Inferring protein interactions from experimental data by association probabilistic method

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MARCH 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.20783 · Source: PubMed

4 AUTHORS:

Luonan Chen
Chinese Academy of Sciences
265 PUBLICATIONS **4,642** CITATIONS
SEE PROFILE

Ling-Yun Wu
Chinese Academy of Sciences
68 PUBLICATIONS **1,161** CITATIONS
SEE PROFILE

Yong Wang
Chinese Academy of Sciences
93 PUBLICATIONS **1,322** CITATIONS
SEE PROFILE

Xiang Zhang
Fourth Military Medical University
759 PUBLICATIONS **22,279** CITATIONS
SEE PROFILE

SHORT COMMUNICATION

Inferring Protein Interactions from Experimental Data by Association Probabilistic Method

Luonan Chen,[1]* Ling-Yun Wu,[2] Yong Wang,[2] and Xiang-Sun Zhang[2]
[1]*Osaka Sangyo University, Osaka, Japan*
[2]*Academy of Mathematics and Systems Science, CAS, Beijing, China*

ABSTRACT **To elucidate protein interaction networks is one of the major goals of functional genomics for whole organisms. So far, various computational methods have been proposed for inference of protein–protein interactions. Based on the association method by Sprinzak et al., we propose an association probabilistic method in this short communication to infer protein interactions directly from the experimental data, which outperformed other existing methods in terms of both accuracy and efficiency despite its simple form. Specifically, we show that the association probabilistic method achieves the highest accuracy among the existing approaches for the measures of root-mean-square error and the Pearson correlation coefficient, and also runs much faster than the LP-based method, by experimental dataset in Yeast. Software is available from the authors upon request. Proteins 2006;62:833–837.** © 2006 Wiley-Liss, Inc.

Key words: protein interaction; domain interaction; protein network; bioinformatics

INTRODUCTION

One of the major goals of functional genomics is to elucidate protein interaction networks for whole organisms. Determining protein interactions provides not only detailed functional insights on characterized proteins, but also an information base for identifying biological complexes and metabolic or signal transduction pathways. The recent emergence of high-throughput proteomics techniques has opened new prospects to systematically characterize physical interactions between proteins. Based on an experimental dataset, many computational algorithms have been developed to infer the protein–protein or domain–domain interactions. For instance, for inferring protein interactions, there are the gene fusion (Rosetta Stone) method,[1,2] the phylogenetic profile method,[3] the interaction domain pair profile method,[4] the probabilistic method,[5] the SVM-based method,[6] and the LP-based approach,[7] whereas for inferring domain interactions, there are the association method,[8] the EM algorithm[9] and the simple ASNM method.[10] However, from the viewpoint of computational complexity, Hayashida et al.[10] have proven that maximizing correctly classified examples of protein–protein interactions is a MAX SNP-hard problem, which indicates the difficulty in nature to minimize the prediction errors of protein–protein interactions.

Despite the relative success, there is much room for improvement of protein interaction inference in terms of prediction quality and computational efficiency, which are also strongly demanded in the biology community. In this article, we propose a new algorithm, association probabilistic method (APM), to infer protein–protein interaction through simple computation procedure, which outperforms other existing methods despite its simple form and achieves an accurate and efficient prediction of protein–protein interactions by the experiment data in Yeast. In addition, both binary and confident ratio data of protein interactions can be considered in the proposed method, thereby significantly extending the available dataset and improving the reliability of the interaction prediction.

METHODS

Probabilistic Model for Inference

In this section, we describe the association-based methods[8] with a probabilistic model for protein–protein interaction inference, which are the essential basis of our algorithm.

Assume that there are N proteins indicated by $P_1,\ldots,P_N$, and M domains in proteins represented by $D_1,\ldots,D_M$. Let P_i also denote a set of domains in the protein i. One protein P_i may include multiple domains D_j. Define P_{ij} and D_{mn} to represent the protein pair (P_i,P_j) and the domain pair (D_m,D_n), respectively. P_{ij} is also used to represent a set of domain pairs in P_i and P_j, i.e., $\{D_{mn}|D_m \in P_i, D_n \in P_j\} \subset P$, where P is a multi set of all protein pairs P_{ij}.

