Assessment of the Reliability of Protein-Protein Interactions Using Protein Localization and Gene Expression Data

Hyunju Lee[1] **Minghua Deng**[2] **Fengzhu Sun**[3] **Ting Chen**[3]

[1]*Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA.*

[2]*School of Mathematics, Beijing University, P. R.China*

[3]*Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA.*

Email: tingchen@usc.edu, fsun@usc.edu

ABSTRACT: Estimating the reliability of protein-protein interaction data sets obtained by high-throughput technologies such as yeast two-hybrid assays and mass spectrometry is of great importance. We develop a maximum likelihood estimation method that uses both protein localization and gene expression data to estimate the reliability of protein interaction data sets. By integrating protein localization data and gene expression data, we can obtain more accurate estimates of the reliability of various interaction data sets.

We apply the method to protein physical interaction data sets and protein complex data sets. The reliability of the yeast two-hybrid interactions by Ito et al. (2001) is 27%, and that by Uetz et al.(2000) is 68%. The reliability of the protein complex data sets using tandem affinity purification-mass spectrometry (TAP) by Gavin et al. (2002) is 45%, and that using high-throughput mass spectrometric protein complex identification (HMS-PCI) by Ho et al. (2002) is 20%. The method is general and can be applied to analyze any protein interaction data sets.

1 INTRODUCTION

High-throughput bio-techniques have generated many protein-protein interaction data sets in several model organisms-for example, *Drosophila melanogaster* [1] and *Caenorhabditis elegans* [2]. Analysis of data sets generated by these techniques has shown that they contain a significant percentage of false positives[3]. In order to use these protein-protein interaction data sets for biological studies, it is important to have an accurate estimate of the fraction of true interactions among the observed interactions, which we refer as the *reliability*. Several methods have been developed to estimate the reliability of different experimentally derived interaction data sets. One of the most widely used approaches is to compare the empirical distribution of gene expression correlation coefficients of any gene pairs among a putative interaction data set with those of random pairs and those of real interacting gene pairs. Mrowka et al. (2001) used a bootstrap method to estimate the number of random gene pairs that need to be added to the reference true interaction data to create the same statistical behavior of gene expression correlation coefficient as that of the putative interaction data[4]. Deane et al. (2002) and Deng et al. (2003) assumed that the putative interactions were a mixture of random gene pairs and real interaction gene pairs[5, 6]. Deane et al. (2002) used a least-square approach, and Deng et al. (2003) used a maximum likelihood estimation (MLE) approach to estimate the reliability of observed putative interactions.

Several other groups developed methods to estimate the probability of interaction for each gene pair based on the topology of the observed interaction network. Saito et al. (2003) defined an "interaction generality measure (IG2)" using a complicated procedure that considered the interaction neighbors of either or both of the genes [7]. They showed that reliable protein interactions had significantly lower IG2 values than less reliable interactions. Goldberg et al. (2003) estimated the probability of interaction for a pair of genes u and v on the basis of their individual interaction partners, N_u and N_v, respectively, and their intersection, $N_u \bigcap N_v$[8]. They defined several neighborhood cohesiveness measures based on $(N_u, N_v, N_u \bigcap N_v)$ and showed that these neighborhood cohesiveness measures strongly correlated with the true interaction status. These measures can be used to estimate the reliability of interaction for each gene pair.

In this study, we proposed a new method to estimate the reliability of protein interaction data sets. Our method is unique in two ways. First, we combined two measures for the reliability of protein interactions data sets: the correlation coefficients in gene expression profiles and co-localization (in cellular locations) of protein interactions. Second, we developed a new method to estimate the reliability using the protein location data, and a new method to combine the two measures so as to lower the variance of the estimates. The protein location data set was obtained from Huh et al. (2003) [9].

