Protein Interactions

TWO METHODS FOR ASSESSMENT OF THE RELIABILITY OF HIGH THROUGHPUT OBSERVATIONS*

Charlotte M. Deane‡§, Łukasz Salwiński‡, Ioannis Xenarios, and David Eisenberg¶

High throughput methods for detecting protein interactions require assessment of their accuracy. We present two forms of computational assessment. The first method is the expression profile reliability (EPR) index. The EPR index estimates the biologically relevant fraction of protein interactions detected in a high throughput screen. It does so by comparing the RNA expression profiles for the proteins whose interactions are found in the screen with expression profiles for known interacting and non-interacting pairs of proteins. The second form of assessment is the paralogous verification method (PVM). This method judges an interaction likely if the putatively interacting pair has paralogs that also interact. In contrast to the EPR index, which evaluates datasets of interactions, PVM scores individual interactions. On a test set, PVM identifies correctly 40% of true interactions with a false positive rate of ~1%. EPR and PVM were applied to the Database of Interacting Proteins (DIP), a large and diverse collection of protein-protein interactions that contains over 8000 *Saccharomyces cerevisiae* pairwise protein interactions. Using these two methods, we estimate that ~50% of them are reliable, and with the aid of PVM we identify confidently 3003 of them. Web servers for both the PVM and EPR methods are available on the DIP website (dip.doe-mbi.ucla.edu/Services.cgi). *Molecular & Cellular Proteomics 1:349–356, 2002.*

One thrust of post-genomic biology is the study of the networks of protein interactions that control the lives of cells and organisms. These networks have been reconstructed by detecting pairwise interactions of proteins. To store and manage this information in a systematic way, databases have been created (1, 2). These databases provide centralized access to curated experimental data. They have also emerged as resources for the investigation of the large scale properties of biological networks, in particular their functional and evolutionary aspects (3).

In this paper we explore the usefulness of the Database of Interacting Proteins (DIP)[1] for assessing the reliability of measurement of protein interaction. Until two years ago, when high throughput screens of protein interaction were developed, the information within interaction databases was collected from the small scale screens in hundreds of individual research papers. The biological relevance of each interaction had often been investigated thoroughly, sometimes with a repertoire of experimental techniques and often with multiple controls (4, 5). These independent, often repeated observations, coupled with controls and curation in the peer-review process, enhanced the reliability of the published data. In the past two years, high throughput, genome-wide detections of protein interactions by yeast two hybrid (Y2H) and mass spectrometric analysis of protein complexes have increased tremendously the experimental coverage. The new methods can generate rapidly more information than was collected by traditional means in more than a decade (6–10). However, the large size of such datasets makes it impractical to verify individual interactions by the same methods used previously in small scale experiments (11, 12). The question then arises, Do these new, high throughput methods of detecting interactions provide information as reliable as the small scale experiments? Verifying the interactions from these high throughput methods is vital (11–15), because only then can the large and small scale data be combined into one self-consistent interaction network useful for further studies.

To address these issues we have analyzed the complete set of 8063 protein-protein interactions identified in yeast, *Saccharomyces cerevisiae*, that are described in DIP as of November 2001. We demonstrate that the subset of interactions obtained through the high throughput Y2H screens differs in several respects from the subset based only on the small scale or multiple, redundant experiments. Most notably, analysis of the coexpression profiles of the interacting partners leads to the conclusion that, overall, only about 30% of the high throughput dataset possesses the same characteristic mRNA expression features as the dataset based on the small scale experiments. To further pinpoint the interactions within the dataset that are likely to be correct, interactions were analyzed between protein pairs that are paralogs of the tested proteins. This resulted in the identification of ~1400 interactions likely to be correct. A reliable, self-consistent set of interactions totaling ~3000 is extracted when these ~1400

From the Howard Hughes Medical Institute, Molecular Biology Institute, UCLA-Department of Energy Laboratory of Structural Biology and Molecular Medicine, University of California, Los Angeles, California 90095-1570

Received, December 26, 2001, and in revised form, March 20, 2002
Published, MCP Papers in Press, April 4, 2002, DOI 10.1074/mcp.M100037-MCP200

[1] The abbreviations used are: DIP, Database of Interacting Proteins; EPR, expression profile reliability; IST, interaction sequence tag; PVM, paralogous verification method; Y2H, yeast 2 hybrid; GY2H, genome-wide Y2H; YPD, Yeast Protein Database.

are combined with the small experiment datasets and with interactions verified by more than one experiment.

EXPERIMENTAL PROCEDURES

Interaction Datasets—The protein-protein interaction datasets analyzed in this work are listed in Table I. They are all, except for the RND sets, subsets of the *S. cerevisiae* protein-protein interaction network (DIP-YEAST; 8063 distinct interactions) extracted from the DIP database on November 19, 2001. The INT set contains all the interactions determined by one or more small scale experiment (defined as an experiment described in a published article listing no more than 100 distinct protein-protein interactions) whereas sets EC2 and EC3 contain interactions determined by, respectively, at least two or three independent experiments. The GY2H set contains all the interactions reported in high throughput protein-protein interaction screens (6–8, 16, 17), and GY2H′ is a subset of GY2H that excludes interactions occurring only in the ITO1 set. The ITO1, ITO2, . . . ITO8 are subsets of GY2H that contain all the interactions reported by Ito *et al.* (7) as identified by at least 1,2, . . . 8 interaction sequence tags (ISTs) in a genome-wide Y2H protein-protein interaction screen. These datasets (ITO1, etc.) contain fewer interactions than the numbers reported in the original paper because of some redundancy of the original dataset (interactions were reported in both directions, P-P′ and P′-P). Also, some of the open reading frames could not be traced unambiguously to a unique SWISS-PROT, PIR (Protein Information Resource), or GenBank™ entry.

