is measured by applying the leave-one-out validation procedure to a functional path matching scheme, which compares the prediction with the GO description of a protein's function from the abstract level to the detail level along the GO structure. For biological process, the leave-one-out validation procedure shows 52% precision and recall of our method, much better than that of the simple guilty-by-association methods.

2 METHODS

Suppose a genome has N proteins $P_1, \ldots, P_N$ and M functional categories $F_1, \ldots, F_M$. Some proteins are known and others are unknown. Let $P_1, \ldots, P_n$ be the unknown proteins and $P_{n+1}, \ldots, P_{n+m}$ be the known proteins, $N = n + m$. Through some biological experiments, we are given several protein–protein interaction networks. Our objective is to assign function to all the unknown proteins based on function of the known proteins and the interaction networks.

Our MRF model based function prediction method is detailed elsewhere (Deng *et al.*, 2002, 2003b). Here we just describe it briefly. Let a function of interest be category 1 and the rest be category 0. All the known proteins can be classified into one of the two categories according to their function. Thus, an interaction between two known proteins can be classified into one of the three groups: (1, 1), (1, 0) and (0, 0). Given a protein physical interaction network (Net1) and a genetic interaction network (Net2), the belief can be represented by a Gibbs distribution (Li, 1995) for this function by considering the classification of all the proteins,

$$\Pr(X \mid \text{Net1}, \text{Net2}) = \frac{\exp[-U(x; \theta)]}{Z(\theta)}, \qquad (1)$$

where

$$U(x; \theta) = -(\alpha N_1 + \beta_1 N_{1_{11}} + \gamma_1 N_{1_{10}} + \kappa_1 N_{1_{00}}$$
$$+ \beta_2 N_{2_{11}} + \gamma_2 N_{2_{10}} + \kappa_2 N_{2_{00}}).$$

$U(x; \theta)$ represents the potential function of the two networks given a functional configuration of $X = (x_1, \ldots, x_N)$. N_1 is the number of proteins for category 1, and $Nk_{ll'}$ is the number of protein interactions between category l and category l' in the k-th network where $k = 1$ for the physical interaction network and $k = 2$ for the genetic interaction network. $\theta = (\alpha, \beta_1, \gamma_1, \kappa_1, \beta_2, \gamma_2, \kappa_2)$ are parameters. $Z(\theta)$ is a normalized constant, which is calculated by summing over all the configurations,

$$Z(\theta) = \sum_x \exp[-U(x; \theta)].$$

$Z(\theta)$ is called the partition function in the general theory of MRF. Note that parameters κ_1 and κ_2 are redundant and can be set to 1. In Deng *et al.* (2002, 2003b), we presented a Gibbs sampler strategy to estimate θ and the posterior probability that an unknown protein has the function of interest given the interaction networks and the functions of known proteins.

The model of Vazquez *et al.* (2003) is a special case of our model. There is an edge between two proteins if they interact. Their energy function is a special case of $U(x, \theta)$ where $\alpha = 0$, $\beta = \kappa = 1$ and $\gamma = 0$. Thus, they gave the same weight to different within-class interactions. The model of Letovsky and Kasif (2003) is essentially the same as Deng *et al.* (2002) although they formulated the problem differently. Here our model is extended to multiple networks.

3 RESULTS

We apply our method to infer the function of unknown proteins in Yeast. In the following, we use genes and proteins interchangeably. We use the functional annotations from GO Consortium (GO Consortium, 2000, 2001). GO is a set of structured vocabularies organized in a rooted directed acyclic graph (DAG), describing attributes of gene products (proteins or RNA) in three categories of 'cellular component', 'molecular function' and 'biological process'. We study these three categories separately. Due to space limitations, we present the results based on 'biological process' only. The results based on the other two categories are provided as supplementary materials. Generally, a gene is annotated by one or multiple GO nodes along the DAG. The nodes at the higher levels correspond to more abstract functional descriptions for gene products. If a gene is annotated with a GO node, we say that this node as well as its parents covers this specific gene. Thus, the nodes at the higher levels cover more genes. Similar to the definition used in (Zhou *et al.*, 2002), we say that a GO node is an informative node if it covers more than 50 genes, and none of its child nodes covers the same number of

