

# Estimating and improving protein interaction error rates

Patrik D'haeseleer

*Lipper Center for Computational Genetics  
Harvard Medical School  
patrik@genetics.med.harvard.edu*

George M. Church

*Lipper Center for Computational Genetics  
Harvard Medical School  
g1m1c1@arep.med.harvard.edu*

## Abstract

*High throughput protein interaction data sets have proven to be notoriously noisy. Although it is possible to focus on interactions with higher reliability by using only those that are backed up by two or more lines of evidence, this approach invariably throws out the majority of available data. A more optimal use could be achieved by incorporating the probabilities associated with all available interactions into the analysis.*

*We present a novel method for estimating error rates associated with specific protein interaction data sets, as well as with individual interactions given the data sets in which they appear. As a bonus, we also get an estimate for the total number of protein interactions in yeast. Certain types of false positive results can be identified and removed, resulting in a significant improvement in quality of the data set. For co-purification data sets, we show how we can reach a tradeoff between the “spoke” and “matrix” representation of interactions within co-purified groups of proteins to achieve an optimal false positive error rate.*

## 1. Introduction

The arrival of large-scale protein-protein interaction data over the past few years opens up the possibility to gain insight in the rich network of interactions inside living cells. Mirroring the excitement engendered by the arrival of large-scale expression data only a few years earlier, it has spurred an explosion of interest in the genome-wide study of protein interaction networks, and (together with advances in mass spectrometry) has given rise to the new field of “proteomics”.

However, even more so than is the case for gene expression data, large-scale interaction assays have

proven to be notoriously noisy, and one simply cannot use the data blindly, without paying attention to the error rates, biases and artifacts involved.

The surprisingly small overlap between high-throughput yeast 2-hybrid data sets was noticed early on by Ito *et al* [1], who speculated that this might be partly due to a lack of saturation of the screens (see also [2]). In other words, if two assays only found a small subset of all protein interactions, we would expect few interactions to be found by both. An alternative hypothesis is that the data sets contain large numbers of false positive errors, few of which would be expected to occur in the overlap between the data sets. It is impossible to distinguish between these two scenarios by comparing only two data sets with each other. However, we will show that by adding a third, reference data set in the mix, we are able to separate out the effects of lack of saturation and false positives, and derive estimates for the error rates involved.

A number of other methods have been proposed to estimate the reliability of protein interactions data. Mrowka *et al* [3], Deane *et al* [4] and Deng *et al* [5] note that high-throughput protein interactions tend to have lower levels of mRNA co-expression than known interactions, and estimate how many random protein pairs would need to be added to achieve the same distribution of co-expression. This method (as pointed out explicitly by Mrowka *et al* [3]) is sensitive to biases in earlier proteome research towards co-expressed proteins (see also von Mering *et al* [6]), and can occasionally yield large variances on the error estimate.

In addition to co-expression, Deane *et al* also uses interactions between paralogs (similar to the notion of “interologs” by Walhout *et al* [7]) to distinguish true positive interactions [4]. This method is applicable to individual interactions and has high selectivity, but low sensitivity because of the absence of paralogs for

some proteins. In addition, they do not provide a direct estimate of the likelihood of a putative interaction.

Sprinzak *et al* [8] uses annotations of localization and cellular role for the putative interacting proteins to estimate the number of false positive interactions in each data set, matching the observed degree of co-localization and co-cellular role to that of a mixture of true and random interactions. This method relies on the accuracy and completeness of the annotations (as the paralogs method above), as well as on assumptions about the degree of co-localization and co-cellular role in real interactions.

Saito *et al* [9], Goldberg & Roth [10] and Bader *et al* [11] suggest measures based on the number of interaction partners shared by a protein pair, and the connectivity of the pair itself. This method could potentially form the basis for an estimate of the probability that two proteins truly interact.