Huh et al. (2003) generated a large-scale yeast protein localization map. By comparing the number of interaction pairs in GRID [10] between pairs of locations with that for one randomized network, defined as the *enrichment*, they showed that protein interactions are strongly enriched among co-localized proteins and proteins between specific cellular locations. Since the number of interaction pairs can be different for different randomized interaction networks, the enrichment defined in Huh et al. (2003) depends upon the realization of the randomized network. In this paper, we first modify the definition of enrichment for the protein interactions of Huh et al. (2003) between any pair of locations. Second, we draw the enrichment map for several different putative interaction data sets and show that the enrichment maps contain information about their reliability. Third, we develop a novel computational method based on an expectation-maximization (EM)

algorithm to estimate the reliability of putative interaction data sets using the enrichment patterns. Finally, we develop an integrated approach for reliability estimation by combining gene expression data and the localization data. By combining all of these data, we can obtain a more accurate estimate of the reliability of protein interaction data sets.

2 Algorithms

2.1 Data Sources

We assessed the reliability of two groups of protein-protein interactions: physical interactions and protein complexes using the MIPS physical interaction data set [11] (http://mips.gsf.de) and the MIPS complex data set as the "gold standard" for each group. The protein physical interaction data sets included the two yeast two-hybrid data sets of Uetz et al. (2000) and Ito et al. (2001), and DIP [12, 13, 14], a collection of protein interactions from the literature and the yeast two-hybrid assays. The protein complex data sets included the data obtained by tandem affinity purification-mass spectrometry(TAP) by Gavin et al. (2003) [15] and the data obtained by high-throughput mass spectrometric protein complex identification (HMS-PCI) by Ho et al. (2003) [16]. The protein localization data were obtained from Huh et al. (2003) [9], and the gene expression data were obtained from Spellman et al. (1998) [17].

2.2 Estimation of Enrichment Ratio

In order to use localization data for reliability estimation, we first provide preliminary evidence that reliability of interactions does correlate with enrichment patterns based on localization data. Proteins in functionally and physically related pairs of locations are more likely to interact with each other than proteins in other pairs of locations. This preference of interactions shows particular patterns that can be used to measure the reliability of protein-protein interaction data. Given a putative protein interaction network, we calculate the enrichment of protein interactions for a pair of locations k and k' as

$$\mathrm{En}(k,k') = \frac{\mathrm{Obs}(k,k')}{\mathrm{Exp}(k,k')}, \tag{1}$$

where $\mathrm{Obs}(k,k')$ is the observed number of interactions and $\mathrm{Exp}(k,k')$ is the expected number of interactions between a pair of locations k and k' for randomized protein networks for which the number of interactions for each protein is the same as that of the given network. A randomized protein network is generated as follows. We randomly select two interactions from the network-i.e., $a-b$ and $c-d$-and shuffle them to $a-c$ and $b-d$. We repeat this process 1,000 times. In summary, $\mathrm{Exp}(k,k')$ is calculated as follows.

1. Generate a randomized network as mentioned above.
2. Calculate $N_{kk'}$, the number of interactions between locations (k,k').
3. Repeat steps 1 and 2 B times.
4. $\mathrm{Exp}(k,k') = \frac{1}{B}\sum_b N_{kk'}(b)$.

In this paper, we let $B = 1{,}000$.

The enrichment measure defined here is a modification of the same measure in Huh et al. (2003), where they used $B = 1$. This new measure eliminates or reduces the variability of enrichment due to the generation of a randomized interaction network.

The other significant difference from Huh et al. (2003) is the calculation of $\mathrm{Obs}(k,k')$, and similarly $N_{kk'}$, considering that a protein may belong to multiple locations. For example, suppose that two proteins, P1 and P2, interact and that protein P1 belongs to locations (bud, bud neck) and protein P2 belongs to locations (bud, ER). Huh et al. (2003) counted the (P1, P2) interaction as four interactions between (bud, bud), (bud, bud neck), (ER, bud), and (ER, bud neck), respectively. In their supplementary materials, they also took the (P1, P2) interaction as one interaction between (bud, bud) only. This practice can be considered a greedy approach in that it takes the most likely location pair as the truth. In this paper, we formalize this problem as a missing value problem and calculate the number of interactions between pairs of locations using an EM algorithm (see supplementary material [19]). The EM approach takes the likelihood of interactions between pairs of locations into consideration and assigns weights for different location pairs according to their likelihood. Therefore, the calculations based on EM should be more reasonable and should give more accurate estimate for the enrichment ratio.