Table 1. Gene ontology informative nodes define 134 functions in functional category 'biological process' (only part of them are listed, see Supplemental material for the full table)

	Genes no.	GO ids	Function description
0	2541	4	biological_process unknown
1	348	7154	cell communication
2	231	9605	response to external stimulus
3	50	7606	chemosensory perception
4	145	9628	response to abiotic stimulus
5	91	9607	response to biotic stimulus
6	137	7165	signal transduction
7	84	7242	intracellular signaling cascade
8	54	7264	small GTPase mediated signal transduction
9	3668	8151	cell growth and/or maintenance
10	98	7114	budding
11	504	7049	cell cycle
12	201	67	DNA replication and chromosome cycle
13	60	7059	chromosome segregation
14	91	6260	DNA replication
15	73	6261	DNA dependent DNA replication
16	269	279	M phase
17	118	87	M phase of mitotic cell cycle
18	116	7067	mitosis
19	57	72	M-phase specific microtubule process
20	210	280	nuclear division
21	107	7126	meiosis
22	265	278	mitotic cell cycle
23	93	84	S phase of mitotic cell cycle
24	105	74	regulation of cell cycle
25	941	16043	cell organization and biogenesis
26	647	7028	cytoplasm organization and biogenesis
27	486	6996	organelle organization and biogenesis
28	234	7010	cytoskeleton organization and biogenesis
29	59	30029	actin filament-based process
30	53	30036	actin cytoskeleton organization and biogenesis
31	88	30012	establishment and/or maintenance of cell polarity (sensu Saccharomyces)
32	86	283	establishment of cell polarity (sensu Saccharomyces)
33	89	7017	microtubule-based process
34	80	226	microtubule cytoskeleton organization and biogenesis
35	93	7005	mitochondrion organization and biogenesis
36	111	7033	vacuole organization and biogenesis
37	59	6623	protein-vacuolar targeting
38	165	42254	ribosome biogenesis and assembly
39	50	42255	ribosome assembly
40	127	7046	ribosome biogenesis
41	100	6364	rRNA processing
42	122	7047	cell wall organization and biogenesis

Nodes with underlines are the terminal informative nodes.

genes as the parent node, and the terminal informative nodes as the informative nodes such that none of their descendants are informative. In this study, we define a functional path for a node as the path from the root to the node. The closer a node is to the root, the more abstract the corresponding function is and the farther away from the root, the more detailed. By definition, if a gene belongs to a functional node, it automatically belongs to all the nodes on its functional paths. It is easy to see that the DAG structure allows multiple functional paths for a given node. In this study, we use the concept of functional paths to define the function of a gene product, so that a gene can be predicted with GO functions at different resolutions.

We downloaded three ontology files from the GO database on December 4, 2002. We also downloaded the SGD gene list with GO annotations from the SGD database. For 'biological process', part of the informative nodes and the terminal informative nodes are given in Table 1. The full table is

available in Supplementary material. It should be noted that the 'Root' node is not used for prediction since it covers all the proteins. The known proteins are those proteins covered by at least one GO node except the 'unknown' node. When we use the informative nodes to define function, some known proteins may not be covered by any of these informative nodes, so we call them as uncovered proteins. The total number of uncovered proteins is quite small. For protein interactions, we downloaded the Munich Information Center for Protein Sequences (MIPS, Mewes *et al.*, 2002, http://mips.gsf.de) physical and genetic interaction data.

For a function of interest, we first test the hypothesis that the number of within-class $(1, 1)$ interaction pairs is higher than expected within a network. For a given network, let K be the number of interaction pairs in which both proteins are known. Let K_{11} be the number of interaction pairs with both proteins having the function of interest. Let p be the fraction of known proteins having the function of interest among all the known proteins of *Saccharomyces cerevisiae*. Under the null hypothesis of no association between protein function and protein interactions, we should expect $E(K_{11}) = Kp^2$. We define the fold number as

$$\text{Fold} = \frac{K_{11}}{Kp^2}.$$

We first calculate the fold number for each informative node based on the physical interaction network and the genetic interaction network separately. The fold numbers are similar for the two networks and, thus, we then combine the two networks to calculate the fold numbers for all the informative nodes. For biological process, all terminal informative nodes have fold number greater than 1, and the average fold number is 35.7. Therefore, the MRF model should be applicable to most functional classes. The prediction accuracy based on the MRF should increase as the fold number increases.