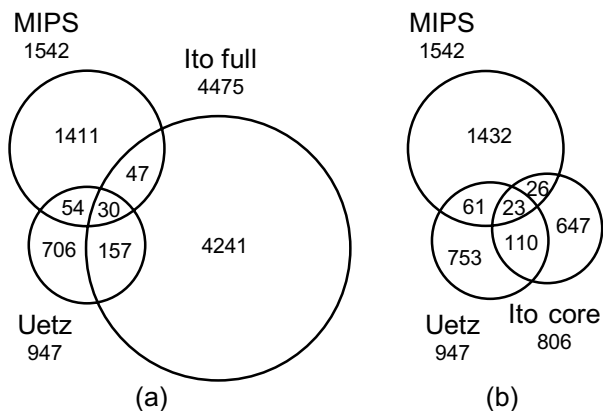
Lastly, Gilchrist *et al* [12] use a Bayesian technique to derive the probability that two proteins occur within the same complex, based on the number of times a pair has been observed to co-purify when using one of them as bait.

Our own approach shares certain features with the analysis presented in the review of large-scale protein interaction data sets by von Mering *et al* [6], in the sense that it uses a trusted reference data set to assess the reliability of various data sets. But whereas von Mering *et al* only calculated an accuracy value that was directly related to the size of the reference set (fraction of the data set covered by the reference set), we estimate absolute error levels by incorporating a third data set into the analysis.

Each of these methods has its own advantages and disadvantages. Given the large number of errors, and the existence of biases both within the experimental data sets, as well as within the reference data sets, literature and annotations, it is important to have a number of independent estimators of the error rates, based on different lines of evidence. As we will show, the different approaches largely agree at least on the overall reliability of the different data sets (if perhaps not always on the reliability of individual interactions). Next-generation methods will integrate these separate lines of evidence into a single reliability/error estimate.

## 2. Error Estimation

The analysis uses two experimental data sets and a reference data set, which we will assume for now is error-free. We will show how the error rates can be

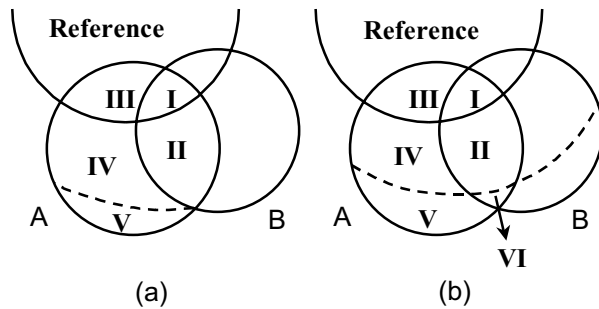


**Figure 1** Relative size of the MIPS reference set, the Uetz, Ito full (a) and Ito core (b) yeast 2-hybrid data sets and their intersections. Numbers refer to the size of the set or subset.

estimated based on the size of the overlap between all three data sets. For our analyses, we used a reference data set of 1542 protein interactions culled from the MIPS table of physical interactions [13] (after removal of a number of high-throughput 2-hybrid data that had already been entered into MIPS, and a small number of interactions annotated as “most probably nonspecific”). The yeast 2-hybrid data sets examined consist of one by Uetz *et al* containing 947 unique protein interactions [14], and one by Ito *et al* containing 4475 putative interactions (referred to as “Ito full” for the rest of this paper), plus a higher quality “core” data set of 806 interactions observed three or more times (further referred to as “Ito core”) [1]. High-throughput co-purification data sets examined consist of one by Gavin *et al* containing 3761 bait-target protein pairs [15], and one by Ho *et al* containing 3618 bait-target pairs [16].

Figure 1 illustrates the sizes of the yeast 2-hybrid data sets compared to the MIPS reference set, and the amount of overlap between them. (See Figure 3 for co-purifications data sets.)

Figure 2 outlines the different subsets used in the calculation of the false positive error rates. As a first approximation, we will assume that the intersection between the two experimental data sets (subsets I and II in Figure 2a) is error-free. If subset V contains all the false positive interactions of data set A (i.e. those protein pairs which are included in the data set but which do not correspond to an actual biological protein-protein interaction), then subsets I, II, III, and IV are all error-free. If the experimental data sets are



