Assessment of the reliability of protein-protein interactions and protein function prediction

M. Deng, F. Sun, T. Chen

Pacific Symposium on Biocomputing 8:140-151(2003)

ASSESSMENT OF THE RELIABILITY OF PROTEIN-PROTEIN INTERACTIONS AND PROTEIN FUNCTION PREDICTION

MINGHUA DENG, FENGZHU SUN, TING CHEN

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

Abstract

As more and more high-throughput protein-protein interaction data are collected, the task of estimating the reliability of different data sets becomes increasingly important. In this paper, we present our study of two groups of protein-protein interaction data, the physical interaction data and the protein complex data, and estimate the reliability of these data sets using three different measurements: (1) the distribution of gene expression correlation coefficients, (2) the reliability based on gene expression correlation coefficients, and (3) the accuracy of protein function predictions. We develop a maximum likelihood method to estimate the reliability of protein interaction data sets according to the distribution of correlation coefficients of gene expression profiles of putative interacting protein pairs. The results of the three measurements are consistent with each other. The MIPS protein complex data have the highest mean gene expression correlation coefficients (0.256) and the highest accuracy in predicting protein functions (70% sensitivity and specificity), while Ito's Yeast two-hybrid data have the lowest mean (0.041) and the lowest accuracy (15% sensitivity and specificity). Uetz's data are more reliable than Ito's data in all three measurements, and the TAP protein complex data are more reliable than the HMS-PCI data in all three measurements as well. The complex data sets generally perform better in function predictions than do the physical interaction data sets. Proteins in complexes are shown to be more highly correlated in gene expression. The results confirm that the components of a protein complex can be assigned to functions that the complex carries out within a cell. There are three interaction data sets different from the above two groups: the genetic interaction data, the in-silico data and the synexpress data. Their capability of predicting protein functions generally falls between that of the Y2H data and that of the MIPS protein complex data. The supplementary information is available at the following Web site: http://www-hto.usc.edu/∼msms/AssessInteraction/.

1 Introduction

The development of high-throughput bio-techniques for functional genomic analysis generated a large amount of protein-protein interaction data. These include the yeast two-hybrid assay [1,2,3] and mass spectrometry [4,5]. Several databases have been developed to collect different sources of protein interaction data, including the Munich Information Center for Protein Sequences (MIPS) [6], the Database of Interacting Proteins (DIP)[7], and the Biomolecular Interaction Network Database (BIND)[8]. However, even using the same experimental technique of the yeast two-hybrid assay, Ito's data and Uetz's data share few overlaps[3], which suggest that errors are present in these data sets. In this paper, we estimate the reliability of these data sets using three different measurements: (1) the distribution of gene expression correlation coefficients, (2) the reliability based on gene expression correlation coefficients, and (3) the accuracy of protein function predictions.

Interacting proteins are more likely to be involved in similar biological processes and functions and thus they are more likely to be co-expressed. Based on this observation, Grigoriev (2001)[9] first showed that the mean correlation coefficient of gene expression profiles for interacting protein pairs is higher than that for random protein pairs. Therefore, the significance of a set of protein interaction data can be tested by comparing the mean correlation coefficient of gene expression profiles with that of random protein pairs [10,11,12]. Using the same idea, Ge et al. (2001) [13] showed that interacting protein pairs are more likely to be in the same cluster of gene expression data than random pairs. Although this idea can be used to study the significance of a set of protein interaction data, it cannot be used to estimate the *reliability* of this set. Here the *reliability* is defined as the fraction of real interactions over the observed protein interactions.

Two methods have been proposed to estimate the reliability of a set of putative protein interactions. Mrowka et al. (2001)[14] used the protein physical interactions in MIPS as the reference of real interactions and estimated the reliability of a putative set by comparing the distributions of correlation coefficients with those of random pairs. They used a bootstrap method to count how many random pairs needed to be added to the reference data to create the same statistical behavior of gene expression correlation coefficients as the putative interaction data, and then they computed the reliability from this sampling data. Deane et al. (2002)[15] used INT, a subset of DIP interactions that are derived from small-scale experiments, as the reference for real interactions. They assumed that the distribution of the square of the Euclidian distance between expression profiles of putative interacting pairs is a mixture

of distributions for real interacting pairs and for random pairs. They then used a least-square approach to estimate the reliability of the putative protein interaction data. Although both ideas are novel and interesting, their estimation methods do not optimally correspond to their models. Other methods have also been proposed to assess the reliability by looking at protein functions [16] and "interaction generality"[17].