We apply the MRF method to predict protein function under the three GO categories. For each informative node, the parameters can be estimated by the quasi-likelihood approach (Li, 1995) using the interaction subnetworks consisting of known proteins. The computation is done in S-PLUS (Venables and Ripley, 1996). With these parameters, the Gibbs sampler computes the posterior probability that an unknown protein has the function of interest. We assign this function to an unknown protein if this posterior probability is above a certain threshold.

For comparison, we also implement the neighbor-counting method (Schwikowski *et al.*, 2000) and the Chi-square method (Hishigaki *et al.*, 2001) to predict protein function. For the neighbor-counting method and $1 \leq k \leq 10$, we assign $\min(k, n_i)$ functions with the top frequencies among the neighbors of the i-th protein, where n_i is the number of functions among the neighbors of the i-th protein. For the Chi-square method and $1 \leq k \leq 10$, we assign k functions with the top k Chi-square values for each protein with at least

one interaction partners. To quantify how the effective use of interaction data can increase the prediction accuracy, we also randomly assign functions to a protein according to the fraction of the known proteins having the function among all the known proteins.

The accuracy of the predictions is measured by a leave-one-out method. The method randomly selects a known protein and assumes it as unknown. We predict its functions by the above methods, and then compare the predictions with the original functions of the protein. For example, in Figure 1, PRP12 is annotated with two GO terminal informative nodes 35 and 41, corresponding to four functional paths, $(-1 \rightarrow 9 \rightarrow 25 \rightarrow 26 \rightarrow 27 \rightarrow 35, -1 \rightarrow 9 \rightarrow 25 \rightarrow 26 \rightarrow 38 \rightarrow 40 \rightarrow 41, -1 \rightarrow 9 \rightarrow 57 \rightarrow 89 \rightarrow 102 \rightarrow 106 \rightarrow 107 \rightarrow 41,$ and $-1 \rightarrow 9 \rightarrow 57 \rightarrow 89 \rightarrow 96 \rightarrow 98 \rightarrow 41)$, and predicted with eight informative nodes of 9, 25, 26, 27, 57, 89, 96 and 98, corresponding to two predicted functional paths $(-1 \rightarrow 9 \rightarrow 25 \rightarrow 26 \rightarrow 27,$ and $-1 \rightarrow 9 \rightarrow 57 \rightarrow 89 \rightarrow 96 \rightarrow 98)$.

We use precision and recall to summarize the comparison as in (Owen *et al.*, 2003). The precision is defined as the fraction of matches between the annotated and the predicted functions among the predictions, and the recall is defined as the fraction of matches between the annotated and the predicted functions among the original function annotations as detailed below. We repeat the leave-one-out experiment for all the known proteins with at least one interaction partner. Because of the specialty of the structured GO annotation, different criteria can be used to count the number of matches between the annotated and the predicted functional paths. In the Appendix, we give three approaches for counting matches between the annotated and predicted functional paths and their limitations. In the first approach, we treat the functional attributes as independent without considering their relationship in the GO structure. In the second approach, we treat functional paths as units of analysis and two functional paths match each other if and only if they are exactly the same. The above two approaches represent two extreme cases for counting matches between annotated and predicted functional paths. In the Appendix, we define new precision and recall measures considering the relationship between functional paths. The following results are based on the new precision and recall measures. However, the correlation between prediction accuracy and fold number is not very strong. This maybe due either to errors in the interaction networks or in the function annotations of known proteins.

Figure 2 shows the relationship between precision and recall of our approach using different thresholds for posterior probabilities for 'biological process'. With the threshold equals to 0.19, the corresponding precision and recall are roughly the same and equal to 52% which is defined as the prediction accuracy.