**Figure 2. Estimation of number of true positives based on ratio of intersections. A and B are experimental data sets. Subsets I-IV contain only true interactions. (a) The size of subset IV is initially estimated based on the ratio between subsets I and II. The remaining subset V determines the false positive error rate. (b) After calculating the false positive rates of A and B, the size of subset II is adjusted, and the error rates are recalculated.**

independent (we will discuss later how all these assumptions can be relaxed), the ratio between subsets I and II on Figure 2a should be equal to the ratio between subsets III and IV. From this we can calculate the size of IV, i.e. the number of true interactions in A that are not included in B or the reference set:  $IV=III \times II/I$ . The remainder (subset V) contains the false positive interactions and thus determines the false positive rate of the data set. The same method can be used to estimate the false positive rate of the second data set B. we can also derive an initial estimate of the total number of real interactions by comparing which part of the intersection of A and B is covered by the reference set. This gives us an estimate of the false negative rate, i.e. the fraction of real interactions that is missed by the datasets. In the case where we take MIPS as our reference set, Uetz as set A, and Ito core as set B (see Figure 1b),  $I = 23$ ,  $II = 110$ ,  $III = 61$ , so  $IV = 61 \times 110 / 23 = 291.7$ , which leaves an expected number of false-positive interactions in Uetz equal to  $V = 461.3$ .

These initial estimates for the error rates of A and B can now be used to correct our initial assumption that the intersection between A and B is error-free. If we define the accuracy  $\alpha$  to be the fraction of real interactions in a data set, and  $K_r$  and  $K_n$  to be the estimated number of real interactions and non-interactions, then analogous to Equation 4 of Deng *et al* [5], we can use Bayes' rule to calculate the probability

that an interactions is true, given the fact that it occurs in both datasets:

$$\Pr(\text{interaction } ij \text{ is real} \mid ij \in A, ij \in B) = \frac{\alpha_A \alpha_B}{\alpha_A \alpha_B + \frac{K_r}{K_n} (1 - \alpha_A)(1 - \alpha_B)}$$

For the example above, we find that  $A \cap B$  is expected to contain only 0.1% false positives (area VI in Figure 2b). We can then use this updated estimate of the number of true interactions in the intersection to recalculate the error rates for A and B. This process is iterated a small number of times until convergence is achieved

The independence assumptions stated above can be significantly relaxed. In fact, for the calculation of the positive error rates, only a conditional independence is required: Within the putative interactions listed in data set A, whether they appear in the reference set should be independent of whether they also appear in data set B, and vice versa for the interactions listed in data set B (i.e. B is independent of the reference given A, and A is independent of the reference given B). Since the portion of the reference set which falls outside the experimental data sets is irrelevant for the calculation of the false positive error rate, the composition and biases in the make-up of the reference set are irrelevant as well, as long as it is not biased differently with respect to either of the two data sets, or with respect to their intersection. This assumption is reasonable when both data sets were generated from screens using the same method (e.g. both are 2-hybrid data sets), but would not necessarily be expected to hold when comparing data sets generated using different methods. For this reason, we make sure to always use similar data sets in the analysis (e.g. two 2-hybrid data sets vs. the reference set, or two co-purifications data sets vs. the reference set).

Likewise, the "trusted" reference set does not need to be 100% error-free to calculate the false positive error rate, as long as the intersections with the experimental data sets are close to error-free. Since even intersections between rather poor data sets yield high-confidence interactions, this assumption will be valid provided the reference set is of a reasonable quality.

For calculation of the total number of protein interactions in yeast, (and thus the negative error rate of the data sets), the composition and error rate of the reference set does come into play. However, we may be able to account for these effects by judicious choice of data sets, and estimation of the error rate of the reference set itself.

### 3. Removing auto-activators

One important source of false positive errors in yeast 2-hybrid systems consists of a relatively small set of proteins that appear to interact with a very large number of other proteins. These can be due to "sticky" proteins with a large number of nonspecific interactions [15],[17], or auto-activator proteins that are able to activate the reporter gene even in the absence of an interaction partner [17],[18],[19].

This type of error is particularly noticeable in the full Ito data set: of the 25 proteins with 30 or more interactions in the combined MIPS, Uetz and Ito data sets, more than 90% of these interactions were only found when the proteins were used as bait in the Ito data set (see also [18]). For example, the single most connected protein across these three data sets is JSN1, an otherwise unremarkable protein involved in mRNA catabolism, with a total of 289 interactions. 285 were found only when using JSN1 as bait in the Ito assay, none when the protein was used as prey in the same assay, only four (and different) interactions with JSN1 were found in the Uetz data set, and no interactions were found in the previous literature (MIPS data set). Since the probability that such an uneven distribution across the data sets would happen at random is negligible ( $p=7.09 \times 10^{-129}$ ), we will assume that the majority of these interactions are false positives caused by an auto-activator bait protein, and reject them as systematic errors in the Ito data set.

