doi:10.1016/S0022-2836(03)00239-0 J. Mol. Biol. (2003) **327**, 919–923

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

COMMUNICATION

How Reliable are Experimental Protein–Protein Interaction Data?

Einat Sprinzak, Shmuel Sattath and Hanah Margalit*

*Department of Molecular
Genetics and Biotechnology
Faculty of Medicine
P.O. Box 12272
The Hebrew University
of Jerusalem, Jerusalem 91120
Israel*

**Corresponding author*

Data of protein–protein interactions provide valuable insight into the molecular networks underlying a living cell. However, their accuracy is often questioned, calling for a rigorous assessment of their reliability. The computation offered here provides an intelligible mean to assess directly the rate of true positives in a data set of experimentally determined interacting protein pairs. We show that the reliability of high-throughput yeast two-hybrid assays is about 50%, and that the size of the yeast interactome is estimated to be 10,000–16,600 interactions.

Keywords: protein–protein interaction; genome-wide analysis; yeast two-hybrid; cellular localization; cellular-role

Databases of protein–protein interactions have expanded significantly during the last years, due to the recent high-throughput methods for identifying protein interactions at a genomic scale. On the one hand, this vast amount of data is considered as a rich source of information, from which new biological insight can be gained.[1–4] On the other hand, the accuracy of the data is often criticized.[5–8] Especially, the high-throughput methods are believed to contain many false positives, i.e. interactions that are identified in the experiment but never take place in the cell.[5,6,8,9] It is therefore essential to obtain an estimate of the reliability of the interactions documented by the various methods.[8,9] Here we provide a simple approach to compute the extent of correct interactions (true positives (TP)) in experimental protein–protein interaction data from *Saccharomyces cerevisiae*, and demonstrate its usefulness in assessing various experimental methods and in estimating the size of the yeast interactome.

We have assembled a database of the yeast *S. cerevisiae* interacting protein pairs from three public databases, MIPS,[10]† DIP,[11]‡ and BIND,[12]§ and from large compilations of interactions determined by genome-wide yeast two hybrid (Y2H) assays.[13,14]‖ After exclusion of redundancies, the entries in our database summed to a total of 9347 pairs of yeast interacting proteins. For each interacting pair, the experimental methods used to determine it were recorded. Information about complexes[10,15,16] was not used to infer novel binary interactions among the complex proteins, because binary relationships between proteins in a complex cannot be extracted without additional information. Involvement of a pair of proteins in a complex was recorded only if it supported binary interactions reported by other methods. Our documentation system has enabled us to assess the reliability of interactions in the database in general, and per experimental method. The latter were not treated individually, but grouped by their nature into several types of method categories: genetic, physical, immunological, biochemical, and the Y2H method (for a description of the methods included in each category see the legend to Table 1). The Y2H interactions were divided according to the different large-scale and "small-scale" studies in which they had been determined.

To assess the quality of the data we used two measures: the fraction of interacting proteins that were documented as localized in the same cellular compartment, and the fraction of interacting proteins that were annotated as having a common cellular-role. We expect that for true interactions, the interacting proteins should be localized to the same cellular compartment, at least at the time of interaction. It is also conceivable that they interact while participating in the same cellular process, i.e. sharing a common cellular-role.[2] Previous studies had used these two properties only

Abbreviations used: TP, true positives.

E-mail address of the corresponding author: hanah@md.huji.ac.il

† http://mips.gsf.de/proj/yeast/tables/interaction/

‡ http://dip.doe-mbi.ucla.edu/

§ http://www.bind.ca/

‖ http://depts.washington.edu/sfields/ and http://genome.c.kanazawa-u.ac.jp/Y2H/

0022-2836/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved

Table 1. Data sets of pairs of interacting proteins

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Homo-dimers were excluded. Values are rounded to integers.