2.3 Maximum Likelihood Estimation of Reliability Using Localization Data

Next we show how to use the localization data to estimate the reliability of putative interaction data sets. As in Deng et al. (2003), we assume that the putative interaction data set is a mixture of true interactions and random pairs. Let α be the fraction of true interactions, and let $O(\cdot)$, $T(\cdot)$, and $R(\cdot)$ be the probability distributions of pairs of locations for observed, true, and random interactions, respectively. Then we have

$$O(\cdot) = \alpha T(\cdot) + (1-\alpha)R(\cdot) \tag{2}$$

The distribution of pairs of locations for observed interactions based on the location data can be calculated as follows. Let $\hat{\theta}_{kk'}(\mathtt{true})$ and $\hat{\theta}_{kk'}(\mathtt{rand})$ be the fractions of true interactions and random pairs, respectively, for a pair of locations k and k', and let $Loc(i)$ be the set of locations to which protein i belongs. For a pair of proteins i and j, we define

$$p_{ij} = \sum_{k\in Loc(i),k'\in Loc(j)} \hat{\theta}_{kk'}(\mathtt{true}),$$

$$q_{ij} = \sum_{k\in Loc(i),k'\in Loc(j)} \hat{\theta}_{kk'}(\mathtt{rand}).$$

The likelihood of the observed data based on the protein localization data is

$$L_{loc}(\alpha) = \prod_{ij}(\alpha p_{ij} + (1-\alpha)q_{ij}). \tag{3}$$

To calculate $\theta_{kk'}(\texttt{true})$ for physical interactions, we use the MIPS physical interactions as the gold standard. Based on the data, we can use the EM algorithm [19] to calculate $\theta_{kk'}(\texttt{true})$. Similarly, we use the MIPS complex data as the gold standard for complex data. We use all of the protein pairs as the random set and employ the EM algorithm to calculate $\theta_{kk'}(\texttt{rand})$.

If $\theta_{kk'}(\texttt{true})$ and $\theta_{kk'}(\texttt{rand})$ are given, we can use a gradient descent algorithm to estimate the parameter α: $\hat{\alpha}$ by maximizing $L_{loc}(\alpha)$. The variance of $\hat{\alpha}$ is

$$Var(\hat{\alpha}) = \frac{1}{\sum_{ij} \frac{(p_{ij}-q_{ij})^2}{(\hat{\alpha}p_{ij}+(1-\hat{\alpha})q_{ij})^2}}. \quad (4)$$

2.4 Estimating Reliability by Combining Localization and Gene Expression Data

Deng et al. (2003) developed a maximum likelihood method to estimate the reliability of putative interaction data sets using gene expression data. By combining gene localization and gene expression data, we can reduce the variance of the reliability estimate. For the gene expression data, we split the values of gene expression correlation coefficients into $N = 20$ equally spaced bins. Let n_k be the number of observed interaction pairs in the k-th bin, and let p_k and q_k be the fraction of true interactions and random pairs in the k-th bin, respectively. Then the likelihood for the observed interaction data based on both the cellular location data and the gene expression data is

$$L_{both}(\alpha) = \prod_k \prod_{ij} (\alpha p_k p_{ij} + (1-\alpha) q_k q_{ij}). \quad (5)$$

The variance of the estimated reliability can be similarly calculated as in equation 4.