In this paper, we develop a maximum likelihood estimation(MLE) method for estimating the reliability of several interaction data sets based on the model of Deane et al. (2002). We studied two groups of interaction data. The first group included the MIPS, Uetz's and Ito's interaction data, all of which contain pairwise physical interactions. In this group, the interactions in MIPS were treated as real interactions, and we estimated the reliability of Uetz's data, Ito's data and Ito's data with multiple IST hits (Ito1IST, $\cdots$, Ito8IST). We show that the reliability of Uetz's interaction data (53%) is much higher than the reliability of Ito's data (17%) and is comparable to that of the Ito2IST data with at least two IST hits (56%). The reliability of Ito's data generally increases with the number of IST hits, which is consistent with our intuition that multiple IST hits could reduce the false positives significantly. The second group contained protein complexes: the MIPS complexes, the TAP data, and the HMS-PCI data. They provide protein components in a complex but do not give the physical interactions. The two groups are different, although a relationship does exist, since proteins in the same complex are more likely to physically interact with one another. The two groups have to be treated differently. For the protein complex data, the meaning of reliability is the fraction of protein pairs that are in the same protein complex in the putative complex data. We used the same MLE approach as above to estimate the reliability of the protein complex data using the MIPS complexes as the true complexes. The reliability of the TAP data and the HMS-PCI data are 58% and 25%, respectively.

Several methods have been developed to predict functions using protein-protein interaction data [18,19,20,21]. The basic assumption behind these studies is that proteins involved in similar functions are more likely to be interacting. In this paper, we consider two methods: the neighborhood-counting method and the chi-square method for the prediction of protein function. We use a leave-one-out method to estimate the accuracy of predictions, and we compare the results from different data sets. As expected, the reference data sets show higher accuracy than others in predicting functions. The MIPS complex data have the highest accuracy (70% sensitivity and specificity), while Ito's data have the lowest accuracy (15% sensitivity and specificity). The results are consistent with the estimation of the reliability of the protein interaction data.

2 Method

2.1 Estimating the reliability of a putative protein interaction data set

The reliability of a set of putative protein interactions is defined as the fraction of real protein interactions over all the putative protein interactions. Let α be the reliability of a given set of putative interactions. Let $O(\cdot)$, $T(\cdot)$ and $R(\cdot)$ be the distributions of the correlation coefficients of the gene expression profiles for the given set of putative interaction pairs, the true interacting pairs and the random pairs, respectively. Then we should have

$$O(\cdot) = \alpha T(\cdot) + (1 - \alpha)R(\cdot).$$

Suppose we split the values of correlation coefficients into K bins. Let n_k be the number of observed interaction pairs in the kth bin. Let p_k and q_k be the fractions of real interactions and random pairs in the kth bin, respectively. Then the likelihood function can be defined as:

$$L(\alpha) = \prod_{k=1}^{L} (\alpha p_k + (1 - \alpha)q_k)^{n_k}. \tag{1}$$

$L(\alpha)$ is a convex function and we can use a classical gradient algorithm to estimate the parameter α, $\widehat{\alpha}$, by maximizing $L(\alpha)$.

To find the precision of the estimation, we use the following formula to calculate the variance of $\widehat{\alpha}$,

$$Var(\widehat{\alpha}) = \frac{1}{\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha}p_k + (1-\widehat{\alpha})q_k)^2}}.$$

2.2 Estimating the reliability of protein interactions from different experiments.

If we have protein interaction data from several experiments, how do we estimate the reliability of the different sets of putative interactions? For example, we have two sets of putative protein interactions, E_1 and E_2, with M_1 and M_2 pairs, respectively. How do we estimate the reliability of the putative interacting pairs in $E_1/E_1 \cap E_2$, $E_2/E_1 \cap E_2$, and $E_1 \cap E_2$? If the number of protein pairs in a set is large, the above approach can be applied. On the other hand, if the number of protein pairs is not too large, we propose the following method.