The RND1–3 sets were generated by randomly selecting 100,000 protein-protein pairs from the yeast genome that are not present in DIP. They are dominated by the non-interacting pairs (less than 0.15% of the true interactions present, assuming ~10 interacting partners per protein) even when overestimating by a factor of two the average number of interacting partners for each protein within the *S. cerevisiae* genome predicted in the recent literature (14, 18).

Functional Correlation—Proteins have been assigned to 44 "cellular role," 58 "functional," and 29 "compartment" categories in the Yeast Protein Database (YPD) (19, 20). Cellular role is defined as the major biological process involving the protein and function as the principal structural, regulatory, or enzymatic function of the protein. The YPD categories are broad, and a large percentage of proteins are associated with more than one cellular role, function, or compartment (subcellular location).

The functional annotation, cellular role, and compartment, if one exists, were collected for all the *S. cerevisiae* open reading frames from the YPD database. We counted a correlation if the two interacting proteins shared one or more annotated function in a manner analogous to Schwikowski *et al.* (15). The background probability that one could expect two proteins to share a common function was calculated using all possible pairs of proteins annotated in a given category.

Expression Profile Reliability Index—The expression profile reliability index (EPR) was extracted from the interaction datasets by solving the equality-constrained linear least squares problem defined by Equation 2 (see "Results") using LAPACK implementation of the GRQ factorization method (21) and a discrete representation of the $\rho(d^2)$ distributions (up to 30 bins, 1.25 units wide; only bins with at least five counts were included in the calculations). χ^2 was calculated assuming binomial distribution of the error for the individual bins in each of the histograms. The accuracy of the fitted parameter was estimated using a bootstrapping approach with 5,000 synthetic datasets as described (22).

The Euclidean expression distance between proteins A and B, d_{AB}, was calculated according to Equation 1,

$$d_{AB}^2 = \sum_i (\log(e_i^A/e_{ref}^A) - \log(e_i^B/e_{ref}^B))^2 \qquad \text{(Eq. 1)}$$

where e_i^N is a log ratio of the expression level of protein *N* under the *i*th conditions as reported customarily by Brown and co-workers (23). The sum is performed over a set of 12 distinct shock conditions using the data provided by Gasch *et al.* (23).

Paralogous Verification Method—The paralogous verification method (PVM) validates interacting pairs using the existence of paralogous interactions. Paralogs were collected by performing intra-proteome comparisons using PSI-BLAST (24). Each predicted open reading frame product of *S. cerevisiae* served as a query sequence against the entire database of *S. cerevisiae*. The PSI-BLAST comparisons were performed using the BLOSUM62 substitution matrix and the seg filter to mask compositionally biased regions in the query sequence. To arrive at the optimal definition of family, different PSI-BLAST conditions were examined, and the coverage and sensitivity were measured.

RESULTS

Yeast Interactions in DIP—The set of known protein-protein interactions in budding yeast (*S. cerevisiae*), as documented in DIP on November 2001, contains ~8000 distinct interactions between 4150 proteins (Table I). Approximately 2000 of these interactions were detected by small scale experiments described in more than 800 research articles. The remainder (~6000) is derived from four independent high throughput Y2H screens (Fig. 1). Comparison of the datasets shows that the overlap of detected interactions obtained in the four studies, as well as between any of these datasets and the set derived from the small scale interaction screens, is petite. This observation, made already by others (12, 14, 15, 25), is the motivation of the present work.

There are many possible reasons for the lack of overlap. Those include the use of different yeast strains, differences in the quantitative measures of interaction, and the use of non-physiological conditions in experiments. Additionally, high throughput protein-protein interaction screens, such as those utilizing Y2H methods, increase the chance of identifying artifactual partners by testing exhaustively arbitrary protein-protein interactions. Those include the partners that can physically interact but that are never in close proximity to one another in the cell because of distinct subcellular localization or expression at different times during the life cycle.

All these factors can lead to the observation of either false negatives (interactions that cannot be detected under the conditions used) or false positives (physical interactions without biological meaning). Here we concentrate on the following two problems: 1) identifying the fraction of false positives within the high throughput datasets (using EPR) and 2) identifying true positives (using PVM). We do this by relating the global properties of these datasets with those of the reference set of biologically relevant interactions extracted from the DIP database. The underlying assumption of this approach is that, by the virtue of its size and diversity, this reference dataset (INT) captures the most prominent features of biologically relevant protein-protein interactions and therefore can be used to judge the quality of other interaction datasets.