Let an interaction between P_i and P_j or between D_m and D_n be represented by a random variable p_{ij} or d_{mn}. Then,

Grant sponsor: project "Bioinformatics," Bureau of Basic Science, CAS

*Correspondence to: Luonan Chen, Osaka Sangyo University, Osaka 574-8530, Japan. E-mail: chen@elec.osaka-sandai.ac.jp

Received 27 May 2005; Revised 11 August 2005; Accepted 1 September 2005

Published online 4 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20783

$p_{ij} = 1$ if P_i and P_j interact with each other, otherwise $p_{ij} = 0$. In the same manner, $d_{mn} = 1$ if D_m and D_n interact with each other, otherwise $d_{mn} = 0$. Based on binary interaction data (i.e., whether or not interaction for each protein pair is observed from the experiment),[7] the association method (ASSOC) defines the score or probability of interaction between D_m and D_n as

$$\lambda_{mn} \equiv \Pr(d_{mn} = 1) = \frac{I_{mn}}{N_{mn}} \in [0,1] \quad (1)$$

where N_{mn} is the number of protein pairs containing domain pair (D_m,D_n) in the training dataset, and I_{mn} is the number of interacting protein pairs containing domain pair (D_m,D_n) in the training dataset, i.e., $N_{mn} = \sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} 1$ and $I_{mn} = \sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} \beta_{ij}$where $\{P_{ij}|D_{mn} \in P_{ij}\}$ represents the set of protein pairs containing domain pair (D_m,D_n) in the training dataset, and $\beta_{ij} = 1$ if the interaction between proteins P_{ij} and P_j is observed in the experiments, otherwise $\beta_{ij} = 0$.

However, for general interaction data (i.e., the ratio of the numbers between observed interactions and total experiments), a ratio of interactions between P_i and P_j[10] is defined as

$$\rho_{ij} = \frac{O_{ij}}{Z} \in [0,1] \quad (2)$$

where O_{ij} is the number that the interaction between proteins P_i and P_j is observed in the experiments, and Z is the total number of the experiments. In other words, ρ_{ij} is a confident ratio of the interaction between proteins P_i and P_j, and is considered as the natural extension of the binary β_{ij} from the discrete {0, 1} to the continuous [0, 1]. Then, based on such ratio data as well as binary interaction data,[10] ASNM (association numerical method) extends the association method by defining the score or probability of interaction between D_m and D_n as

$$\lambda_{mn} \equiv \Pr(d_{mn} = 1) = \frac{\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} \rho_{ij}}{N_{mn}} \in [0, 1] \quad (3)$$

Figure 1 illustrates an example for two proteins with four domains. Therefore, the probability of interactions between P_i and P_j is estimated by

$$\Pr(p_{ij} = 1) = 1 - \prod_{\{D_{mn}\in P_{ij}\}} (1 - \lambda_{mn}) \in [0,1] \quad (4)$$

This probabilistic model assumes that domain–domain interactions are independent, and two proteins interact if at least one domain pair from the two proteins interacts.

APM

Clearly, Eqs. (2) or (3) are efficient from the computational viewpoint because of their simple form, but the prediction accuracy is not satisfactorily high.[10] However, the domain is a basic function unit in a protein. In that sense, the probabilistic model assumes that a protein pair interacts provided that there is at least one domain pair that interacts between the two proteins. However, the domain interaction rates for ASSOC of Eq. (2) and ASNM of Eq. (3) are based on the average rates for all domains, which are not consistent with the probabilistic model of Eq. (4), in particular for the ratio data. The key problem to infer protein interaction is to estimate λ_{mn} accurately from the given experimental data ρ_{ij} or β_{ij}. Analogous to the probabilistic model of Eq. (4), we propose a new estimation method for domain interactions, i.e., association probabilistic method (APM), as follows:

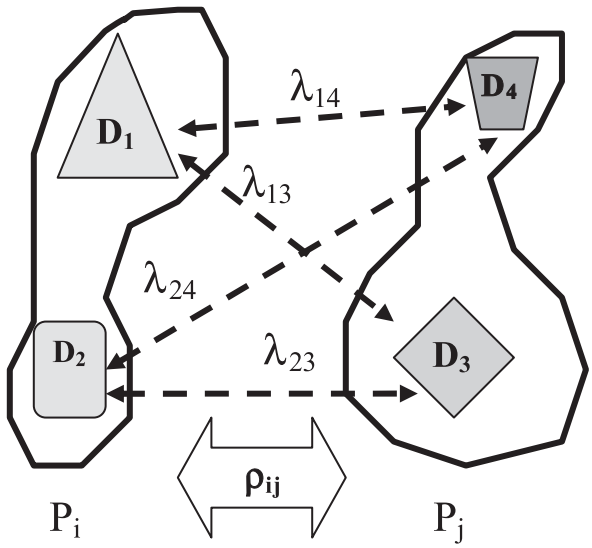

Fig. 1. An illustrative example for two proteins with four domains. Proteins: $P_i = \{D_1, D_2\}$, $P_j = \{D_3, D_4\}$; Domains: D_1, D_2, D_3, D_4; Domain pairs: D_{13}, D_{14}, D_{23}, D_{24}; $P_{ij} = \{D_{13}, D_{14}, D_{23}, D_{24}\}$.

$$\lambda_{mn} \equiv \Pr(d_{mn} = 1) = \frac{\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} [1 - (1 - \rho_{ij})^{1/|P_{ij}|}]}{N_{mn}} \in [0,1] \quad (5)$$

where $|P_{ij}|$ represents the number of domain pairs in P_{ij}. If the ratio ρ_{ij} for each protein pair (P_i,P_j) takes either 0 or 1, Eq. (5) is identical to Eq. (1) or Eq. (3) because of $\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} [1 - (1 - \rho_{ij})^{1/|P_{ij}|}] = \sum_{\{P_{ij}|D_{mn}\in P_{ij}\}} \rho_{ij} = I_{mn}$.λ_{mn} of Eq. (5) can be viewed as a reverse function of $\Pr(p_{ij} = 1)$ of Eq. (4) when all of λ_{mn} in P_{ij} take an identical value. Thus, the protein interaction of APM is obtained by substituting λ_{mn} of Eq. (5) into Eq. (4).

Clearly, both λ_{mn} and $\Pr(p_{ij} = 1)$ are straightforward equal to ρ_{ij} for $|P_{ij}| = 1$ (i.e., there is only one interacting domain pair between proteins P_i and P_j). However, all the domain pairs have the equal opportunity to contribute the interactions between P_i and P_j for $|P_{ij}| > 1$ if there is no prior information under the assumption of independence for domain–domain interactions.

In major datasets of protein interaction, such as DIP database[11] and YIP (Ito's Yeast Interacting Proteins)[12] database, a large number of domain pairs exist only in a single protein pair or a few protein pairs. Such facts make the prediction of domain interactions by Eqs. (1) and (3) far from the consistency with the probabilistic model Eq. (4), and also are the main reason why the accuracy of Eqs. (1) or (3) is poor for the probabilistic model. For instance, if domain pairs D_{13},D_{14},D_{23},D_{24} exist only in protein pair P_{ij}

as shown in Figure 1, and the ratio of interactions between proteins P_i and P_j by experiments is $\rho = 0.1$, then the score of interaction between each domain pair is $\lambda_{13} = \lambda_{14} = \lambda_{23} = \lambda_{24} = 0.1$ according to Eq. (3). Hence, the probability of interactions between P_i and P_j is $\Pr(p_{ij} = 1) = 1 - (1 - 0.1)^4 = 0.3439$ by the probabilistic model Eq. (4), which is clearly different from the experimental ratio $\rho = 0.1$. However, the score of interaction between each domain pair is $\lambda_{13} = \lambda_{14} = \lambda_{23} = \lambda_{24} = \lambda = 1 - (1 - 0.1)^{1/4}$ according to the proposed Eq. (5) because of $|P_{ij}| = 4$, which in turn gives $\Pr(p_{ij} = 1) = 1 - (1 - \lambda)^4 = 0.1$ that is consistent with the experimental ratio $\rho = 0.1$. Similarly, when a domain pair exists in multiple protein pairs, Eq. (5) can approximately make the inference of domain interactions consistent with that of the corresponding protein interactions.