As in Deng et al. [10], we define the false positive rate (fp) and the false negative rate (fn) for a specific data set, where the *false positive* rate is the probability that two proteins do not interact in reality but are observed to be interacting in the experiment. The *false negative* rate is the probability that two proteins interact in reality but are not observed to be interacting in the experiment. Let O_{ij} and P_{ij} be the variables for the observed and the real interaction for proteins P_i and P_j, respectively, with value 1 for interaction and 0 for no interaction. Then

$$fp = \Pr(O_{ij} = 1 \,|\, P_{ij} = 0), \qquad fn = \Pr(O_{ij} = 0 \,|\, P_{ij} = 1). \tag{2}$$

Thus, the probability for the observed protein-protein interaction is

$$\Pr(O_{ij} = 1) = \Pr(P_{ij} = 1)(1 - fn) + \Pr(P_{ij} = 0)fp.$$

As in Mrowka et al.[14] and Deane et al. [15], the reliability of a protein-protein interaction data set is measured by the fraction of real interactions in the data set, denoted by α. Let K_r and K_n be the sizes of *real interactions* and *non-interactions*, respectively, and let M be the size of the data set. We assume that the observed interactions and non-interactions are random samples from the real interaction set and the non-interaction set, respectively. The false positive rate and false negative rate can be estimated as

$$fp = \frac{M(1-\alpha)}{K_n}, \qquad fn = 1 - \frac{\alpha M}{K_r}. \tag{3}$$

Using the above formula, we can estimate the reliability of a putative protein-protein pair given two protein-protein interaction sets. We use the following notation for the kth data set:

$O_{ij}^{(k)}$: observed interaction result for P_i and P_j,
$fp^{(k)}$ and $fn^{(k)}$: false positive rate and false negative rate,
α_k: reliability,
M_k: the number of interaction pairs.

From equations 2 and 3, we have,

$$\begin{aligned} &\Pr(P_{ij} = 1 | O_{ij}^{(1)} = 1, O_{ij}^{(2)} = 1) \\ =&\frac{\Pr(P_{ij} = 1)(1 - fn^{(1)})(1 - fn^{(2)})}{\Pr(P_{ij} = 1)(1 - fn^{(1)})(1 - fn^{(2)}) + \Pr(P_{ij} = 0)fp^{(1)}fp^{(2)}} \\ =&\frac{\alpha_1\alpha_2}{\alpha_1\alpha_2 + \frac{K_r}{K_n}(1-\alpha_1)(1-\alpha_2)}. \end{aligned} \tag{4}$$

$$
\begin{aligned}
&\Pr(P_{ij}=1|O_{ij}^{(1)}=1,O_{ij}^{(2)}=0)\\
=&\frac{\Pr(P_{ij}=1)(1-fn^{(1)})fn^{(2)}}{\Pr(P_{ij}=1)(1-fn^{(1)})fn^{(2)}+\Pr(P_{ij}=0)fp^{(1)}(1-fp^{(2)})} \qquad (5)\\
=&\frac{\alpha_1(1-\frac{\alpha_2 M_2}{K_r})}{\alpha_1(1-\frac{\alpha_2 M_2}{K_r})+(1-\alpha_1)(1-\frac{(1-\alpha_2)M_2}{K_n})}.
\end{aligned}
$$

Therefore, if $K_r << K_n$ (which is true for yeast), the reliability for the intersection of both interaction sets will be very high, close to 1. If $M_k << K_r$ and $M_k << K_n$ such as in the yeast two-hybrid assays of Ito's and Uetz's, the reliability for $E_1/E_1 \cap E_2$ is very close to α_1, and the reliability for $E_2/E_1 \cap E_2$ is very close to α_2.