Functional Correlation—We began by asking what level of functional resemblance we can find between two interacting *S. cerevisiae* proteins in DIP. For this study, we divided the

TABLE I
Protein-protein interaction datasets

DIP-YEAST is the entire *S. cerevisiae* protein-protein interaction network extracted from the DIP database on November 19, 2001. INT contains all the interactions determined by at least one small scale experiment, whereas sets EC2 and EC3 contain interactions determined by at least two or three independent experiments, respectively. GY2H contains all the interactions reported in genome-wide protein-protein interaction screens, and GY2H′ is a subset of GY2H that excludes interactions occurring only in the ITO1 set. The ITO1, ITO2, . . . ITO8 are subsets of GY2H that contain all the interactions reported by Ito *et al.* (7, 16) as identified by at least 1, 2, . . . 8 ISTs. RND1–3 sets were generated by randomly selecting 100,000 protein-protein pairs that are not present in DIP. PVM is the subset of DIP-YEAST scored by the PVM method. CORE is the amalgamation of the INT, E2, and PVM subsets. DIP subsets. See "Experimental Procedures" for further details. SS, small scale experiments; HT, high throughput screens.

Dataset	SS	HT	Number of interactions			
			All	Expression data available[a]	PVM subset[b]	Interactions with paralogs
DIP-YEAST			8063	7225	1428	6083
INT	√		2246	1806	913	1857
EC2	√		1179	976	448	910
EC3	√		377	322	158	300
GY2H		√	6114	5494	648	4474
GY2H′		√	3244	2882	527	2436
PVM			1428	1172	1428	1428
CORE			3003	2435	1428	2480
RND1			100000	87368	1062	99980
RND2			100000	87324	1060	99975
RND3			100000	87330	1060	99983
ITO1		√	4337	3935	365	3083
ITO2		√	1454	1289	236	1033
ITO3		√	795	696	173	550
ITO4		√	571	493	131	388
ITO5		√	494	404	112	314
ITO6		√	462	324	90	246
ITO7		√	371	268	75	202
ITO8		√	307	233	68	173
ITO9		√	270	198	60	150

[a] Data reported by Brown and co-workers (23).
[b] Those interactions selected from each of the datasets by PVM as correct.

interacting pairs into four datasets; DIP-YEAST includes all pairs, EC3 and EC2 are datasets with greater than or equal to three or two observations supporting the interaction, respectively, and INT is the set of interactions observed in at least one small scale experiment. A full description of the subsets is given under "Experimental Procedures."

Fig. 2 shows the percentage agreement of function, cellular role, and compartment as defined by the YPD (19, 20) for the pairs. The *horizontal black line* gives the background percentage agreement. It shows that if we pick two proteins at random from the set with known functions, the members of ~18% of pairs agree in function. The difference between the observed agreement and this background is large in all cases.

These results can be compared with those of other investigators. We find that ~66% of the DIP-YEAST pairs share one or more annotated compartments compared with the 78% found by Fields and co-workers (15) in a global analysis of 2,709 published interactions of *S. cerevisiae* proteins. Correlation of function was also tested in a different way by Vidal and co-workers (27) who examined whether interacting pairs were found within the same gene expression cluster. These gene expression clusters are generally believed to correspond to functional categories (27–30). As with the results here, correlation of the functional categories based on the gene

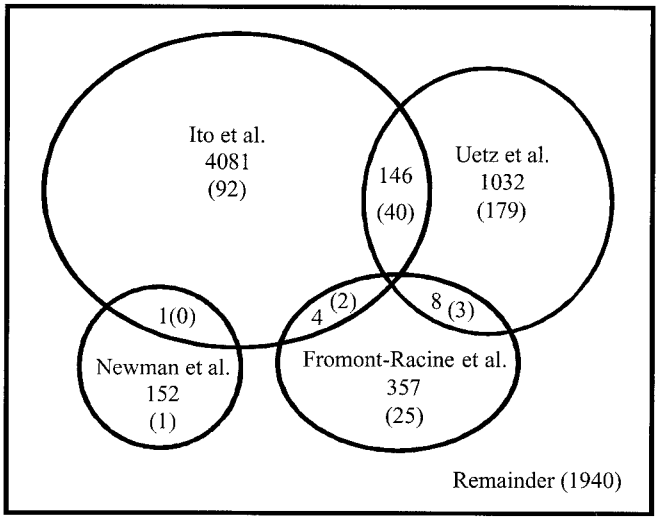

FIG. 1. **Interacting yeast proteins as detected in several studies.** A Venn diagram illustrates the overlap between the datasets in YEAST-DIP. Each *oval* represents a high throughput Y2H study, and the overlaps between the Y2H studies are given at the intersections. The *number* in *parentheses* represents those interactions that have been determined by small scale methods (see "Experimental Procedures" for more details). Thus, the *numbers* within *parentheses* represent the INT set. Notice the small overlap among the datasets.