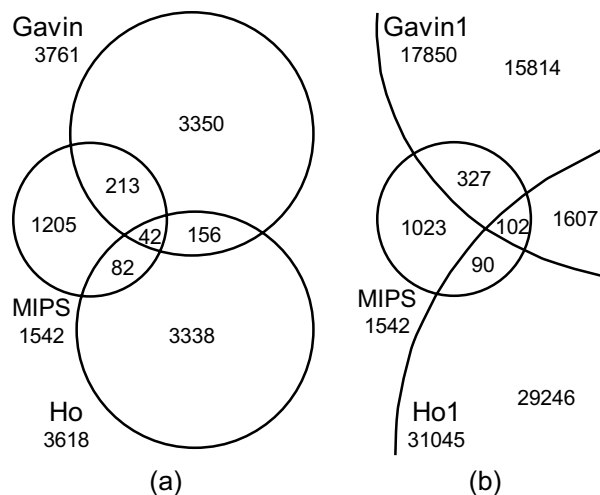
Based on this criterion, we can identify a total of 32 proteins (listed in Table 1) whose interactions are very significantly ( $p < 0.01$  after Bonferroni correction) over-represented as baits in the Ito full data set. Since they also tend to be the most highly connected proteins, these false positives cover almost 2000 interactions, or close to half the entire Ito full data set. As we will see, removal of these false positives significantly decreases the false positive error rate of the data set. Interestingly, although many of these spurious interactions were preferentially removed by the criterion used to generate the Ito core data set, a good number of them remain: Ito core still contains 246 interactions where these proteins were used as bait (close to 30% of the data set). This indicates that even the already higher quality core Ito data set can be further improved by removal of these false positives.

### 4. Co-purification data sets

**Table 1. Over-represented baits in Ito full data**

Gene	Total	MIPS	Uetz bait	Uetz prey	Ito full bait	Ito full prey	Ito core bait	Ito core prey	p-value
JSN1	289	0	0	4	285	0	22	0	7.09E-129
ATP14	124	0	0	0	123	1	7	0	2.04E-55
NUP116	127	3	0	0	125	0	15	0	2.71E-53
SUA7	100	1	0	0	99	0	2	0	9.25E-44
BZZ1	91	0	0	0	91	0	18	0	1.12E-41
SER3	96	0	0	1	95	2	14	1	5.90E-39
VMA6	88	0	0	1	88	1	7	1	6.12E-37
SRB4	101	5	0	0	95	3	5	0	1.34E-33
SRP1	140	8	17	0	122	1	55	0	2.03E-32
RIF2	80	2	0	0	78	1	1	1	6.99E-31
LYS14	63	0	0	0	63	0	2	0	5.74E-28
SOH1	69	0	0	1	67	1	9	0	6.87E-27
MEC3	82	4	1	1	75	3	4	1	8.81E-24
RPB9	44	0	0	0	44	0	2	0	1.15E-18
SAS10	32	0	0	0	32	0	3	0	8.62E-13
VID22	31	0	0	0	31	0	1	0	2.66E-12
MCM21	34	0	0	0	33	1	0	1	6.51E-12
DID4	41	0	1	0	38	3	4	1	2.45E-11
CRM1	38	5	0	0	33	1	6	0	9.49E-08
TEM1	72	0	24	0	54	0	20	0	1.21E-07
DSE3	26	0	2	0	25	0	1	0	3.84E-07
TRM7	20	0	0	0	20	0	0	0	6.46E-07
YNL092W	29	0	0	1	26	2	6	1	8.90E-07
MUK1	30	0	0	1	27	3	4	2	1.70E-06
GCD7	23	0	1	1	22	0	8	0	8.92E-06
TFB1	24	0	3	1	23	0	7	0	8.74E-05
BUD7	15	0	0	0	15	0	0	0	1.81E-04
SRB2	18	0	0	0	17	1	0	0	2.38E-04
KAP95	35	8	0	0	27	0	4	0	2.87E-04
NUP84	19	0	0	0	17	2	1	1	1.57E-03
ADY3	31	0	3	0	25	5	14	1	1.63E-03
GTS1	23	0	0	0	19	4	4	0	4.07E-03