[a] Abbreviations for method categories by which the interacting pairs were determined: All, all interacting protein pairs in the data; A, "small-scale" yeast two-hybrid method; A0, genome-scale yeast two-hybrid screening (GY2H) based on Uetz *et al.*;[13] A1, unpublished data of Uetz *et al.* (claimed by the author to be filtered for false positives less rigorously than the published data); A2, filtered core data from the genome-scale yeast two-hybrid screening by Ito *et al.*;[14] A3, data from Ito *et al.* not including core data;[14] B, physical methods (e.g. X-ray, mass spectrometry); C, genetic methods (e.g. suppression, synthetic lethals); D1, biochemical *in vitro* methods: (*in vitro* binding, cross-linking); D2, biochemical chromatography methods (e.g. affinity column, co-purification, gel filtration, chromatography); E1, direct immunological methods (e.g. co-immunoprecipitation); E2, indirect immunological methods (e.g. immunostaining, immunolocalization); 2M, 3M, 4M, interacting protein pairs discovered by at least two, three, four different method categories, respectively. (The number of interacting pairs in the subsets can be lower than in the original publications due to redundancies in the data or inconsistencies with Swissprot accession numbers.)

[b] Percentage co-localization (co-cellular-role) for each method category was calculated by dividing the number of interacting pairs with co-localized (shared cellular-role) pair-mates by the total number of annotated interacting protein pairs in the group (× 100).

descriptively, either to assess the protein–protein interaction data,[8,9] or to evaluate the usefulness of extracting knowledge from interacting proteins for annotations of protein function.[17] Unlike these previous studies, here we utilize the cellular localization and cellular-role properties to provide a sound quantitative estimate of the extent of TP in an experimental data set of pairs of interacting proteins.

The following formulation is developed for co-localization, but it applies to the measure of shared cellular-role as well:

$$D = \mathrm{TP} \times I + (1 - \mathrm{TP}) \times R \qquad (1)$$

where TP is the fraction of true interacting pairs in a data set of experimentally determined interacting proteins, D is the fraction of pairs with co-localized pair-mates in this data set, R is the fraction of pairs with co-localized pair-mates in a random set of protein pairs, and I is the fraction of pairs with co-localized pair-mates in true interacting pairs. (We include I in the formulation and do not set it necessarily to 1 because categories of localization may include boundaries where interactions can take place, and because of incompleteness in the annotations. This issue is elaborated below.) The rate of the TP in each data set of experimentally determined interacting protein pairs can be computed if D, R and I are given:

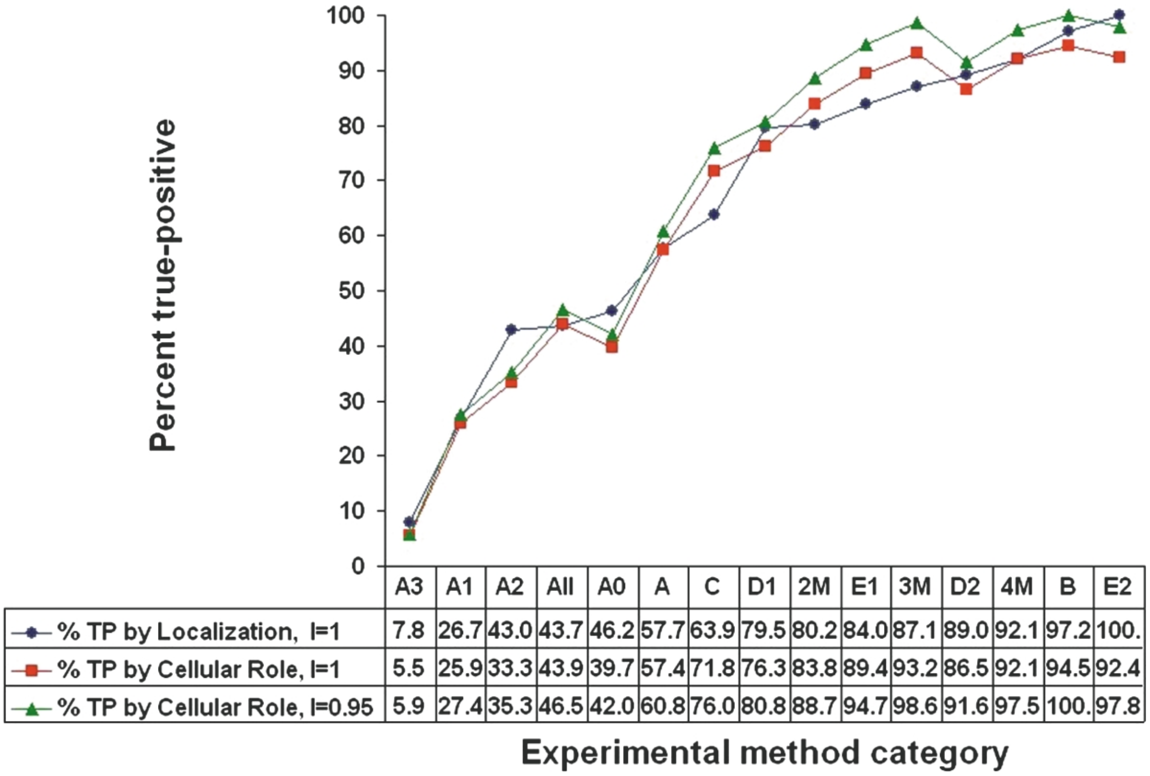
$$\mathrm{TP} = (D - R)/(I - R) \qquad (2)$$

The formulation in equation (1) asserts that the rate of co-localization of a data set (D) is composed of a co-localized fraction of the true interactions ($\mathrm{TP} \times I$), and a co-localized fraction of the false positive interactions $(1 - \mathrm{TP}) \times R$. The latter component is derived on the assumption that the false positive interactions in the data set (that are actually non-interacting proteins) show a fraction of co-localized pairs that is similar to that of random pairs. This assumption is correct if the selection process of the protein pairs in the assessed data set is not biased towards co-localization of pair-mates. For biased data sets the fraction of co-localized pairs among the false positives cannot be based on that of random pairs, and it is R' such that typically $R' > R$, and $\mathrm{TP}' = (D - R')/(I - R')$. By expressing TP and TP′ by formula (2) it follows that $\mathrm{TP}' - \mathrm{TP} = (R - R')(I - D)/(I - R)(I - R')$. Since $R' > R$ and $I > D, R, R'$ this difference is always negative, implying that TP decreases as R increases. This implies that for such biased data sets formula (2) provides an upper bound to the value of TP. Since the data sets of interacting proteins in our study were not derived in a manner that is biased in its rate of co-localization, the values computed here provide exact estimates of the TP rates.

The cellular localization was annotated on the basis of the YPD database[18]† and the data of Kumar *et al.*,[19]‡ and the cellular-role was based on

† http://www.incyte.com/sequence/proteome/

‡ http://ygac.med.yale.edu/ygac-cgi/front_page_OE.html



	A3	A1	A2	All	A0	A	C	D1	2M	E1	3M	D2	4M	B	E2
◆ % TP by Localization, I=1	7.8	26.7	43.0	43.7	46.2	57.7	63.9	79.5	80.2	84.0	87.1	89.0	92.1	97.2	100.
■ % TP by Cellular Role, I=1	5.5	25.9	33.3	43.9	39.7	57.4	71.8	76.3	83.8	89.4	93.2	86.5	92.1	94.5	92.4
▲ % TP by Cellular Role, I=0.95	5.9	27.4	35.3	46.5	42.0	60.8	76.0	80.8	88.7	94.7	98.6	91.6	97.5	100.	97.8

Figure 1. True positive rates in various data sets that are distinguished by the experimental method for determining protein–protein interaction. Blue, percentage of TP based on co-localization ($I = 1$); red, green, percentage of TP based on shared cellular-role for $I = 1$ (red) and $I = 0.95$ (green). I is the fraction of pairs with co-localized pair-mates in true interacting pairs (see the text). For the method categories see the legend to Table 1.