EXPERIMENTAL RESULTS

In this section, we show that APM is much superior to the existing methods in terms of accuracy and efficiency of protein interaction prediction by experimental data.

Data and Manipulation

In this study, we compared the proposed APM with LP-based method (LPNM),[7] association based methods (ASNM,[10] ASSOC[8]), and EM[9] method. Among those existing methods, the LPBN, ASSOC, and EM are developed for the binary data whereas LPNM and ASNM can be applied directly on experiment ratio data. To make the comparison in a unified framework for both binary and ratio data, the full data of Ito's Yeast Interacting Proteins (YIP) database[12] are adopted as a first example. Although the database provides the numerical interaction (ratio) data for protein pairs based on the number of IST (Interaction Sequence Tags) hits, the binary interaction dataset can be extracted from the numerical interaction dataset by a defined threshold according to the confidence degree of experiment. With such a comparison framework, all of the methods are applied to the database to generate two sequences, i.e., the sequence of observed probability of interaction and the sequence of predicted probability of interaction for protein pairs. Thus, we evaluate each method by fivefold cross validation and assess the prediction accuracy by three measures: root-mean-square error (RMSE), sensitivity and specificity, and correlation coefficient.

RMSE Measure

The quality of prediction is evaluated by RMSE between the predicted probability $\Pr(p_{ij} = 1)$ of Eq. (4) and the observed value ρ_{ij} or β_{ij}

$$\mathrm{RMSE} = \sqrt{\sum_{\{P_{ij} \in P\}} (\Pr(p_{ij} = 1) - \rho_{ij})^2 / |P|} \qquad (6)$$

where $|P|$ means total number of protein pairs in P or whole dataset. The performance of each method of RMSE and elapsed training time for five numerical interaction datasets are summarized in Table I. According to Table I, the proposed APM not only has highest accuracy or minimal errors for both training (Train) and prediction (Test) of protein interactions, but also is very efficient with the same CPU level as ASSOC and ASNM. In terms of accuracy, t test of one-tail probability for the five tests (total 347 samples) is also performed between APM and each existing method for RMSE criterion in Table I, and indicates that APM is significantly better than others except LPNM from the statistical viewpoint. In addition, we also compared RMSE measure (fivefold cross-validation with IST number = 3) for Uetz's two-hybrid yeast screen dataset (THY),[13] in the same manner as Table I, as indicated in Table II, which also shows similar results as those of the YIP database.

TABLE I. Comparisons for RMSE and Training Time for Ito's Yeast Interaction Datasets (YIP)

	LPNM	EM	ASSOC	ASNM	**APM**
Train (RMSE)					
1st	0.0139	0.4872	0.4625	0.0411	0.0125
2nd	0.0132	0.4856	0.4624	0.0375	0.0116
3rd	0.0141	0.4718	0.4425	0.0395	0.0116
4th	0.0127	0.4751	0.4471	0.0382	0.0104
5th	0.0139	0.4932	0.4672	0.0430	0.0122
Average	0.0136	0.4826	0.4564	0.0399	0.0117
Time (s)	1.5516	1.6586	0.0104	0.0096	0.0118
Test (RMSE)					
1st	0.0368	0.6862	0.6592	0.0633	0.0376
2nd	0.0465	0.6171	0.5792	0.0612	0.0445
3rd	0.0502	0.6404	0.5914	0.0767	0.0491
4th	0.0505	0.6289	0.5846	0.0708	0.0487
5th	0.0362	0.5882	0.5563	0.0524	0.0365
Average	0.0441	0.6322	0.5942	0.0649	0.0434

t test	LPNM/APM	EM/APM	ASSOC/APM	ASNM/APM
t score	0.0838	8.7708	7.4353	2.9942
probability	0.4665	<0.0001	<0.0001	0.0015