The type of data generated by the high throughput co-purification assays is fundamentally different in nature from the pairwise interactions derived from yeast 2-hybrid assays. Each bait will typically co-purify with a small cluster of other proteins, without any additional information on how this small set of proteins interact with each other. In order to integrate



**Figure 3** Relative size of the MIPS reference set, the Gavin and Ho co-purification data sets and their intersections. (a) Using the spoke model for interactions within a co-purification. (b) Using the matrix (fully connected) model, which results in a much larger number of putative interactions. Numbers refer to the size of the set or subset.

this data with pairwise interaction data sets, two different approaches have been used: The “spoke” model only assumes that the bait protein interacts with each of the co-purified target proteins individually, whereas the “matrix” model assumes that *all* the co-purifying proteins interact [20]. Obviously, the latter approach yields a much larger number of interactions, but it is also susceptible to a larger number of false positives [20]. Figure 3 illustrates the size of the co-purification data sets under the spoke and matrix models, compared with the MIPS reference set.

We propose an alternative approach, which scales smoothly between the two earlier models, by adding to the “spoke” interactions only those protein pairs that occur in  $N$  or more different co-purifications. For large  $N$ , this reduces to the spoke model. The most consistently co-purifying protein pairs will get added first, but as  $N$  decreases, the criterion for inclusion becomes less stringent and more protein pairs are added, until at  $N=1$  it reduces to the matrix model where all co-purifying protein pairs are included. Intermediate values such as  $N = 2$  or  $3$  allow us to include consistently co-purifying proteins, without resorting to the fully connected assumption of the matrix model.

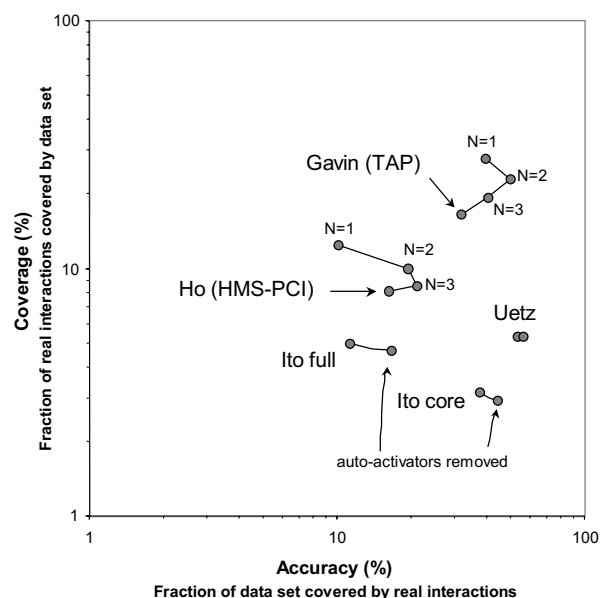
**Table 2** False positive rates of high-throughput protein interaction data sets.

Data set	Size	False pos. rate (%)	Number of true pos.	Number of false pos.
Uetz	947	46	507	440
Ito full	4475	89	504	3971
Ito full -32	2527	83	417	2110
Ito core	806	62	303	503
Ito core -32	557	55	248	309
Gavin	3761	68	1197	2564
Gavin3	5140	59	2085	3055
Gavin2	6800	50	3415	3385
Gavin1	17850	60	7059	10791
Ho	3618	83	582	3036
Ho3	4427	79	923	3504
Ho2	7595	81	1468	6127
Ho1	31045	90	3159	27886

## 5. Results

Table 2 shows the resulting false-positive rates (more formally, the False Discovery Rate (FDR): the estimated proportion of false positives in the data set) for all the data sets examined. “Ito full -32” and “Ito core -32” refer to the filtered versions of the Ito data sets, with all interactions with one of the 32 auto-activators used as bait removed. “Gavin1” and “Ho1” refer to the “matrix” interpretation of the Gavin and Ho co-precipitation data. “Gavin2” etc. refers to the  $N=2$  intermediate between the spoke and matrix model for the Gavin data set, whereas “Gavin” itself refers to the spoke model for this data set. Error rates for all the intersections between data sets are not shown, but are typically very low (with a high of around 20% false positives for the intersection between the two lowest quality data sets, Ho1 and Ito full). To allow for easy comparison, the same information is also represented in Figure 4, in a similar format as Figure 2 in von Mering *et al* [6]. (“Accuracy” in Figure 4 is equal to one minus the false positive rate).