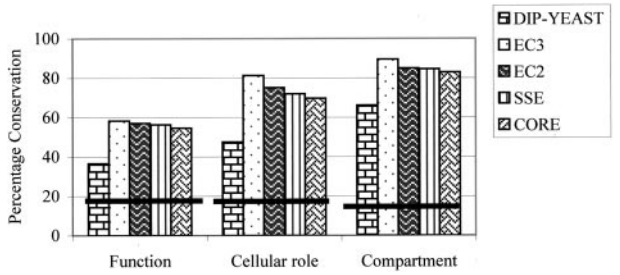

FIG. 2. **The differences for five datasets of pairs of putatively interacting proteins in percentage correlation of function, cellular role, and compartment of the two members of each pair.** The datasets tested are DIP-YEAST and the subsets INT, EC2, EC3, and CORE (see "Experimental Procedures" for details). All datasets show correlation well above the background with the strongest correlation seen for compartment. The subset of DIP-YEAST interactions believed to be correct (CORE) shows a pattern of correlation that is higher than the entire DIP-YEAST dataset and appears closer to that of the INT, E2, and E3 datasets, which are believed to contain exclusively biologically relevant interactions.

expression clusters was higher than random but still relatively low.

Notice in Fig. 2 that the INT set and the EC2 and EC3 sets show substantially higher correlation than the DIP-YEAST set. The relative lack of agreement of compartment within the DIP-YEAST data (63%) could be, in part, because of the large number of interactions between nuclear and cytoplasmic proteins (15); these are expected as there are many reports of proteins shuttled between these compartments through the nuclear pore (31). The INT dataset may show higher correlation because of a better relationship between functional annotation and protein interactions described in the small scale studies. However, if we select random pairs of proteins from INT, as opposed to the entire set, a similar level of random correlation is observed. This points to a similar level of multiple annotation and possible cross-talk in both cases.

It should also be remembered that the annotations in these categories may have been transferred from homologous proteins without experimental confirmation and as such are subject to error. However, when we calculate the percentage correlations for the set of experimentally annotated proteins calculated they are similar to the results described above.

Function (the principal structural, regulatory, or enzymatic function) is the least conserved of the three properties. This is not surprising, as an interaction between two proteins does not demand that they share an identical function; rather it demands that they are linked in a functional network. Thus, the linkages observed between functional groups could well be biologically meaningful. For example, Schwikowski *et al.* (15) found that there are a large number of interactions between the categories of protein folding and protein translocation. Therefore, in the assessment of an individual interaction, identical assignments of function or cellular role should not always be expected; rather consideration should be given to the relationships between the functions of the proteins.

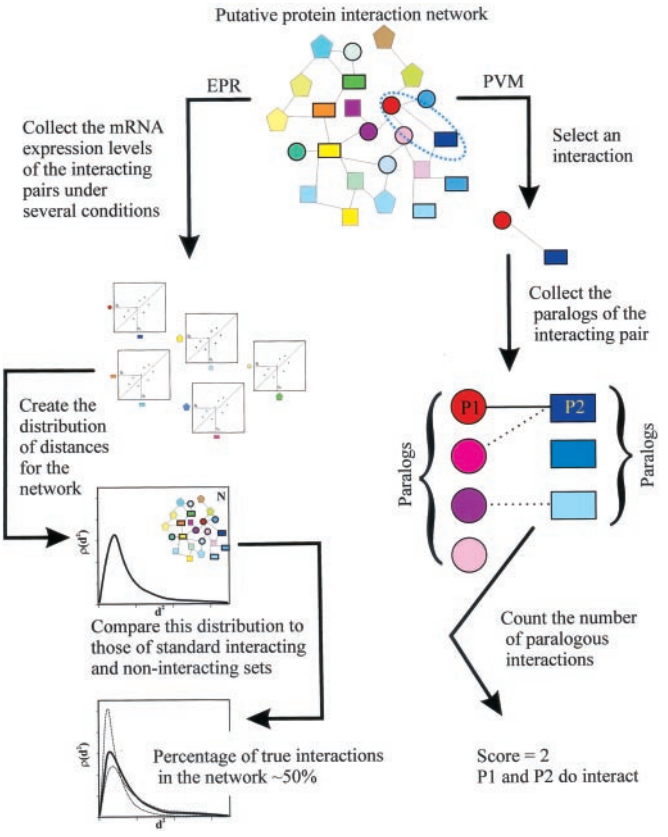

FIG. 3. **A flow chart of the EPR and PVM methods used to test the reliability of complete interaction datasets and individual interactions, respectively.** The EPR index calculation was as follows: the similarity of the expression patterns of two proteins is evaluated by calculating expression distance *d*. This distance is calculated for all pairs of proteins within the set of interest. The distribution of these distances is interpreted as a linear combination of the probability distributions of standard interacting and non-interacting sets resulting in the expression profile reliability index, α_{EPR} (see Equation 2). The PVM procedure was as follows: if two proteins, P1 and P2, are considered to interact the paralogous families of P1 and P2 are collected. The number of interactions between these families within the DIP database is counted, excluding the P1 to P2 link. This count is the score of the interaction. In this case the link P1 to P2 scores 2. If this score is greater than zero the interaction is predicted to be true.