YPD. When proteins were annotated as localized to more than one cellular compartment (or as having more than one role), we considered two pair-mates to be co-localized (or to share a cellular-role), if at least one of the localizations (roles) was shared. By the above annotation, D could be automatically computed for the different method categories (see Table 1). R was computed from all possible pairs of proteins in our data set, and found to be ~0.36 for co-localization and ~0.1 for shared cellular-role, significantly lower than the values listed in Table 1 for the interacting protein pairs. As for the value of I, since we do not know the actual set of true interacting protein pairs we can take one of two approaches in order to obtain a value for I. (i) We could take a large enough set of interacting proteins where we are very confident there is a little noise and assess I in that set. (ii) We can try a range of values for I. However, the values of I can fluctuate in a very narrow range: it cannot exceed 1, and setting it at some distance below 1 yields for some data sets estimates of TP that are above the value of 1. Therefore, for practical purposes, the value of I regarding co-localization must be set to 1, and for shared cellular-role it can fluctuate between 0.95 and 1.

Figure 1 shows the TP rates obtained by the above computation for each method category. It is remarkable that the computations by the different measures yielded very close estimations for the true positive rates. This consistency cannot be accounted for by the partial dependence between the co-localization and shared cellular-role (reflected in the Pearson correlation coefficient $r = 0.12$, and in the Kendall rank correlation coefficient $\tau = 0.20$). Moreover, we repeated the analysis using a different property of the pairs of proteins, based on the expression profiles of the genes encoding them as clustered by Ihmels *et al.*[20] and obtained TP rates that were highly correlated with those presented in Figure 1 ($r = 0.95$, data not shown).

Notably, the rates of TP for all physical, biochemical and immunological methods are above 80% and reach even 100% (for a small set of interactions determined by the immunological methods). The genetic interactions and those determined by "small-scale" Y2H show a medium accuracy, with ~60–70% TP. The more filtered high-throughput Y2H assays show true positive rates of about 50%, consistent with the estimations by Mrowka *et al.*[6] and Ito *et al.*[14] based on different considerations. Our assessment strongly supports the intuitive conjecture, confirmed also by Deane *et al.*[9] and von Mering *et al.*,[8] that interactions which were revealed by more than one method are the most reliable, and the more methods revealing an interaction the higher its reliability (see Figure 1). The estimated TP values that were calculated by us for the 2M and 3M groups are very

Table 2. Estimation of the yeast interactome

Genome-scale data set	Interactions[a] (A)	Revealed[b] (%)(B)	True positives[c] (%) (C)	Interactome estimate[d]
Uetz *et al.*[13] (published data)	956	4.7	~50	~10,000
Ito *et al.*[14] (core data)	798	2.4	~50	~16,600

[a] Number of interactions in the genome-wide Y2H data sets (A0 and A2, from Table 1).
[b] Percentage of the more reliable interactions that are also revealed by the genome-wide Y2H.
[c] The rate of TP in the genome-wide data sets. The TP values for both data sets were close to 50% (Figure 1); in order to avoid small possible biases because of the incompleteness of the localization and role annotations we used the approximate value of 50% for both data sets.
[d] The estimated number of interactions in the yeast cell is obtained by $A/B \times C$.

similar to the estimated values described by Deane *et al.*[9] who have used a similar measure based on gene expression profiles.