Figure 3 shows the relationship between precision and recall for the four different methods discussed above: the
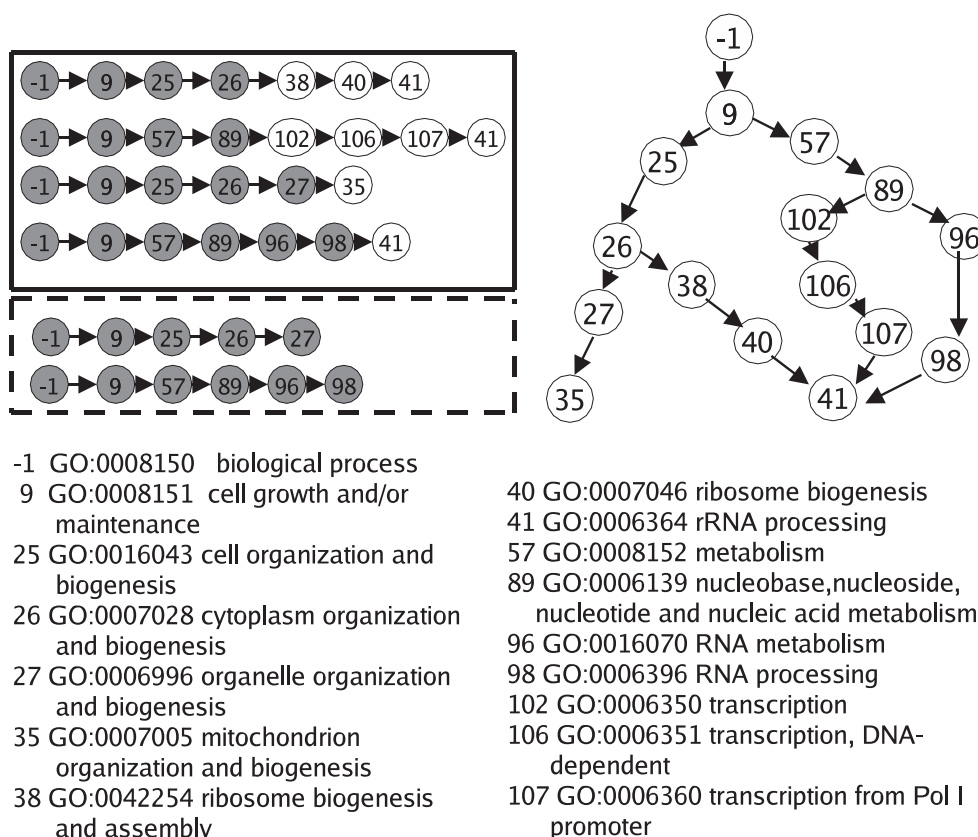
-1 GO:0008150 biological process
9 GO:0008151 cell growth and/or
maintenance
25 GO:0016043 cell organization and
biogenesis
26 GO:0007028 cytoplasm organization
and biogenesis
27 GO:0006996 organelle organization
and biogenesis
35 GO:0007005 mitochondrion
organization and biogenesis
38 GO:0042254 ribosome biogenesis
and assembly

40 GO:0007046 ribosome biogenesis
41 GO:0006364 rRNA processing
57 GO:0008152 metabolism
89 GO:0006139 nucleobase,nucleoside,
nucleotide and nucleic acid metabolism
96 GO:0016070 RNA metabolism
98 GO:0006396 RNA processing
102 GO:0006350 transcription
106 GO:0006351 transcription, DNA-
dependent
107 GO:0006360 transcription from Pol I
promoter