The poorer conservation of function, compartment, and cellular role within the DIP-YEAST dataset than the INT, EC2, and EC3 datasets suggests that small scale studies yield more reliable results than high throughput studies; this calls for methodologies, which determine the reliability of a dataset and the reliability of any given interaction. Here we introduce two computational methods that use mRNA expression data and sequence analysis, respectively, to assess reliability of the high throughput datasets and to identify protein-protein interactions that are likely to be correct. An overview of the two methods is offered in Fig. 3.

mRNA Expression Profiles of Interacting Pairs: the EPR Method—It has been demonstrated numerous times that functionally related genes tend to be expressed in a concerted

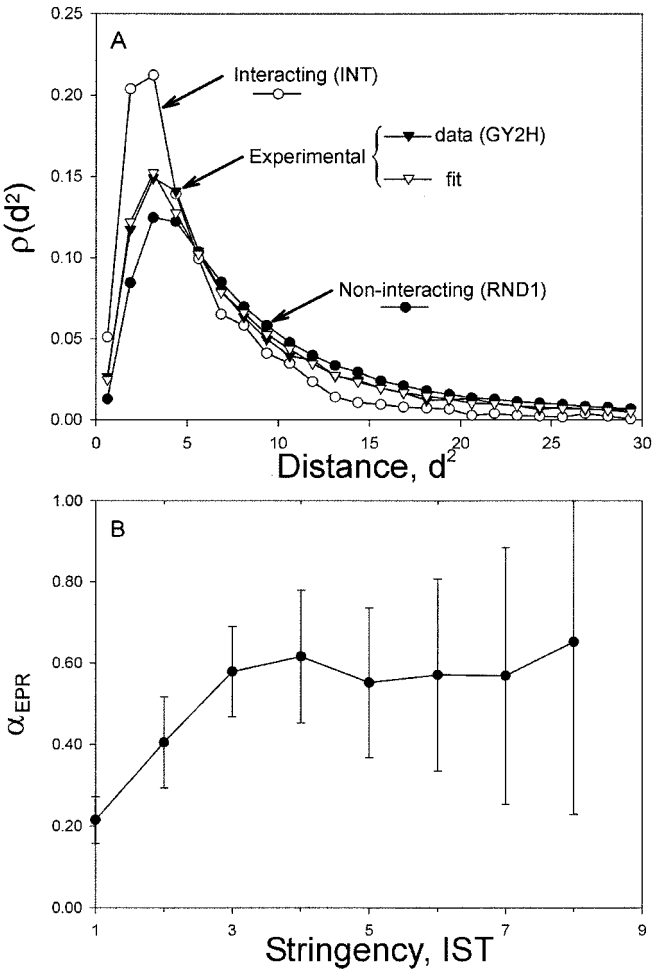

FIG. 4. **Evaluation of genome-wide Y2H interaction data in DIP using the EPR index.** *A*, distribution of the expression distances for the INT (*open circles*), RND1 (*closed circles*), and GY2H (*closed triangles*) datasets. The curve fitted to the GY2H distribution using Equation 1 is marked with *open triangles* ($\alpha_{EPR} = 0.315$). The INT distribution represents interacting proteins, and RND1 represents non-interacting proteins. Notice that all the curves are normalized to a unit area. *B*, the dependence of the expression profile reliability index (α_{EPR}) calculated for the subsets of the genome-wide yeast two hybrid data from Ref. 7 on the stringency of the selection procedure as reflected by the number of IST observed. Notice that α_{EPR} tends to increase with higher stringency of selection of interacting proteins. Notice also that the uncertainty of α_{EPR} grows with higher stringency, because there are fewer interactions.

fashion (27–29). Here, we utilize this observation to assess the quality of datasets of interacting proteins. Specifically, we define a distance measure d^2 between the expression levels of the mRNAs encoding for the members of an interacting pair (see equations under "Experimental Procedures"). Then we characterize a dataset of protein interactions by plotting the fraction of pairs having each value of d^2. This is the basis of the EPR index method illustrated on the *left* of Fig. 3.

Fig. 4*A* shows the normalized distribution of expression level distances (d^2) for several sets of protein interaction data. The curve RND1 gives the distribution for randomly generated sets of protein pairs. Notice that it is the broadest distribution shown, with the lowest peak. The curve INT is for the small scale dataset and is seen to have the highest peak and sharpest distribution. Those differences are statistically significant (confidence level $p = 10^{-140}$), as inferred from a Kolmogorov-Smirnof test. We take the INT set to be a reference set of interacting proteins and the RND1 set to be representative of non-interacting proteins.

On the basis of the $\rho(d^2)$ distribution curves, we define a parameter, α_{EPR}, that characterizes the expected accuracy of a dataset of protein interactions. To do so we notice that the expression-distance profile of the GY2H set appears to be intermediate between the reference interacting (INT) and non-interacting (RND1) sets. The simplest model explaining this behavior assumes that the Y2H experiments result in two types of protein-protein pairs, the true positive (biologically relevant interactions) pairs, drawn randomly from the interacting population, and false positives, drawn randomly from the non-interacting population. The resulting, overall distribution of expression distances obtained for an experimental set, ρ_{exp}, is then described by Equation 2,

$$\rho_{exp}(d^2_{AB}) = \alpha_{EPR} \cdot \rho_i(d^2_{AB}) + (1 - \alpha_{EPR}) \cdot \rho_n(d^2_{AB}) \quad \text{(Eq. 2)}$$

where ρ_i and ρ_n are the expression distance probability distributions for the interacting and non-interacting protein pairs, and the expression profile reliability index, α_{EPR}, corresponds to the fraction of the true positives in the experimental dataset.