2.3 Protein function prediction using protein-protein interaction data

We used two simple methods to predict protein functions based on protein-protein interaction data. One is referred to as the "neighborhood-counting method" [18], which assigns k functions to a protein with the k largest frequencies in its interacting partners. The other, referred to as the "chi-square method" [19], assigns k functions to a protein with the k largest chi-square scores. The chi-square score for a function j and a protein P_i is defined as

$$S_i(j) = \frac{[n_i(j) - e_i(j)]^2}{e_i(j)},$$

where $n_i(j)$ is the number of interaction partners of protein P_i having function j, $e_i(j) = n_i(j) \times p_j$ is the expected number of partners having function j, and p_j is the fraction of proteins having function j among all the proteins.

3 Results

Our study assessed the reliability of two different groups of protein-protein interaction data: the protein physical interaction data and the protein complex data. The protein physical interaction data included two yeast two-hybrid data sets, by Uetz et al. [1] and Ito et al. [2,3], and DIP (http://dip.doe-mbi.ucla.edu), a collection of protein interactions from the literature and the yeast two-hybrid assays. A special feature of Ito's protein-protein interactions is that each interaction is accompanied by an IST number, indicating how many times the interaction was observed. We compared these data with the experimentally determined MIPS physical interaction data set (http://mips.gsf.de). The protein complex data included two data sets, the TAP data [4] and the HMS-PCI data [5],

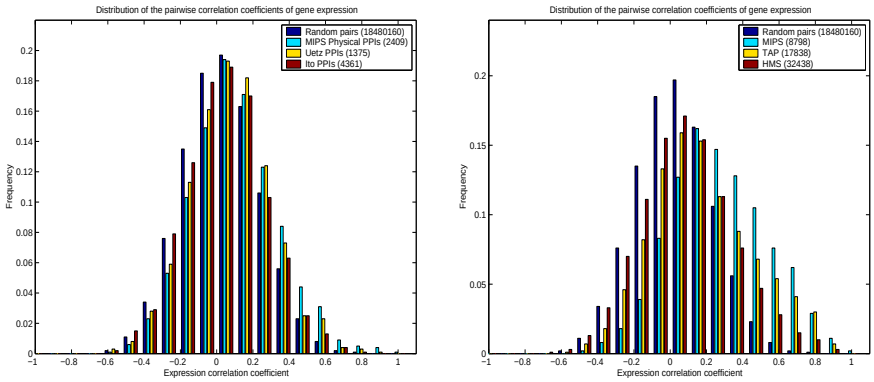

Figure 1: Distribution of expression coefficients for Uetz, Ito, MIPS physical interactions and random protein pairs (left), and that for TAP, HMS-PCI, MIPS complex data and random protein pairs (right).

obtained by systematic purification of protein complexes and protein identification via mass spectrometry. We compared them with a set of experimentally determined protein complexes called "MIPS Complex" [6]. For completeness, we also included the data obtained by other methods: co-expressed proteins measured at mRNA levels ("SynExpress"), computationally predicted interactions ("In-silico") and genetic interactions [16] ("Genetic"). We used the YPD [22] protein names in all data sets.

3.1 Distribution of gene expression correlation coefficients

We computed the correlation coefficient for every interacting protein pair using the cell cycle gene expression data [23], which contains 6,080 genes with 77 data points (2 cln3, 2 clb, 18 alpha, 24 cdc15, 17 cdc28, and 14 elut). Figure 1 shows the distributions of pairwise correlation coefficients. One is for all gene pairs, the MIPS physical interactions, Uetz's data and Ito's data. The other is for all gene pairs, the MIPS complex protein pairs, the TAP data and the HMS-PCI data. The distributions for the other three sets: "SynExpress", "In-silico", and "Genetic", are not included, as they cannot be easily classified as being in either of the above two groups.

The statistical significance for the difference between the mean expression correlation coefficient of a putative interaction set and that of random pairs is measured by the T-score and the P-value for the null hypothesis of no difference between the sample mean and the mean of random gene pairs. The T-scores

are calculated as the standard two sample T-test statistic.

Table 1 gives some descriptive statistics for the distributions of correlation coefficients for different data sets in the three groups: the protein physical interaction data, the protein complex data, and the data obtained by other methods. As expected, the MIPS complex data has the largest mean and the largest T-score among all the data sets. In the first group, the MIPS physical interaction data has the largest mean, while Ito's data have the smallest mean and the smallest T-score. Generally, as the IST number increases, the mean for Ito's data increases as well. In the second group, the TAP data shows a higher mean than the HMS-PCI data.