The 46% false positive rate for the Uetz data set was estimated using Ito core as the second experimental data set. Using the much larger and more error-prone Ito full data set to estimate the Uetz error rate yields a value of 43%, which illustrates the robustness of the method with respect of the third data set used. (Both estimates for the Uetz data set are shown in Figure 4.) In general, the error estimates vary little depending on which second experimental



**Figure 4. Comparison of the Accuracy and Coverage of the different data sets. N=1 to N=3 refers to intermediates between the “spoke” and “matrix” models, with N=1 being the matrix model itself, and the unlabeled point at the bottom being the spoke model.**

data set was chosen (Table 2 uses Ito core for Uetz, Gavin2 for Ho, and Ho3 for the Gavin datasets).

When comparing the error rates associated with the Ito full data set and the “Ito full -32” version generated by removal of the bait interactions with the top 32 auto-activators, we notice that the estimated number of false positives drops by 1861 (from 3971 to 2110), whereas the estimated number of true positives only drops by 87. This indicates that the interactions that were filtered out consisted of 95.5% false positives. Note that even the Ito core data set shows a significant reduction in false positive rate after filtering out the auto-activators. Comparing the numbers of true and false positives, we estimate that the filtered-out interactions consisted of 78% false positives (194 false and 55 true positives). This provides an independent verification that our method of removing auto-activators based on overrepresentation in a single data set is valid, even though it may also result in the removal of a small fraction of true positives.

While the number of true positives in the co-purification data sets strictly increases when moving from the spoke model (Gavin, Ho) to the matrix model (Gavin1, Ho1), it does so at the expense of increasingly large numbers of false positives. Moving gradu-

**Table 3. Estimates for the total number of protein interactions in yeast**

Date set 1	Data set 2	N
Uetz	Ito full	10127
Uetz	Ito full -32	8868
Uetz	Ito core	9564
Uetz	Ito core -32	8535
Ho	Gavin	7257
Ho3	Gavin3	10816
Ho2	Gavin2	14829
Ho1	Gavin1	25440

Source	N
Tucker <i>et al</i> [21]	8-12,000
Sprinzak <i>et al</i> [8]	10-16,600
Legrain <i>et al</i> [22]	15-20,000
Grigoriev [23]	16-26,000 <sup>1</sup>
Walhout <i>et al</i> [7]	< 18,000 <sup>2</sup>
Bader & Hogue [20]	20,000
von Mering <i>et al</i> [6]	> 30,000 <sup>3</sup>

<sup>1</sup> The lower estimate was achieved when removing the JSN1, SRP1 and TEM1 proteins, which were also flagged as auto-activators in our own analysis.

<sup>2</sup> Based on 6 hits/bait, without double-counting.

<sup>3</sup> Based on a “matrix” definition of interactions.

ally from the spoke to the matrix model, the false positive rate first improves when we add the protein pairs that co-occur in N=3 or more co-purifications (Gavin→Gavin3 and Ho→Ho3), because these highly consistently co-purified protein pairs are actually of a higher quality than the “spoke” interactions themselves. The false positive rate reaches a minimum (in Figure 4, the accuracy reaches a maximum) for a low value of N, and then increases again as more spurious interactions are added. For the Gavin data set, the false positive error rate is optimal when we add those protein pairs that occur in at least two separate co-purifications (although it should be noted that for this data set, the “matrix” model is strictly better than the “spoke” model in both accuracy and coverage). The Ho data set requires a slightly more stringent approach, and reaches an optimal false positive rate when only those protein pairs that co-occur in at least three separate co-purifications are added.