The ρ_n distribution can be obtained as the distribution of expression distances for all protein pairs within a genome, because the full genome distribution is of vast size ($\sim 9 \cdot 10^6$ for *S. cerevisiae*) and must be dominated by the non-interacting pairs. The ρ_i can be approximated by the distribution of the expression distances for all the reliable interactions present in DIP-YEAST (for example INT). The latter assumption seems to be valid as the set of interactions described in DIP-YEAST is in the majority of cases obtained in a manner that did not rely on the expression levels of the interacting partners. Therefore, it can be treated as a representative sample of the entire protein-protein interaction set, random with respect to the expression levels of the interacting proteins.

A linear least-squares fit of the GY2H dataset to the model described by Equation 1 allows us to evaluate the α_{EPR} parameter. The α_{EPR} is calculated as 31 ± 3% (Table II) for the GY2H data, suggesting that ~70% of the reported pairs in this set are, in fact, false positives. To verify that α_{EPR} indeed reflects the expected accuracy of the experimental results, subsets of the GY2H corresponding to varying stringency of selection were constructed as reported by Ito *et al.* (7). Ito *et al.* (7) created these sets by identifying those interactions with at least 1, 2, . . . 8 ISTs, labeled here as ITO1 to ITO8, respectively. As expected, the accuracy of the resulting subsets, as evaluated by α_{EPR}, increases with increased selection

TABLE II
EPR index

EPR index, α_{EPR}, calculated for several subsets of DIP-YEAST (see "Results" and "Experimental Procedures" for details) using INT and RND1 subsets as representative for the interacting and noninteracting protein populations, respectively, is shown. The values of χ^2 and *N*, the number of degrees of freedom, are given.

Dataset	α_{EPR}	χ^2	*N*
DIP-YEAST	0.48 ± 0.03	9.07	29
EC2	0.85 ± 0.06	1.65	16
EC3	0.88 ± 0.17	3.05	10
GY2H	0.31 ± 0.04	14.84	29
GY2H′	0.50 ± 0.03	14.09	29
PVM	0.78 ± 0.13	5.85	16
CORE	0.92 ± 0.03	1.69	19
ITO1	0.22 ± 0.06	19.4	29
ITO2	0.41 ± 0.11	12.6	19
ITO3	0.58 ± 0.11	10.1	16
ITO4	0.62 ± 0.16	9.5	14
ITO5	0.55 ± 0.18	8.8	14
ITO6	0.57 ± 0.24	7.1	12
ITO7	0.57 ± 0.32	6.0	10
ITO8	0.65 ± 0.42	4.6	7

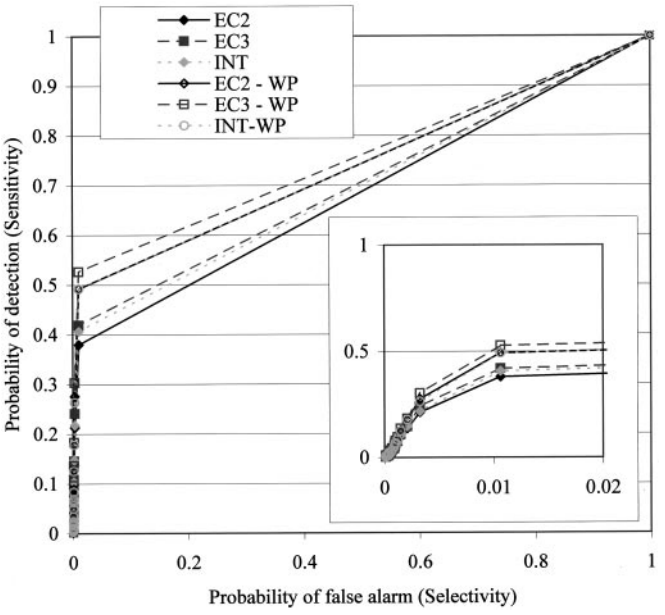

FIG. 5. **A receiver operator characteristic curve showing that the PVM technique has high selectivity (the probability of detecting a false interaction is very low) but relatively low sensitivity (only around 50% of the correct interactions are identified).** EC2 and EC3 refer to datasets with greater than two or greater than three pieces of experimental evidence supporting them, respectively. The INT set is all interactions that are established by any technique other than genome-wide yeast two hybrid experiments. The WP subsets of these are where at least one of the pair of proteins involved in a given interaction has greater than one paralog, making a score greater than zero and thus detection of the interaction possible. The *inset graph* shows a magnification of the low probability of false alarm region of the main graph. At an error rate of ~1%, ~40–50% of correct interactions are detected. Notice that the WP subsets all show around 10% higher sensitivity. This means that even a single addition pair of interacting proteins with paralogs increases significantly the sensitivity of the PVM method.

stringency (Fig. 4*B*). This indicates that the EPR index can be used to characterize the accuracy of experimental, large scale protein-protein interaction datasets, and corresponds crudely to the fraction of pairs that is meaningful biologically. However, the error on α_{EPR} increases rapidly with decreasing dataset size, therefore limiting the applicability of EPR in general to large (>500 interaction) datasets.