3 Result

Figures 1 and 2 show the cellular location maps of protein interactions in each pair of the 22 cellular locations based on the enrichment defined in the methods section with the calculation of Obs(k,k') as the number of interactions between pairs of locations using an EM algorithm.

The MIPS physical in Figure 1 shows the cellular location map for the MIPS physical interaction data set. The strong enrichment of protein interactions along the diagonal shows that proteins co-localized in some cellular locations are more likely to interact than others. We also observe the enrichment of protein interactions between off-diagonal pairs of cellular locations: Microtubule and Spindle pole, Bud Neck and Cell periphery, and Endosome and Vacuolar Membrane. The enrichment of these three pairs of cellular locations were also observed in the map by Huh et al. (2003). However, there are also significant differences between the enrichment map in this study and that in Huh et al. (2003): we did not find any statistical significance in the other 15 pairs of locations. Partly, this was due to the lack of protein interactions for pairs of locations such as Nuclear periphery and ER to Golgi, Golgi and Action, and Late Golgi and Early Golgi, but for most of them, they do not show a statistically significant enrichment of protein interactions.

Figure 1 also shows the cellular location maps of the DIP data set [14], Uetz's yeast two-hybrid interaction data set [12], and the Ito's yeast two-hybrid interaction data sets [13]. The maps for the DIP data set and the Uetz data set are similar to that for the MIPS physical interaction data. On the other hand, the map for Ito1 looks like a mixture of the map for MIPS physical and a random map. One of important factors in the enrichment map is the number of interactions. There are only 792, 712, and 368 interactions in the Uetz, Ito2, and Ito8 interaction sets, respectively. We have a total of 21*(21 + 1)/2 = 231 pairs of locations. A large number of pairs of cellular locations have less than 5 observed interactions: ER to Golgi (one protein having interaction with punctate composite) in the Uetz, microtubule (nine proteins having one interaction with cell periphery) in the Ito2, and peroxisome (nine proteins having one interaction with mitochondrion) in the Ito8. Therefore, the calculated enrichment ratios have large variation for the three data sets, and the enrichment itself is not reliable. For that reason, we indicate the statistical significance by the red color, meaning the p-value is less the 0.01.

Figure 2 shows the cellular location maps of protein complex data sets: MIPS, TAP [15], and HMS-PCI [16]. Bader et al. (2002) suggested two models to define protein-protein interaction in a complex set: the "spoke model" and the "matrix model"[18]. The spoke model assumes that the bait directly interacts with other proteins in a complex, and the matrix model assumes that all proteins in a complex interact. Because the MIPS complex does not have the bait, we used the matrix model for our result. For the TAP and HMS-PCI data, we also used the matrix model. The enrichment map of both the spoke model and the matrix model of the TAP and HMS-PCI data obtained from Bader et al. (2002) are shown in the supplementary materials [19]. These two different models give similar enrichment maps.

We estimate the reliability of the different data sets using an maximum likelihood estimation based on the distribution of protein interactions within the same cellular locations or between two cellular locations (see details in the methods section). We study two groups of interaction data sets and results are described in Table 1 along with the enrichment ratio of protein interactions within the same cellular locations.

The first group includes the MIPS, DIP, Uetz, and Ito interaction data sets, all of which contain pairwise physical interactions. The MIPS physical interactions are used as a true interaction data set. Similar to the results of the protein location enrichment ratios, the reliability of the Uetz data set (68%) is higher than that of the Ito1IST data set (27%), and is comparable to those of the DIP data (59%) and the Ito8IST data (61%). The second group includes the protein complexes such as the MIPS complex data, the TAP complex data, and the HMS-PCI complex data. We treat the MIPS complex data as a true protein complex data set. The reliability of the TAP data is 45%, and that of the HMS-PCI data is 20%.