Table 3 shows the total number of protein interactions in yeast, estimated using different combinations of two experimental data sets, plus the MIPS reference

set. The estimates are centered around 10,000 interactions, with some bias relative to the size of the experimental data sets used. We are currently working on deriving a single consensus estimate based on all available data sets. These estimates seem fairly low, but they are well within the range of estimates derived by other methods, as shown by the estimates from other sources in the lower half of the table. Note that some of the higher estimates such as Walhout *et al* [7] and Bader & Hogue [20] do not take the 50-80% error rates of the data sets into account. Likewise, the largest estimate of >30,000 by von Mering *et al* [6] is based on co-complexed proteins, not necessarily directly interacting proteins.

## 6. Future Directions

The combined MIPS, Uetz, Ito full, Gavin1 and Ho1 data sets cover a total of 53133 unique putative protein-protein interactions, although only 2449 of these are supported by more than one data set (two interactions are actually supported by all five data sets: CKA2-CKB2, and HAP2-HAP5). If we focus only on those data sets with the lowest false positive rates (MIPS, Uetz, Ito core -32, Gavin2 and Ho3), this reduces to 13075 unique interactions, with 1031 covered by more than one data set. This implies that if we were to focus only on those interactions backed up by multiple lines of evidence, we would be throwing out 92-95% of the available data.

On the other hand, if we compare the estimate of the total number of protein interactions in yeast with the estimates of the number of true positive interactions included in the data sets (see Table 2), it seems plausible that many—if not most—real interactions have already been included in some high-throughput data set. Additional high-throughput data sets of the same level of quality would help increase the number of real interactions covered by multiple data sets. For example, the combined MIPS, Uetz, Ito full and Ho1 data sets (minus Gavin1) only contain 512 interactions that are covered by two or more data sets. In other words, the addition of the relatively large and accurate Gavin1 data set caused an almost five-fold increase in the number of well-supported protein interactions. Nevertheless, at current error rates of around 50% (best-case), the number of spurious interactions will still rise much faster. Further advances in the field may come from development of higher-accuracy assays, rather than from collecting more data with similar error rates.

Meanwhile, in order to complete the protein interaction network, it may be useful to focus experimental efforts on the more than 1000 yeast ORFs that have so far not been touched by the high-throughput interaction assays. This may prove to be an exercise in frustration, as many of these ORFs are likely to be hard to clone, or have failed in some other way in the previous assays. Nevertheless, they pose a significant gap in our knowledge of the yeast proteome.

In the introduction, we already mentioned several methods that have been used to assess the reliability of protein interaction data sets, or of individual protein interactions. Rather than collecting more interaction data, integrating and refining these error models should allow us to get much more mileage out of the already existing data. If it is true that many of the real interactions have already been sampled, better error models will allow us to pinpoint them among the multitudes of false positive interactions. The recent work by Jansen *et al* [24] and Asthana *et al* [25] on Bayesian modeling of co-complexed proteins might be extended to direct protein-protein interactions. And for interactions that have not yet been sampled, error models that are based on other types of data may even be able to predict likely interacting proteins, as shown by Goldberg & Roth [10].

In addition to pointing out where the true interactions may be found, a reliable estimate of the probability that individual interactions are real can also be useful for various probabilistic analyses. Nearly any type of analysis that has already been done on the network of binary protein interactions could be rephrased as acting on a weighted network of putative interactions, where the interaction weights are the probabilities of the individual protein-protein interactions. Recent work by Bader [26] and Asthana *et al* [25] uses protein interaction probabilities to extend partially known protein complexes. In an earlier analysis, Steffen *et al* [27] traced signaling pathways through the network of protein interactions. Given probabilities of the individual interactions in the network, we should be able to find the most likely pathway from receptor A to transcription factor B, where the probability of a linear pathway is simply the product of probabilities of the protein interaction links that make up the pathway.

## 7. Acknowledgements

The authors would like to thank John Aach and Yuan Gao for helpful discussions and critical reading of the manuscript. Patrik D'haeseleer is a

PhRMA/Harvard CEIGI fellow. This work was supported in part by the US Department of Energy (DE-GF02-87ER60565).