The error-prone high throughput Y2H screens can be filtered by excluding the least reliable protein pairs that occur only in the ITO1 set. In fact, the overall reliability of the resulting GY2H′ set increases to roughly 50%, as judged by the EPR index (Table II). However, this improved reliability comes at the price of reducing its size by nearly one-half.

Using Paralogous Interactions to Verify Protein-Protein Interactions: the PVM Method—The reliability of a given protein interaction can be evaluated by the presence of paralogous interactions. The basis for this is that if two proteins are paralogs then the proteins that they are observed to interact with are often also paralogs. This observation is related to the notion of interologs proposed by Vidal and co-workers (9).

To validate a given interaction between a pair of proteins, P1 and P2, all the paralogs of P1 and P2 are collected, and the number of interactions observed in DIP between these two families, excluding the interaction P1 to P2, are counted (Fig. 3). This count is the PVM score.

To ascertain the ability of this method (PVM) to identify true interactions and ignore false interactions, the behavior on datasets of interacting proteins must be compared with the behavior on datasets of non-interacting proteins. We generated the datasets of non-interacting proteins computationally because of the difficulty in crafting such a set from reports within the literature. The three random sets of protein interactions (RND1, RND2, RND3) described under "Experimental Procedures" were used as the non-interacting sets; although these sets will not be entirely free of interactions, the percentage should be very small (see "Experimental Procedures").

Three sets of protein interactions were used as true interaction sets, the INT, EC2, and EC3 sets (see "Experimental Procedures"). The EC2 and EC3 sets are smaller than the INT set (Table I) and can be used by PVM but are not suitable as reference datasets for EPR, because the uncertainty in α_{EPR} is large for such small datasets (Table II).

The efficacy of the PVM method can be illustrated by a selectivity-sensitivity curve (also known as a receiver-operator characteristic curve) shown in Fig. 5. It shows that a score that selects few (~1%) false positives is sensitive to ~40% of the true interactions. That is, the method shows high specificity but a lower sensitivity. This lack of sensitivity in part reflects the lack of paralogs of some proteins. Such interactions cannot score >0. Thus if the INT, EC3, and EC2 sets are modified to consider only those pairs where at least one of each of the

pairs has more than one paralog (Fig. 5) an improvement in sensitivity of ~10% is observed. The low sensitivity is therefore not caused solely by the lack of paralogs but is perhaps because of both the lack of experimental data and, in a number of cases, a lack of any paralogous interactions.

There is another possible source of error in PVM because of the erroneous identification of interactions in Y2H experiments. For example if P1′ and P2′ are paralogs of P1 and P2, respectively (where P denotes a protein), it is possible that *in vivo* only the P1/P2 and P1′/P2′ interactions take place. However, Y2H may detect interactions between P1 and P2′ and P2 and P1′, as well as the true interactions P1-P2 and P1′-P2′. The calculated error rate of PVM of ~1% suggests that this problem is small. However, as with any computational prediction technique the results should be considered in the light of other data such as sub-cellular localization or function of the proteins.

The receiver-operator characteristic curve also demonstrates that the magnitude of the score is unimportant, merely that a score greater than zero indicates a high probability that an interaction exists. Thus, if a given low reliability interaction (such as Y2H (32)) has paralogs but a score of zero, it can be validated either directly or by testing for a paralogous interaction.

It is clear that PVM can only be used in cases where the proteins involved in the interaction have paralogs. In *S. cerevisiae* 3130 of the 6356 proteins have paralogs (~ 50%). This level of paralogs appears to be typical. Koonin *et al.* (33) found that 46% of the *Escherichia coli* genome has paralogs, and ~⅔ of the proteins within the COG database (34) are found to have paralogs.

DISCUSSION

Uses of EPR and PVM—EPR can assess the overall quality of an interaction dataset but cannot assess the quality of individual interactions. Fig. 4*A* demonstrates that, the similarity in the expression levels of interacting (INT) and non-interacting sets (RND1), as judged by the changes in the mRNA levels, can vary over a large range of d^2 and overlap significantly with one another. Therefore, it is generally not possible to use the similarity of the expression profiles as a predictor of protein-protein interactions without using other sources of information. However the profiles do allow an estimation of the percentage of biologically relevant interactions within a set.

PVM, on the other hand, is able to assess the quality of individual protein-protein interactions. However, it can also estimate the total number of biologically relevant interactions within a dataset. This estimation is based on the observation that in the subsets of EC2, EC3, and INT with paralogs, ~50% of the interactions are identified by PVM (Table I). Thus, PVM should identify ~50% of the biologically relevant interactions within any given dataset. The number of true interactions within a set can, therefore, be estimated as twice the number given by PVM. In the DIP-YEAST set only 1428 of the 6083 interactions that could score did. Thus the expected number of true interactions is around 2800 of the subset with paralogs. This suggests that ~2800 of ~6000 interactions are valid, giving an error rate for the overall DIP-YEAST of around 50%. This compares well with the EPR estimation of ~47% given in Table II.

The ability of PVM to identify roughly half the true interactions within a given dataset means that it can also be used to indicate the quality of a dataset, by means of the percentage of identified interactions. The different Ito *et al.* (7) subsets described under "Results" and "Experimental Procedures" were examined separately using PVM, and it was found that as the number of independent observations of the interactions increased from 1 to 8 the percentage of the dataset identified as correct by PVM increased (Table I) much as the EPR index improves (Fig. 4*B*). The efficacy of PVM can also be demonstrated by examining the EPR of the subset of DIP-YEAST selected by PVM. It demonstrates that this dataset behaves within experimental error like the INT set (Table II).