Fig. 1. The annotation and the prediction for protein PRP12. The paths in the solid rectangle are the original GO functional annotation, while the paths in the dash rectangle are the predicted functional paths. If we count the matches by the individual function, the precision and recall are overestimated as 100.0% and 53.3%, respectively. If we count the exact matches by the functional paths, both of the precision and recall are underestimated as 0.0% since no exact matched path exists. If we use the path match by the levels, the precision and recall are estimate as 25.0% and 13.0%, respectively.
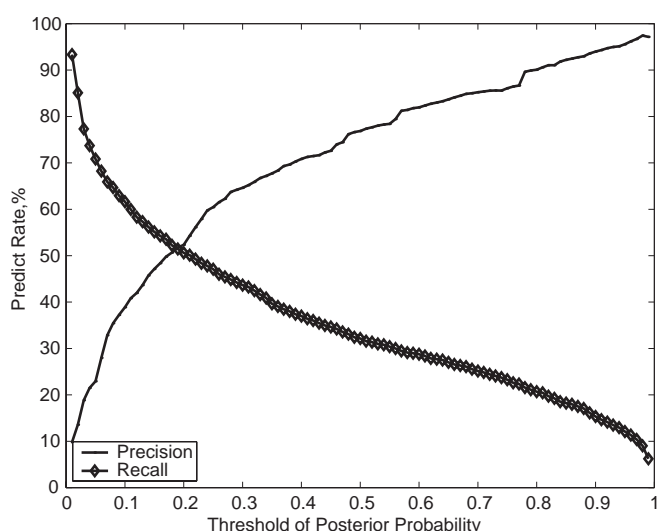

Fig. 2. Precision and recall of the predictions for different posterior probability thresholds based on 'biological process' and proteins with at least one interaction partner.

MRF method, the Chi-square method, the neighbor-counting method and random function assignment. The figure indicates that for any given precision, the recall of the MRF method is higher than that of the neighbor-counting method, the Chi-square method and random function assignment for 'biological process'. It is interesting to see that the performance of the Chi-square method is even worse than random assignment. The reason for this might be due to the relatively small number of interaction partners a protein has and a Chi-square statistic is not appropriate in this situation.

Figure 4 shows the relationship between the fold number [the ratio of the observed number of (1, 1) interaction pairs over the expected] and the prediction accuracy for the terminal informative nodes. It shows a trend that the prediction accuracy increases as the fold number increases.

4 DISCUSSION

We apply the MRF method for function prediction of unknown proteins based on multiple protein–protein interaction networks and the functional annotations of known
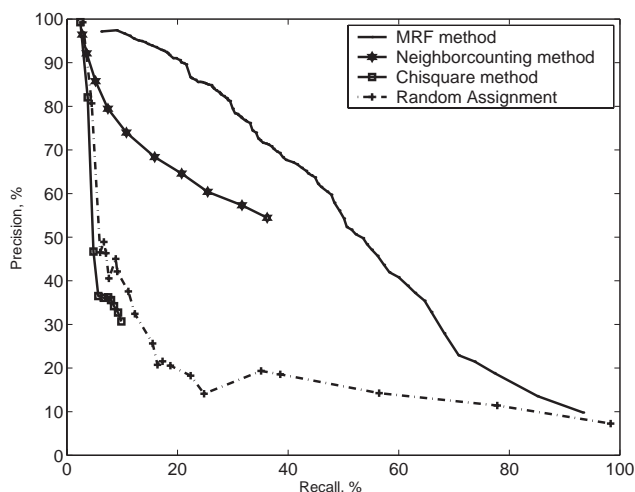
Fig. 3. Precision and recall of predictions for the four different methods based on 'biological process' and proteins with at least one interaction partner.
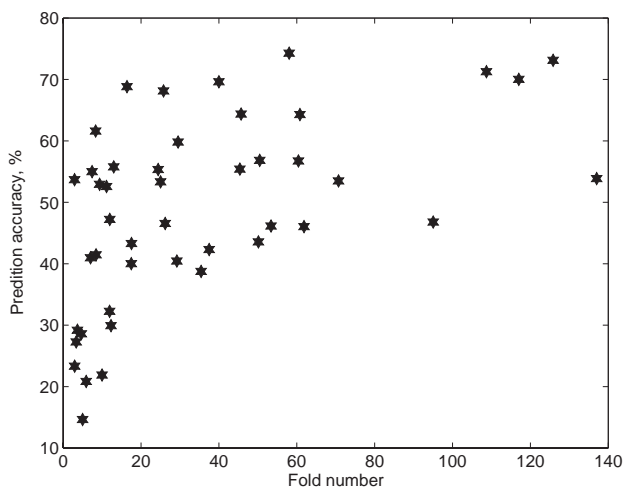

Fig. 4. Relationship between fold number and the prediction accuracy for the terminal informative nodes based on 'biological process'.

proteins in GO. 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 develop here is that it is a global approach taking all the interaction networks and the functions of known proteins into consideration.