We compare the reliability estimates based on the protein localization data with the reliability estimates of protein interaction data sets based on cell cycle gene expression data [17] by Deng et al. (2003) and a combination of the gene expression data and the protein localization data. The results are shown in Figure 3. Figure 3 shows that the reliability estimates based on the gene expression data only have the higher

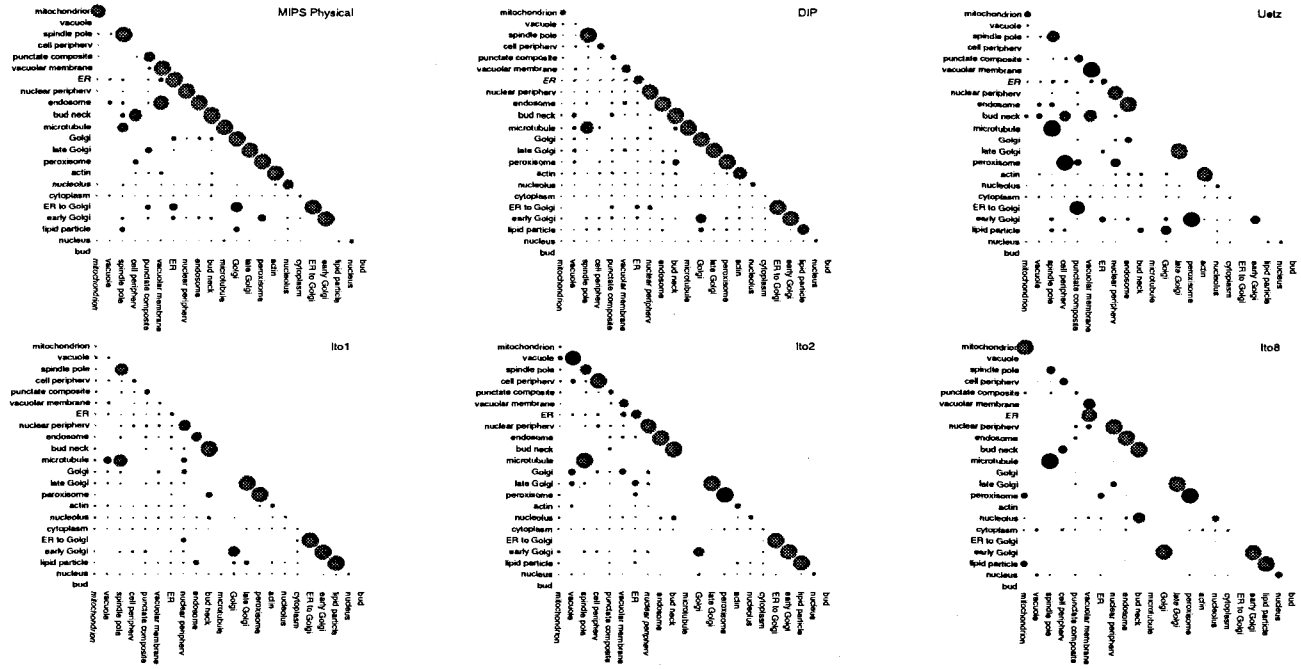

Figure 1: Enrichment ratio calculated with the prior of location probability of the protein for protein physical interactions in cellular locations. The size of the circle is proportional to the enrichment ratio, and the red color indicates that the p-value < 0.01. MIPS Physical: MIPS protein physical interactions; DIP: DIP protein interactions; Uetz: Uetz yeast two-hybrid interactions; Ito1: Ito's interaction with at least one hit; Ito2: Ito's interaction with at least two hits; Ito8: Ito's interaction with at least eight hits. For color figures, see [19].