TABLE II. Comparisons for Average RMSE and Training Time for Uetz's Two- Hybrid Yeast Screen Dataset (THY)

	LPNM	EM	ASSOC	ASNM	**APM**
Train (RMSE)	0.0135	0.8130	0.7952	0.0840	0.0099
Time (s)	1.3494	0.3022	0.0026	0.0024	0.0046
Test (RMSE)	0.0453	0.8138	0.7616	0.0716	0.0414

Sensitivity and Specificity

Although the RMSE measure has the advantage of being defined for nonbinary variables, the main drawback is that the value of the quadratic distance poorly reflects the proportion of positive prediction to a given class.[14] To compare sensitivity and specificity for both binary and ratio data, a threshold is adopted to process the input data and the result. Specifically, EM and ASSOC need to apply a threshold to numerical data for binary treatment before the estimation of domain–domain pair interaction whereas LPNM, ASNM, and APM apply the threshold after the

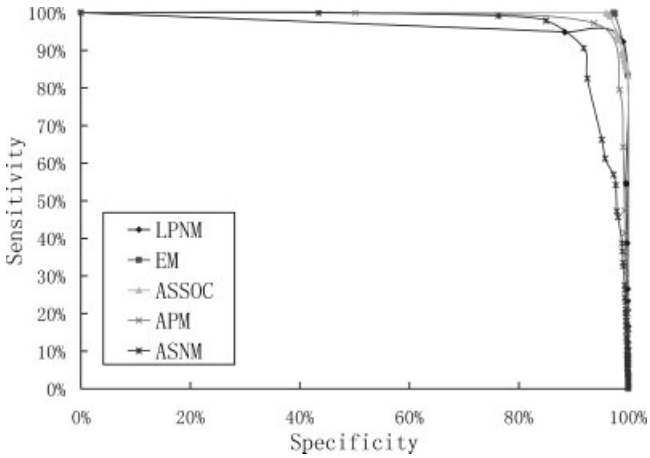

Fig. 2. Comparison of specificity and sensitivity on training data of YIP.

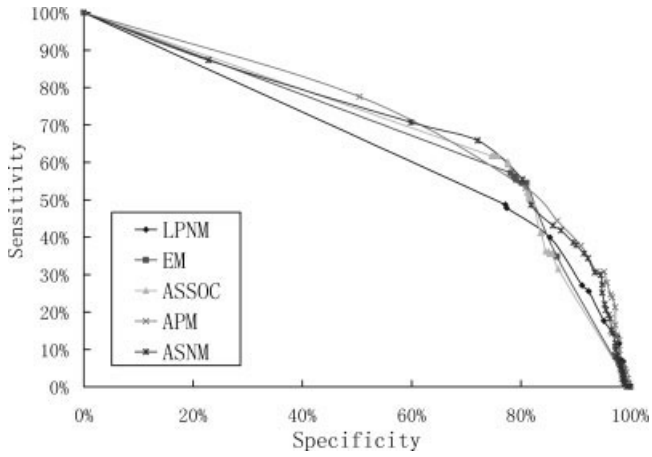

Fig. 3. Comparison of specificity and sensitivity on test data of YIP.

computation only for assessment. Because the ratio of the number of IST hits to the number of experiments is given for each pair of proteins in the numerical interaction data, it is natural to set the IST number as the threshold.