Two types of protein networks are considered here: the physical interaction network and the genetic interaction network. Other kinds of networks or networks generated by other techniques can be easily incorporated into our Bayesian framework (Deng *et al.*, 2003b).

We define new precision and recall measures considering the relationship between different functional paths and assess the accuracy of the predictions by comparing functional paths. The prediction results based on MRF outperform the results of the Chi-square method and the neighbor-counting method. We show that, for a given precision, the recall of our method is higher than that of the other two methods. We also notice that different functional classes can be predicted with different accuracy. The fold number can be used as a preliminary indicator for the prediction accuracy.

There are several limitations of our approach. We do not consider any false positives in the protein–protein interaction network in our model. Thus, only the MIPS interaction networks, which is believed to be real interactions, are used in our analysis. The actual number of interacting protein pairs might be much higher than what have obtained in MIPS. As more and more data being generated, our model will perform better. Several protein–protein interaction data such as, Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/) (Xenarios *et al.*, 2002), Biomolecular Interaction Network Database (BIND, http://www.bind.ca/) (Bader *et al.*, 2003), The General Repository for Interaction Datasets (GRID, http://biodata.mshri.on.ca/grid) (Breitkreutz *et al.*, 2003) are available with different reliability (Deane *et al.*, 2002; Deng *et al.*, 2003a). It is important to develop methods to estimate the reliability of each possible interaction and incorporate these reliability into our model. It is a topic for future research.

ACKNOWLEDGEMENTS

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. *Nucleic Acids Res.*, **25**, 3389–3402.

Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.

Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol.*, **4**, R23.

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.

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. and Kondu,P. *et al.* (2001) YPD™, PombePD™, and WormPD™: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,I. (2002) Protein interactions: two methods for assessment of the reliability of high-throughput observation. *Mol. cell. proteomics*, **1**, 349–356.

Deng,M., Zhang,K., Mehta,S., Chen,T. and Sun,F. (2002) Prediction of protein function using protein–protein interaction data. *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002)*, 197–206.

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

Deng,M., Chen,T. and Sun,F. (2003b) An integrated probabilistic model for functional prediction of proteins. *Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB2003)*, 95–103.

Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M. Sherlock,G. *et al.* (2002). *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

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.

Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.W. and Hani,J. (2000) Integrative analysis of protein ineraction data. *Proceedings of the Eighth International Conference on Intelligent System for Molecular Biology (ISMB2000)*, 152–161.

The Gene Ontology (GO) Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

The Gene Ontology (GO) Consortium (2001) Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.

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.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19** (Suppl.1), 197–204.

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. (1999a) 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. (1999b) 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. *Nucleic Acids Res.*, **30**, 31–34.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Owen,A.B., Stuart,J., Mach,K., Villeneuve,A.M. and Kim,S. (2003) A gene recommender algorithm to identify coexpressed genes in *C. elegans Genome Res.*, **13**, 1828–1837.

Pavlidis,P. and Weston,J. (2001) Gene functional classification from heterogeneous data. *Proceedings of the Fifth International Conference 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.

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.

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

Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks *Nat. Biotechnol.*, **21**, 697–700.

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

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP: the database of interacting proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Zheng,Y., Roberts,R.J. and Kasif,S. (2003) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.*, **3**, 0060.1-0060.9.

Zhou,X., Kao,M. and Wong,W. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.

APPENDIX

In this appendix, we present three approaches for counting overlaps between annotated and predicted functional paths and give three ways for calculating precision and recall. The first naive approach is to compare the set of predicted functions $\mathcal{P}$ with the set of annotated functions $\mathcal{A}$. For the example given in Figure 1, $\mathcal{P} = \{9, 25, 26, 27, 57, 89, 96, 98\}$ and $\mathcal{A} = \{9, 25, 26, 27, 35, 38, 40, 41, 57, 89, 96, 98, 102, 106, 107\}$. The overlap between $\mathcal{P}$ and $\mathcal{A}$ is $\mathcal{P} \cap \mathcal{A} = \{9, 25, 26, 27, 57, 89, 96, 98\}$. The precision (Prec) and recall (Rec) for this protein can be defined as

$$\text{Prec} = |\mathcal{P} \cap \mathcal{A}|/|P| = 8/8 = 100.0\%$$
$$\text{Rec} = |\mathcal{P} \cap \mathcal{A}|/|A| = 8/15 = 53.3\%.$$