Data	#Pairs	Mean	Variance	T-score	P-value
Random	18480160	0.0305	0.200	0.00	–
Physical Interactions					
MIPS Physical	2409	0.0985	0.224	16.71	6.31e-063
DIP	14351	0.0852	0.236	32.78	5.63e-236
Uetz	1375	0.0692	0.210	7.18	3.70e-013
Ito1IST	4361	0.0410	0.209	3.47	2.64e-004
Ito2IST	1408	0.0714	0.214	7.69	7.82e-015
Ito3IST	751	0.0833	0.223	7.23	2.42e-013
Ito4IST	541	0.0941	0.217	7.40	6.85e-014
Ito5IST	442	0.0979	0.223	7.09	6.96e-013
Ito6IST	351	0.0821	0.210	4.84	6.75e-007
Ito7IST	291	0.0883	0.217	4.94	4.04e-007
Ito8IST	257	0.0938	0.223	5.08	1.95e-007
Protein Complex					
MIPS Complex	8798	0.2560	0.250	105.90	0.00
TAP	17838	0.1642	0.270	89.31	0.00
HMS-PCI	32438	0.0801	0.245	44.69	0.00
Other Methods					
SynExpress	16063	0.1650	0.238	85.28	0.00
In-silico	7152	0.1111	0.234	34.10	3.89e-255
Genetic	878	0.0990	0.240	10.16	1.67e-024

Table 1: Statistics of distributions of gene expression correlation coefficients for different protein-protein interaction data sets.

3.2 Reliability of the different data sets

We estimated the reliability of the different data sets using maximum likelihood estimation on the distribution of gene expression correlation coefficients. As in equation 3, we needed to specify the distribution for real interactions and that for non-interactions. We used the distribution for random pairs as that for the non-interactions, since it is believed that the size of real interactions is much smaller than the size of the non-interaction pairs. We chose the MIPS physical interactions as the reference for the physical interaction data and the MIPS complex data as the reference for the TAP and the HMS-PCI complex data.

Data	Pairs	PairsExp	α	Variance
Physical interactions				
Uetz	1436	1375	0.529	0.0843
DIP	14454	14351	0.815	0.0244
Ito1IST	4443	4361	0.167	0.0383
Ito2IST	1469	1408	0.558	0.0831
Ito3IST	802	751	0.753	0.1144
Ito4IST	584	541	0.895	0.1436
Ito5IST	476	442	0.964	0.1567
Ito6IST	379	351	0.676	0.1768
Ito7IST	312	291	0.791	0.1942
Ito8IST	276	257	0.878	0.2054
Protein Complex				
TAP	17962	17838	0.585	0.0081
HMS-PCI	32667	32438	0.248	0.0053

Table 2: Reliability of the physical interaction data (Uetz's, DIP, and Ito's with different IST hits) and the protein complex data (TAP and HMS-PCI).

The results are listed in Table 2. The table shows that the Ito5IST data, with ≥ 5 IST hits, are the most reliable, with $\alpha = 0.96$, while the Ito1IST data are the least reliable with $\alpha = 0.17$. The Ito2IST data are as reliable as the Uetz data. The TAP data are more reliable than the HMS-PCI data. Again, the maximum likelihood approach cannot be applied to the other three data sets because they cannot be easily classified into either of the above two groups.

3.3 Cellular role prediction based on different data.

We applied the neighborhood-counting method and the chi-square method[20] to predict protein functions based on the protein-protein interaction data. Both methods assign functions to a protein based on the functions of its immediate interaction proteins. The functional annotations were obtained from YPD, which assigns protein to three functional categories: "cellular role", "subcellular localization", and "biochemical function". Here, we considered the functional annotation based on the cellular role. Up to April 8, 2002, YPD included 6,416 proteins, among which 3,894 proteins have been assigned to one or more functions of 43 cellular roles, and the rest 2,522 proteins are unknown.