With the binary representation of the observed and predicted interaction of protein pairs, the following numbers are counted:

- TP = the protein pair is both predicted and observed (true positive)
- FP = the protein pair is predicted but not observed (false positive)
- TN = the protein pair is neither predicted nor observed (true negative)
- FN = the protein pair is not predicted but observed (false negative)

The sensitivity is defined as $TP/(TP + FN)$ and specificity is defined as $TN/(TN + FP)$. The comparison of specificity and sensitivity on training data and test data is given in Figures 2 and 3, respectively. The IST hit number is fixed to 3 for observed interaction. From the figures, it is clear that APM has a very good balance between the sensitivity and specificity with a high accuracy both in training and test datasets. It should be noted that the sensitivity in test data is not as high as in the work of Hayashida et al.[7] because we did not remove the protein pairs in the test dataset which do not have domain pairs appearing in the positive training dataset. Although the sensitivity increases significantly by removing the pairs whose scores are always 0, the reliability of the protein–protein interaction is also affected. This fact implies that the number and reliability of the dataset are crucial to the prediction accuracy.

As indicated by Deng et al.,[15] the reliability of the Ito dataset depends on the IST hit number. Thus, we analyzed the change of sensitivity and specificity of test data based on the reliability dataset in Figure 4 by varying the IST hits from 1 to 8. The top and bottom subfigures of Figure 4 correspond to the IST hit numbers 1 and 8, respectively.

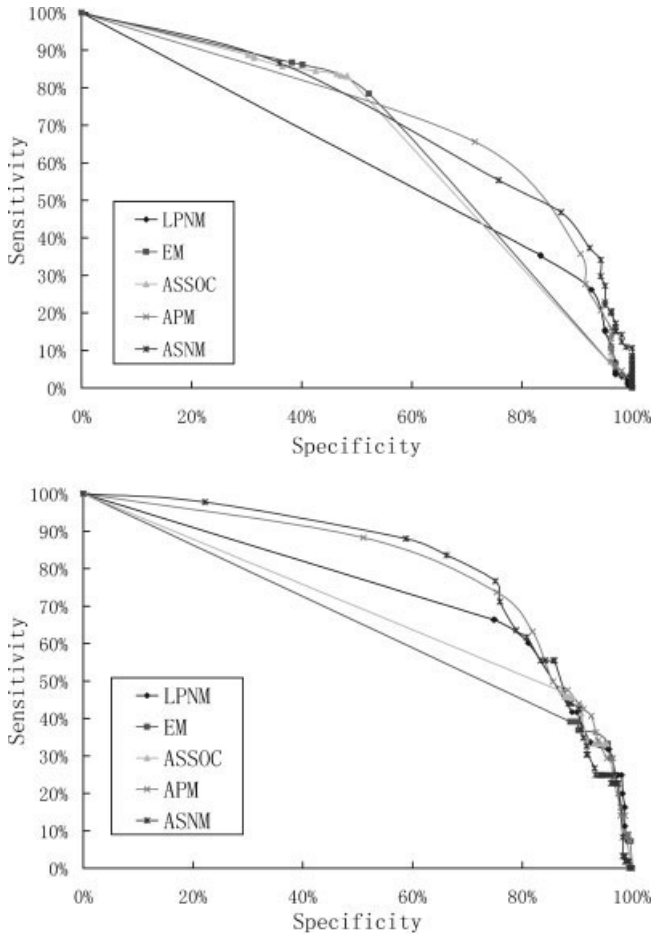

Fig. 4. Comparison of specificity and sensitivity on test data of YIP for different reliability.

All the figures indicate that APM has the best predictive accuracy. In particular, when the reliability of the dataset changes, the performances of LPNM, ASSOC, and EM vary greatly in contrast to APM and ASNM that maintain stable predictive accuracy. In other words, APM is fairly robust to withstand the inconsistency of data. Such feature