The overall precision and recall can be defined as the mean of the corresponding quantities over all the proteins in the test set. The problem with this criterion is that it does not take the level of predictions into consideration. High level functional attributes are generally much easier to predict than low level

detailed attributes. This criterion also ignores the relationship among functional attributes.

The second approach is to consider exact matches among the predicted functional paths with the annotated functional paths. For the example in Figure 1 of Section 3, the annotated functional paths are given in the solid box and the predicted functional paths are given in the dashed box in Figure 1. The overlap between the predicted and annotated functional paths can be found. For the above example, no exact matched path exists, so Rec = Prec = 0.0%. However, if we trace a functional path along the GO DAG, some of the predictions are in fact related to the original annotation even if they are not exactly the same. For example, annotated functional path $-1 \rightarrow 9 \rightarrow 25 \rightarrow 26 \rightarrow 38 \rightarrow 40 \rightarrow 41$ and predicted functional path $-1 \rightarrow 9 \rightarrow 25 \rightarrow 26 \rightarrow 27$ have the same top three level attributes and differ starting from the fourth level. Ignoring the overlapping information among functional paths may not be reasonable.

Here, we define new precision and recall measures that take into consideration overlaps between functional paths. For a known protein, suppose the original annotated functional paths are $\{O_1, O_2, \ldots, O_n\}$ with $O_i = \{O_{i0} \rightarrow O_{i1} \rightarrow \cdots \rightarrow O_{i,k_i}\}$, where O_{il} is the l-th level attribute from the root of the i-th annotated functional path. Similarly, let the predicted functional paths be $\{P_1, P_2, \ldots, P_m\}$ with $P_j = \{P_{j0} \rightarrow P_{j1} \rightarrow \cdots \rightarrow P_{j,k'_j}\}$, where P_{jl} is the l-th level attribute from the root of the j-th predicted functional path.

For each annotated functional path O_i and predicted functional path P_j, we first define an overlap function by

$$\text{overlap}(O_i, P_j) = \max\{L; O_{il} = P_{jl} \text{ for any } 1 \leq l \leq L\},$$

where we define the maximum of an empty set to be $-\infty$. We then define the predicted depth (pre-depth) for functional path O_i as the maximum of the overlap between O_i with all the predicted functional paths, i.e.

$$\text{pre-depth}(O_i) = \max_{j=1}^{m} \text{overlap}(O_i, P_j).$$

Similarly, we define the annotated depth (annot-depth) of P_j as

$$\text{annot-depth}(P_j) = \max_{i=1}^{n} \text{overlap}(O_i, P_j).$$

The general idea of our new recall measure is to give a score $4^{-(k_i - \text{pre-depth}(O_i))}$ for the original functional path O_i. Thus, if all the nodes for O_i are predicted correctly, we give a score 1, and if the very top node is not correctly predicted, we give a score 0. The score increases as the the number of nodes correctly predicted increases. The recall (Rec) can then be defined as

$$\text{Rec} = \frac{\sum_{i=1}^{n} 4^{-[k_i - \text{pre-depth}(O_i)]}}{n}.$$

Similarly, we can define the precision as follows. For predicted functional path P_j, we give it a score $4^{-[k'_j - \text{annot-depth}(P_j)]}$. The precision can be defined as

$$\text{Prec} = \frac{\sum_{j=1}^{m} 4^{-[k'_j - \text{annot-depth}(P_j)]}}{m}.$$

For the example in Section 3, $\text{Rec} = \left[\frac{1}{4} + \left(\frac{1}{4}\right)^3 + \frac{1}{4} + \left(\frac{1}{4}\right)^4\right]/4 = 13.0\%$ and $\text{Prec} = (\frac{1}{4} + \frac{1}{4})/2 = 25.0\%$.