## References

- [1] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proc Natl Acad Sci U S A*, 2001 Apr 10, 98(8):4569-74. .
- [2] Hazbun TR, Fields S, "Networking proteins in yeast", *Proc Natl Acad Sci U S A*, 2001 Apr 10, 98(8):4277-8.
- [3] Mrowka R, Patzak A, Herzog H, "Is there a bias in proteome research?", *Genome Res*, 2001 Dec, 11(12):1971-3.
- [4] Deane CM, Salwinski L, Xenarios I, Eisenberg D, "Protein interactions: two methods for assessment of the reliability of high throughput observations", *Mol Cell Proteomics*, 2002 May, 1(5):349-56.
- [5] Deng M, Sun F, Chen T, "Assessment of the reliability of protein-protein interactions and protein function prediction", *Pac Symp Biocomput*, 2003, 140-51.
- [6] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P, "Comparative assessment of large-scale data sets of protein-protein interactions", *Nature*, 2002 May 23, 417(6887):399-403.
- [7] Walhout, AJ, Boulton, SJ, and Vidal, M, "Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm", *Yeast*, 2000, 17:88-94.
- [8] Sprinzak E, Sattath S, Margalit H, "How reliable are experimental protein-protein interaction data?", *J Mol Biol*, 2003 Apr 11, 327(5):919-23.
- [9] Saito R, Suzuki H, Hayashizaki Y, "Interaction generality, a measurement to assess the reliability of a protein-protein interaction", *Nucleic Acids Res*, 2002 Mar 1, 30(5):1163-8.
- [10] Goldberg DS, Roth FP, "Assessing experimentally derived interactions in a small world", *Proc Natl Acad Sci U S A*, 2003 Apr 15, 100(8):4372-6.
- [11] Bader J, Chaudhuri A, Rothberg JM, and Chant J, "Gaining confidence in high-throughput protein interaction networks", *Nat Biotech*, 2004 Jan, 22:78-85.
- [12] Gilchrist MA, Salter LA and Wagner A, "A statistical framework for combining and interpreting proteomic datasets", *Bioinformatics*, 2004, 20(5):689-700.
- [13] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B, "MIPS: a database for genomes and protein sequences", *Nucleic Acids Res*, 2002 Jan 1, 30(1):31-4.
- [14] Uetz P *et al*, "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*", *Nature*, 2000 Feb 10, 403(6770):623-7.
- [15] Gavin AC *et al*, "Functional organization of the yeast proteome by systematic analysis of protein complexes", *Nature*, 2002 Jan 10, 415(6868):141-7.
- [16] Ho Y *et al*, "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry", *Nature*, 2002 Jan 10, 415(6868):180-3.
- [17] Serebriiskii IG, Golemis EA, "Two-hybrid system and false positives. Approaches to detection and elimination", *Methods Mol Biol*, 2001, 177:123-34.
- [18] Aloy P, Russell RB, "Potential artifacts in protein-interaction networks", *FEBS Lett*, 2002 Oct 23, 530(1-3):253-4.
- [19] El Housni H, Vandebroere I, Perez-Morga D, Christophe D, Pirson I, "A rare case of false positive in a yeast two-hybrid screening: the selection of rearranged bait constructs that produce a functional gal4 activity", *Anal Biochem*, 1998 Aug 15, 262(1):94-6.
- [20] Bader GD, Hogue CW, "Analyzing yeast protein-protein interaction data obtained from different sources", *Nat Biotech*, 2002 Oct, 20(10):991-7.
- [21] Tucker CL, Gera JF, Uetz P, "Towards an understanding of complex protein networks", *Trends Cell Biol*, 2001 Mar, 11(3):102-6.
- [22] Legrain P, Wojcik J, Gauthier JM, "Protein-protein interaction maps: a lead towards cellular functions", *Trends Genet*, 2001 Jun, 17(6):346-52.
- [23] Grigoriev A, "On the number of protein-protein interactions in the yeast proteome", *Nucleic Acids Res*, 2003 Jul 15, 31(14):4157-61.
- [24] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M, "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data", *Science*, 2003 Oct 17, 302:449-453.
- [25] Asthana S, King OD, Gibbons FD, and Roth FP, "Predicting Protein Complex Membership Using Probabilistic Network Reliability", *Genome Res*, 2004, 14:1170-1175.
- [26] Bader J, "Greedy building protein networks with confidence", *Bioinformatics*, 2003, 19:1869-1874.
- [27] Steffen M, Petti A, Aach J, D'haeseleer P, Church G, "Automated modeling of signal transduction networks", *BMC Bioinformatics*, 2002 Nov 1, 3(1):34.