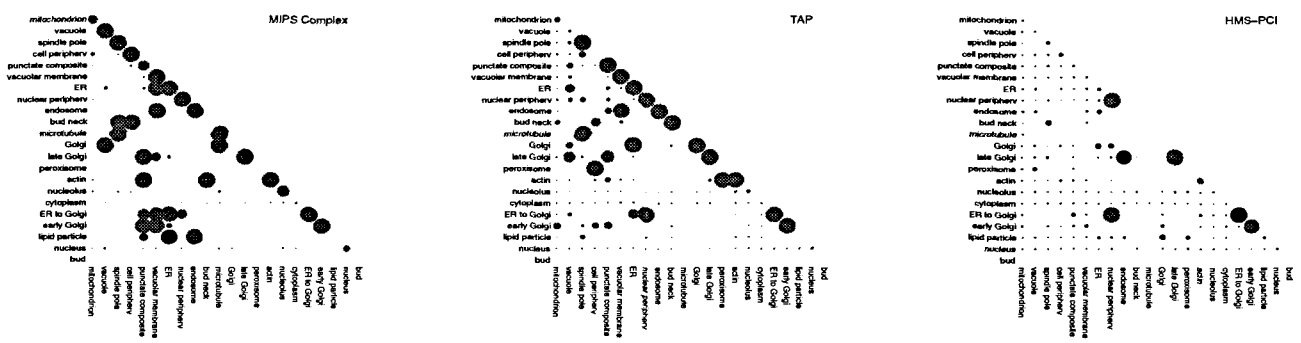

Figure 2: Enrichment ratio calculated with the prior of location probability of the protein; enrichment ratios for protein complexes in cellular locations. MIPS Complex: MIPS protein complex; TAP: TAP protein complexes; HMS-PCI: HMS-PCI protein complexes.

standard errors than the other two that included the protein localization data. The reliability estimates based on the protein localization data have much lower standard errors, and the reliability values are similar to the those based on a combination of the gene expression data and the protein localization data across all the protein interaction data sets. The benefit of combining two measures is to lower the standard errors. In general, these three estimates are consistent with each other (see supplementary materials for more information [19]). The results in this paper are also qualitatively consistent with that of Lee et al. (2004) [20].

4 Discussion

We studied the reliability of protein-protein interactions of observed physical interactions and protein complexes based on protein cellular location data using the maximum likelihood estimation method. We also integrated the cellular location information with gene expressions to increase the accuracy of reliability estimation of protein interactions.

First, we developed a new method to estimate protein interaction enrichment in each pair of locations. We showed the strong enrichment of protein interaction within the same cellular locations and pairs of locations having related functions based on the highly reliable MIPS interactions. We also obtained the protein interaction enrichment map for several observed protein interaction data sets. The enrichment maps for Uetz and DIP are more similar to that of MIPS than Ito, indicating that the enrichment maps can be used for reliability estimation. Second, we developed a maximum likelihood method for estimating the reliability of putative interaction data sets by representing the putative interaction data as a mixture of true interactions and random protein pairs. The difficulty in dealing with the localization data is that the protein can have

Physical Interaction Data		MIPS Physical	Uetz	DIP	Ito1	Ito2	Ito8
Localization	Pairs with locations	2,559	792	8,245	2,108	712	368
	Reliability	1.00	0.68	0.59	0.27	0.41	0.61
	Standard Deviation	-	0.0273	0.0082	0.0140	0.0259	0.0572
Gene Expression	Reliability	1.00	0.529	0.815	0.167	0.558	0.878
	Standard Deviation	-	0.0843	0.0244	0.0383	0.0831	0.2054
Both	Reliability	1.00	0.699	0.619	0.293	0.470	0.684
	Standard Deviation	-	0.0257	0.0076	0.0133	0.0253	0.0514

(a)

Complex Data		MIPS Complex	TAP	HMS-PCI
Localization	Pairs with location	7,091	11,100	18,158
	Reliability	1.00	0.45	0.20
	Standard Deviation	-	0.0063	0.0042
Gene Expression	Reliability	1.00	0.585	0.248
	Standard Deviation	-	0.0081	0.0053
Both	Reliability	1.00	0.516	0.205
	Standard Deviation	-	0.0056	0.0037

(b)

Table 1: Enrichment ratio of protein interactions within the same cellular locations and reliability based on cellular locations for (a) protein physical interaction data sets (MIPS, DIP, Uetz, and Ito), and (b) for protein complex data sets (MIPS, TAP and HMS-PCI). Ito1: Ito's interaction with at least one hit; Ito2: Ito's interaction with at least two hits; Ito8: Ito's interaction with at least eight hits.

multiple locations. Therefore, we estimated the probability of proteins in each location using an EM algorithm. Third, we developed an integrative method to combine the cellular location information with the gene expressions for reliability estimation. The integrated method resulted in small standard error for the reliability estimate.