We used a leave-one-out method to measure the accuracy of the predictions. The leave-one-out method randomly selects a protein with known functions, assuming its functions as unknown, and then uses the neighborhood-counting method or the chi-square method to predict its functions. Finally, the predictions were compared with the actual functions of the protein. We repeated the leave-one-out experiment for K known proteins, $P_i, \cdots, P_K$. Let n_i be the number of functions for protein P_i in YPD, m_i be the number of *predicted* functions for protein P_i, and k_i be the overlap between them. The

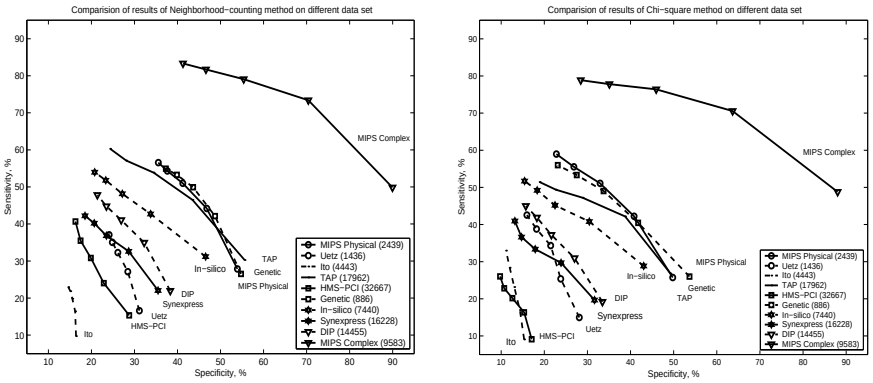

Figure 2: Sensitivity and specificity of functional predictions for different protein-protein interaction data sets using the neighborhood-counting (left) and the chi-square (right) methods.

specificity (SP) and the sensitivity (SN) can be defined as

$$SP = \frac{\sum_i^K k_i}{\sum_i^K m_i}, \qquad SN = \frac{\sum_i^K k_i}{\sum_i^K n_i} \tag{6}$$

Figure 2 shows the relationship between specificity and sensitivity for the neighborhood-counting method and the chi-square method on different protein interaction data sets. As expected, the MIPS protein complex data have the best performance, while the Ito1IST data do the worst. Excluding the MIPS protein complex data, a group of three data sets (the TAP data, the MIPS Physical data and the Genetic Interaction data) shows better performance than the others. Overall, the results are very consistent with those of the distributions of expression correlation coefficients and the reliability estimation.

4 Discussions

We studied two groups of protein-protein interaction data, the physical interaction data and the protein complex data, and estimated the reliability of these data sets using three different measurements: the distribution of gene expression correlation coefficients, the reliability based on gene expression correlation coefficients, and the accuracy of protein function prediction. We separated protein complex data from protein physical interaction data because of their obvious difference: not all protein pairs in a complex interact with one another, and not all physically interacting protein pairs are in the same complex. Many

protein complexes such as ribosomes and RNA Polymerases are essential for a cell, and the interactions within a complex are generally more stable and stronger and have a longer life span than most other physical interactions, while other physical interactions include other important interactions such as signal transductions. Our results confirm that the components of a protein complex can be assigned to functions that the complex carries out within a cell. The complex data sets generally perform better in function predictions than do the physical interaction data sets. Meanwhile, proteins in complexes are shown to be more highly correlated in gene expression, as well.

The results of the three measurements are consistent with one another. For example, the MIPS protein complex data have the highest mean gene expression correlation coefficient (0.256) and the highest accuracy in predicting protein functions (70% sensitivity and specificity), while Ito's Y2H data have the lowest mean (0.041) and the lowest accuracy (15% sensitivity and specificity); Uetz's data are more reliable than Ito's data in all three measurements, and the TAP data are more reliable than the HMS-PCI data in all three measurements. The Ito1IST data containing many interactions with only 1 IST hit are believed to contain many false positives and are the least reliable. However, the Ito2IST data containing interactions with at least 2 IST hits have better performance in all three measurements. This confirms that multiple ISTs can reduce false-positives in Y2H assays significantly.