DIP Yeast Interactions Estimated to be Correct—There are about 5600 interactions within the DIP-YEAST dataset identified solely in the genome-wide Y2H screens. These include roughly 3000 interactions that were reported by Ito *et al.* (7) as based on only single IST. Although these interactions are expected to contain many false positives (26) the results in Tables I and II demonstrate that they still contain a significant proportion of true positives, and the method such as PVM is suited ideally to identify at least some of them.

A subset of the DIP-YEAST interactions believed to be correct can be identified by merging the PVM (1428), INT (2246), and EC2 (1179) sets (Table I); this gives a total of 3003 interactions. This set is denoted as the CORE and is available on the DIP website (dip.doe-mbi.ucla.edu). Four hundred fifty-four of the CORE interactions are identified by PVM alone and as such could not be validated by any other method. Fig. 2 shows that this CORE set of interactions has a correlation of function pattern that is similar to the sets believed to be correct (INT, EC2, and EC3). The gross number of interactions predicted to be correct based on the EPR index of DIP-YEAST is ~4000. Thus though PVM is able to identify putatively correct interactions with very high selectivity it is unable even with the inclusion of INT and EC2 to extract from DIP-YEAST all those interactions, which are estimated to be correct by EPR.

Acknowledgments—We thank Robert Grothe and Parag Mallick for discussions.

* This work was supported in part by National Institutes of Health and the Department of Energy. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Contributed equally to this work.

§ Supported by a Wellcome Trust fellowship.

¶ To whom correspondence should be addressed: Howard Hughes Medical Inst., Molecular Biology Inst., UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, University of California, Los Angeles, P.O. Box 951570, Los Angeles, CA 90095-1570. Tel.: 310-825-3754; Fax: 310-206-3914; E-mail: david@mbi.ucla.edu.

REFERENCES

1. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001) BIND - The biomolecular interaction network database. *Nucleic Acids Res.* **29,** 242–245
2. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* **28,** 289–291
3. Jeong, H., Mason, S. P. Barabási, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411,** 41–42
4. Xenarios, I., and Eisenberg, D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.* **12,** 334–339
5. Golemis, E. A., Ed. (2001) *Protein-Protein Interactions: A Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York
6. Fromont-Racine, M., Mayes, A. E., Brunet-Simon, A., Rain, J. C., Colley, A., Dix, I., Decourty, L., Joly, N., Richard, F., Beggs, J. D., and Legrain, P. (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17,** 95–110
7. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4569–4574
8. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627
9. Walhout, A. J., Boulton, S. J., and Vidal, M. (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **17,** 88–94
10. Newman, J. R., Wolf, E., and Kim, P. S. (2000) From the cover: A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 13203–13208
11. Walhout, A. J., and Vidal, M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24,** 297–306
12. Hazbun, T. R., and Fields, S. (2001) Networking proteins in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4277–4278
13. Legrain, P., Wojcik, J., and Gauthier, J. M. (2001) Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet.* **17,** 346–352
14. Tucker, C. L., Gera, J. F., and Uetz, P. (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.* **11,** 102–106
15. Schwikowski, B., Uetz, P., and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18,** 1257–1261
16. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 1143–1147
17. Fromont-Racine, M., Rain, J. C., and Legrain, P. (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* **16,** 277–282
18. Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287,** 116–122
19. Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., Lengieza, C., Lew-Smith, J. E., Tillberg, M., and Garrels, J. I. (2001) YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ library, an integrated resource for protein information. *Nucleic Acids Res.* **29,** 75–79
20. Costanzo, M. C., Hogan, J. D., Cusick, M. E., Davis, B. P., Fancher, A. M., Hodges, P. E., Kondu, P., Lengieza, C., Lew-Smith, J. E., Lingner, C., Roberg-Perez, K. J., Tillberg, M., Brooks, J. E., and Garrels, J. I. (2000) The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28,** 73–76
21. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999) *LAPACK Users Guide*, 3rd Ed., Society for Industrial and Applied Mathematics, Philadelphia
22. Press, H. P., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992) *Numerical Recipes in FORTRAN*, p. 687, Cambridge University Press, Cambridge
23. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257
24. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
25. Legrain, P., and Selig, L. (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.* **480,** 32–36
26. Vidal, M., and Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.* **27,** 919–929
27. Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29,** 482–486
28. Ferea, T. L., and Brown, P. O. (1999) Observing the living genome. *Curr. Opin. Genet. Dev.* **9,** 715–722
29. Lockhart, D. J., and Winzeler, E. A. (2000) Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836
30. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12,** 37–46
31. Michael, W. M. (2000) Nucleocytoplasmic shuttling signals: two for the price of one. *Trends Cell Biol.* **10,** 46–50
32. Mrowka, R., Patzak, A., and Herzel, H. (2001) Is there a bias in proteome research? *Genome Res.* **11,** 1971–1973
33. Koonin, E. V., Tatusov, R. L., and Rudd, K. E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 11921–11925
34. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29,** 22–28