The methods proposed in this paper can only be used to estimate the reliability of a putative set of protein interactions. They cannot, however, be used to estimate the probability that a particular pair of proteins interact. Methods such as those in Lee et al. (2004) [20] can be used to achieve this objective.

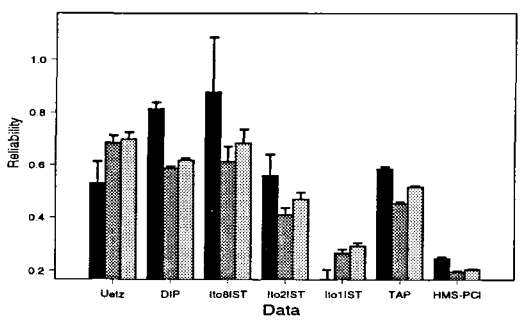

Figure 3: Reliability of protein interaction data sets based on gene expression data (black), cellular localization data (dark), and a combination of both (light grey). Ito8IST, Ito2IST, and Ito1IST are the sets of protein interactions with at least 8, 2 and 1 observations

REFERENCES

[1] L. Giot, J. S. Bader, C. Brouwer, and A. Chaudhuri. A protein interaction map of Drosophila melanogaster. Science, 302(5651):1727–1736, 2003.

[2] S. Li, C. M. Armstrong, and N. Bertin. A map of the interactome network of the metazoan C. elegans. Science, 303(5657):540–543, 2003.

[3] C. von Mering, R. Krause, R. Snel, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417:399, 2002.

[4] R. Mrowka R, A. Patzak, and H. Herzel. Is There a Bias in Proteome Research? Genome Research, 11:1971–1973, 2001.

[5] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics, 1(5):349–356, 2002.

[6] M. Deng, F. Sun, and T. Chen. Assessment of the Reliability of Protein-protein Interactions and Protein Function Prediction. Pacific Symposium of Biocomputing (PSB2003), pages 140–151, 2002.

[7] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with

a new interaction generality measure. Bioinformatics, 19(6):756–763, 2003.

[8] D. Goldberg and F. Roth. Assessing experimentally derived interactions in a small world. Proc. Natl. Acad. Sci. USA, 10(8):4372–4376, 2003.

[9] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J.S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. Nature, 425(6959):686–691, 2003.

[10] B. J. Breitkreutz, C. Stark, and M. Tyers. The GRID: The General Repository for Interaction Datasets. Genome Biol., 4:R23, 2003.

[11] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a Database for Genomes and Protein Sequences. Nucleic Acids Res, 30:31–34, 2002.

[12] P. Uetz, et al. A comprehensive analysis of proteinprotein interactions in Saccharomyces cerevisiae. Nature, 403:623–627, 2000.

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

[14] I. Xenarios, I. Salwinski, X. J. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 30:303–305, 2002.

[15] A. Gavin, M. Böche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. Michon, C. Cruciat, et al. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. Nature, 415:141–147, 2002.

[16] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 415(6868):180–183, 2002.

[17] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell, 9:3273–3297, 1998.

[18] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology, 20(10):991–997, 2002.

[19] http : //www – scf.usc.edu/ ~hyunjul/localsupple.html

[20] I. Lee, S. Date, A. Adai, and E. Marcotte. A probabilistic functional network of yeast genes. Science, 306:1555–1558, 2004.