The ability to compute a quantitative measure for the true positive rate in an experimental data set of interacting proteins has significant implications for biological inference that relies on such data. For example, we found that interactions that involve one protein that interacts with many other proteins contain only a small fraction of TP. This result supports the suspicion that some of the interactions detected by high-throughput Y2H assays involve "sticky" proteins, and are not necessarily correct.[5] We can also use this computation to more accurately estimate the size of the yeast interactome, the total number of protein–protein interactions in yeast. This intriguing question has addressed considerable attention, as it provides a general estimate for the complexity of the cellular networks in the yeast *S. cerevisiae*. Previous publications provided different estimates for the number of interactions in the yeast cell, ranging from at least 30,000 interactions[8] down to 12,000–20,000 interactions.[21,22] As described below, our computation supports the lower estimate.

In the last few years there have been two attempts to reveal all interactions in yeast using the genome-wide Y2H method (data sets A0[13] and A2[14] in Table 1). The fact that these data sets are genome-wide makes it possible to use them as the basis for estimating the interactome size. Had we known the fraction of true interactions that were revealed by the large-scale methods, we could use this information to compute the size of the interactome. Since we have shown that the biochemical, physical and immunological methods are most reliable, we computed the fraction of those interactions that were revealed also by the large-scale Y2H methods. This provided us with an estimate of the fraction of the interactome revealed by the large-scale methods, as detailed in Table 2. If all interactions in the Y2H data were true, we could obtain an estimate for the interactome size by computing A/B in Table 2. However, since we have shown that the TP rate among these Y2H data is about 50%, this estimate must be multiplied by a factor of 0.5. Thus, the yeast interactome is estimated to range from ~10,000 to ~16,600 interactions (Table 2), consistent with the estimates by Tucker *et al.*[21] and Legrain *et al.*[22]

Acknowledgements

We thank Itay Onn, Nir Friedman and Norman Grover for useful discussions. Joel Sussman is thanked for providing computing facilities to E.S. while in Rehovot. This study was supported by the Israeli Ministry of Science, The Israel Science Foundation (administered by the Israeli Academy of Sciences), and The European Commission (EC grant QLRT-CT-2001-00015-TEMBLOR). E.S. is supported by the Hurvitz Foundation.

References

1. Fellenberg, M., Albermann, K., Zollner, A., Mewes, H. W. & Hani, J. (2000). Integrative analysis of protein interaction data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 152–161.
2. Schwikowski, B., Uetz, P. & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnol.* **18**, 1257–1261.
3. Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* **311**, 681–692.
4. Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29**, 482–486.
5. Legrain, P., Jestin, J. L. & Schachter, V. (2000). From the analysis of protein complexes to proteome-wide linkage maps. *Curr. Opin. Biotechnol.* **11**, 402–407.
6. Mrowka, R., Patzak, A. & Herzel, H. (2001). Is there a bias in proteome research? *Genome Res.* **11**, 1971–1973.
7. Saito, R., Suzuki, H. & Hayashizaki, Y. (2002). Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucl. Acids Res.* **30**, 1163–1168.
8. Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. *et al.* (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
9. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.
10. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M. *et al.* (2002). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **30**, 31–34.
11. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M. *et al.* (2001). DIP:

the database of interacting proteins: 2001 update. *Nucl. Acids Res.* **29**, 239–241.

12. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001). BIND—the biomolecular interaction network database. *Nucl. Acids Res.* **29**, 242–245.
13. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al.* (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
14. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
15. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
16. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L. *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
17. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. & Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.
18. Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P., Kranz, J. E., Mangan, M. *et al.* (2002). Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol.* **350**, 347–373.
19. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S. *et al.* (2002). Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719.
20. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature Genet.* **31**, 370–377.
21. Tucker, C. L., Gera, J. F. & Uetz, P. (2001). Towards an understanding of complex protein networks. *Trends Cell Biol.* **11**, 102–106.
22. Legrain, P., Wojcik, J. & Gauthier, J. M. (2001). Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.* **17**, 346–352.

Edited by G. von Heijne

(Received 22 October 2002; received in revised form 20 December 2002; accepted 30 January 2003)