TABLE III. Performance on Correlation Coefficient of YIP

	LPNM	EM	ASSOC	ASNM	**APM**
Train					
1st	0.8807	0.5343	0.5223	0.6673	0.9052
2nd	0.8532	0.5531	0.5476	0.6643	0.8863
3rd	0.8455	0.5594	0.5552	0.6688	0.9019
4th	0.8595	0.5304	0.5251	0.6808	0.9013
5th	0.8574	0.5510	0.5518	0.6874	0.8983
Average	0.8593	0.5456	0.5404	0.6737	0.8986
Test					
1st	0.3092	0.3974	0.4017	0.2560	0.4021
2nd	0.4705	0.3086	0.2240	0.2523	0.6463
3rd	0.1946	0.2310	0.2533	0.1638	0.3182
4th	0.3005	0.3428	0.2855	0.1881	0.2410
5th	0.0477	0.1149	0.1162	0.1049	0.0686
Average	0.2645	0.2789	0.2561	0.1931	0.3353

is also very important for the protein–protein interaction data because the IST hit number is generally small compared with the number of experiments, e.g., in the Ito dataset. According to Figures 2–4, some algorithms outperform APM for training data, but APM is generally better than others for test data, which implies that the other algorithms may be over-fitting the data.

Correlation Coefficient

One of the standard measures used by statisticians is the correlation coefficient, also called the Pearson correlation coefficient[14]:

$$C(\boldsymbol{D},\boldsymbol{M}) = \sum_i \frac{(d_i - \bar{d})(m_i - \bar{m})}{\sigma_D \sigma_M} \tag{7}$$

where $\boldsymbol{D}$ and $\boldsymbol{M}$ denote the observed and predicted sequence, respectively. $(\bar{d}, \bar{m})$ are the averages, and (σ_D, σ_M) are the corresponding standard deviations of $\boldsymbol{D}$ and $\boldsymbol{M}$. The correlation coefficient is always between -1 and $+1$. Unlike many measures, the correlation coefficient has a global form rather than being a sum of local terms. Because of information of all four numbers (*TP*, *TN*, *FP*, *FN*), the Pearson correlation coefficient often provides a much more balanced evaluation of prediction.

The performance on correlation coefficient measure is listed in Table III. Again, the proposed APM outperforms other methods in both training and test experiments.

CONCLUSION

We developed a novel inference method (APM) to improve the quality and efficiency of protein–protein interaction prediction for both binary and ratio data by a simple procedure, which not only deals with experimental data directly but also significantly improves the RMSE and correlation coefficient measures of the interaction prediction. In addition to the highest accuracy among the existing approaches, APM outperforms the LP-based method for CPU consumption about 130 times by experimental dataset in Yeast, although computing time is still not a major problem for all of the existing methods with the size of current databases.

It is well known that the accuracy of a predictive algorithm depends on both the dataset and predictive model. The numerical results show that the bottleneck of protein–protein interaction inference problem lies in the dataset, which means that the task of estimating the reliability of different datasets becomes increasingly important.

ACKNOWLEDGMENTS

We are grateful to M. Hayashida, N. Ueda, T. Akutsu, R. Mrowka, and M. Deng for kindly giving us software and data.

AVAILABILITY

Software is available from the authors upon request.

REFERENCES

1. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomics based on gene fusion events. Nature 1999;402:86–90.
2. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.
3. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96: 4285–4288.
4. Wojcik J, Schachter C. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics 2001;17: S296–S305.
5. Gomez GM, Lo SH, Rzhetsky A. Probabilistic prediction of unknown metabolic and signal-transduction networks. Genetics 2001;159:1291–1298.
6. Dohkan S, Koike A, Takagi T. Support vector machines for predicting protein-protein interactions. Genome Informatics 2003; 4:502–503.
7. Hayashida M, Ueda N, Akutsu T. Inferring strengths of protein-protein interactions from experimental data using linear programming. Bioinformatics 2003;19:ii58–ii65.
8. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 2001;311:681–692.
9. Deng M, Mehta S, Sun F, Chen T. Inferring domain-domain interactions from protein-protein interactions. Genome Res 2002; 12:1540–1548.
10. Hayashida M, Ueda N, Akutsu T. A simple method for inferring strengths of protein-protein interaction. Genome Informatics 2004; 15:56–68.
11. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S, Eisenberg D. DIP: the database of interacting proteins. A search tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30:303–305.
12. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98:4569–4574.
13. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000;403:623–627.
14. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16(5):412–424.
15. Deng M, Sun F, Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. Pac Symp Biocomput 2003:140–151.