There are three interaction data sets different from the above two groups: the genetic interaction data, the In-silico data, and the Synexpress data. Their capability of predicting protein functions generally falls between that of the Y2H data and that of the MIPS protein complex data. It should be noted that these interactions contain not only real physical interactions but also other protein pairs that are functionally associated. These data are important in understanding protein functions on a global scale.

We used three different ways to assess protein-protein interaction data. Although they show consistency with one another, there are some limitations to the reliability estimation. First, we assumed that the MIPS interactions and complexes were unbiased real interactions. However, the MIPS data sets may contain errors and may be biased to certain functions, cell compartments, mRNA expression levels, and so on, and thus the measurements may be inaccurate. Second, we used the gene expression profiles as a measurement. Ideally, a good measurement should itself be unbiased for assessing the reliability of interactions, and the data collection process should be unbiased as well. In this study, the gene expression data we used are certainly biased. It is also well-known that only a small set of interacting protein pairs are correlated at the mRNA expression levels. Another limitation is the assumption that

known proteins have all their functions annotated. Based on this assumption, we can estimate the accuracy. In reality, the proteins may have un-discovered functions.

Acknowledgments

This research was partially supported by National Institutes of Health Grant DK53392, National Institutes of Health Grant 1-R01-RR16522-01, National Science Foundation EIA-0112934, and the University of Southern California.

References

1. P. Uetz, et al. *Nature* **403**, 623 (2000).
2. T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara and Y. Sakaki. *Proc. Natl. Acad. Sci. USA* **97**, 1143 (2002).
3. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki. *Proc. Natl. Acad. Sci. USA* **98**, 4569 (2001).
4. A. Gavin, M. Böche, R. Krause, P. Grandi,M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A. Michon, C. Cruciat *et al.*. *Nature* **415**, 141 (2002).
5. Y. Ho, *et al.*. *Nature* **415**, 180 (2002).
6. H.W. Mewes, D. Frishman, U. Guldener,G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd and B. Weil. *Nucleic Acids Research* **30**, 31 (2002).
7. I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S. Kim and D. Eisenberg. *Nucleic Acids Research* **30** 303 (2002).
8. G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, C.W. Hogue. *Nucleic Acids Research***29**, 242 (2001).
9. A. Grigoriev. *Nucleic Acid Res.* **29**, 3513 (2001).
10. M. Deng, S. Mehta, F. Sun and T. Chen. *Proceedings of the Sixth International Conference on Computational Molecular Biology (RECOMB2002)*, 117 (2002).
11. R. Jansen, D. Greenbaum and M. Gerstein. *Genome Research* **12**, 37 (2002).
12. P. Kemmeren, N.L.V. Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma and F.C.P. Holstega. *Molecular Cell* **9**, 1133 (2002).
13. H. Ge, Z. Liu, G.M. Church and M. Vidal. *Nature Genetics*, **29**, 482 (2001).
14. R. Mrowka, A. Patzak and H. Herzel. *Genome Research* **11**, 1971 (2001).
15. C.M. Deane, L. Salwinski, I. Xenarios and D. Eisenberg. *Molecular and cellular proteomics* (2002)
16. C.V. Mering, R. Krause, M. Snel, S.G. Oliver, S. Fields and P. Bork. *Nature* **417**, 399 (2002).
17. Saito, R., Suzuki, H. and Hayashizaki, Y.. 2002. *Nucleic Acids Research* **30:** 1163-1168.
18. B. Schwikowski, P. Uetz and S. Fields. *Nature Biotechnology* **18**, 1257 (2000).
19. H. Hishigaki, K. Nakai, T. Ono, A. Tanigami and T. Takagi. *Yeast* **18**, 523(2001).
20. M. Deng, K. Zhang, S. Mehta, T. Chen and F. Sun. *Proceeding of IEEE Computer Society Bioinformatics Conference*, Page 197-206 (2002).
21. M. Fellenberg, K. Albermann, A. Zollner, H.W. Mewes, and J. Hani. In *Proc. of the Eighth Int. Conf. on Intelligent System for Molecular Biology (ISMB2000)*, 152 (2000).
22. M.C. Costanzo, et al. *Nucleic Acids Research* **29:**, 75 (2001).
23. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher. *Molecular Biology of the Cell* **9**, 3273 (1998).