BIOINFORMATICS

Vol. 20 no. 5 2004, pages 689–700
DOI: 10.1093/bioinformatics/btg469

A statistical framework for combining and interpreting proteomic datasets

Michael A. Gilchrist[1,*], Laura A. Salter[2] and Andreas Wagner[1]

[1]Department of Biology and [2]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87106, USA

Received on August 27, 2003; accepted on September 25, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: To identify accurately protein function on a proteome-wide scale requires integrating data within and between high-throughput experiments. High-throughput proteomic datasets often have high rates of errors and thus yield incomplete and contradictory information. In this study, we develop a simple statistical framework using Bayes' law to interpret such data and combine information from different high-throughput experiments. In order to illustrate our approach we apply it to two protein complex purification datasets.

Results: Our approach shows how to use high-throughput data to calculate accurately the probability that two proteins are part of the same complex. Importantly, our approach does not need a reference set of verified protein interactions to determine false positive and false negative error rates of protein association. We also demonstrate how to combine information from two separate protein purification datasets into a combined dataset that has greater coverage and accuracy than either dataset alone. In addition, we also provide a technique for estimating the total number of proteins which can be detected using a particular experimental technique.

Availability: A suite of simple programs to accomplish some of the above tasks is available at www.unm.edu/~compbio/software/DatasetAssess

INTRODUCTION

Numerous experimental approaches are available to characterize protein function on a genome-wide scale. They include two-hybrid assays which detect direct protein–protein interactions (Ito *et al*., 2001; Uetz *et al*., 2000), mass spectroscopy of purified protein complexes which detect associations of proteins within a complex (Gavin *et al*., 2002; Ho *et al*., 2002) and gene deletions that assess a proteins impact on metabolism and fitness (Steinmetz *et al*., 2002; Allen *et al*., 2003). Each of these approaches provides information on a particular level of biological organization (i.e. direct interactions versus complex composition versus phenotypic effects). Additional sources of information, each with its own strengths and shortcomings, such as gene expression data, evolutionary comparisons and network topology can provide further insight into protein function (Tong *et al*., 2002; Deane *et al*., 2002; Saito *et al*., 2002, 2003). This wealth of information, however, also has its problem. Specifically, each experimental technique to characterize protein function has its own source of both random and systematic errors. Such errors can lead to contradictory results within and between high-throughput experiments. This problem underscores the need for a cohesive framework to integrate data from multiple sources to understand a proteins role within a cell.

We believe that this study represents an important step in the development of such a framework. Below, we develop a simple Bayesian approach which permits the integration of information within and between protein interaction datasets. We illustrate our approach with data on protein complexes, which allows us to calculate the probability that two proteins occur within the same complex. Our approach, however, can just as easily be applied to direct protein–protein interaction datasets which, in contrast, would permit calculation of the probability that two proteins interact directly.

Because our framework is Bayesian in nature, we can integrate information from replicated experiments using one experimental technique, as well as information from experiments using different experimental techniques. Other recent studies have also employed Bayesian techniques to evaluate or identify possible protein–protein interactions (Edwards *et al*., 2002; Goldberg and Roth, 2003). Our approach is currently limited in that it can only be applied to datasets which provide information on the same level of biological organization (e.g. direct interactions or protein complexes). Although our framework permits integration of data from different such levels, to do so is beyond the scope of this study.

Here, we analyze the results of two recently published high-throughput experiments that purified hundreds of protein complexes in the yeast *Saccharomyces cerevisiae*. The study by Gavin *et al*. (2002) used tandem affinity purification (TAP) while the study by Ho *et al*. (2002) used high-throughput mass-spectrometric protein complex identification (HMS-PCI). We will refer to the datasets from these studies by the specific purification technique used.

*To whom correspondence should be addressed.

Both the TAP and the HMS-PCI datasets use techniques to purify protein complexes in which a focal 'bait' protein is modified by integrating a standard polypeptide 'hook' into the protein via standard recombinant DNA techniques. The bait protein is then expressed inside a cell where it may carry out its function as part of one or more protein complexes. To detect the other proteins in a complex, the complex is purified from a cell lysate via affinity chromatography using the hook of the bait protein. The purified proteins retained on the affinity column are then separated and identified using mass-spectrometry. These proteins are often referred to as 'prey' proteins, a convention we employ here. Because the prey proteins are thought to associate with the bait within a protein complex, we define a 'true' association as occurring between proteins which are members of the same complex.

The systematic and random errors encountered in protein complex characterization fall into two categories, false negative and false positive errors. False negative errors occur when an experiment fails to identify all members of a protein complex. Conversely, false positive errors occur when an experiment identifies additional proteins that are not part of the protein complex. Previous work (e.g. Edwards *et al*., 2002; von Mering *et al*., 2002) has shown that the false negative and false positive error rates for any given technique to identify protein interactions can be quite high.

Our goal in this study is to build a statistical model, which takes random errors into account, allowing us to calculate the probability that two proteins co-occur in the same complex give the available experimental data. Our statistical model is based on a mechanistic description of how the data in a single protein complex characterization experiment is generated. In addition, our model also allows us to estimate the false positive and false negative random error rates of a dataset without the use of a protein complex reference set.

Below, we first illustrate our approach with a simple hypothetical example. We then apply our model to the TAP and HMS-PCI datasets individually and jointly. This allows us to combine the TAP and HMS-PCI datasets to increase further both accuracy and coverage. We then evaluate our models ability to predict protein–protein associations by comparing our results to a set of known associations. Our comparison shows that the model can predict the probability that two proteins are truly associated to a remarkable degree. The model produces a complete weighted graph of pairwise protein associations. We show how this weighted graph can be converted into an unweighted graph, and that this graph has statistical properties consistent with that found in other studies of yeast protein interaction networks.

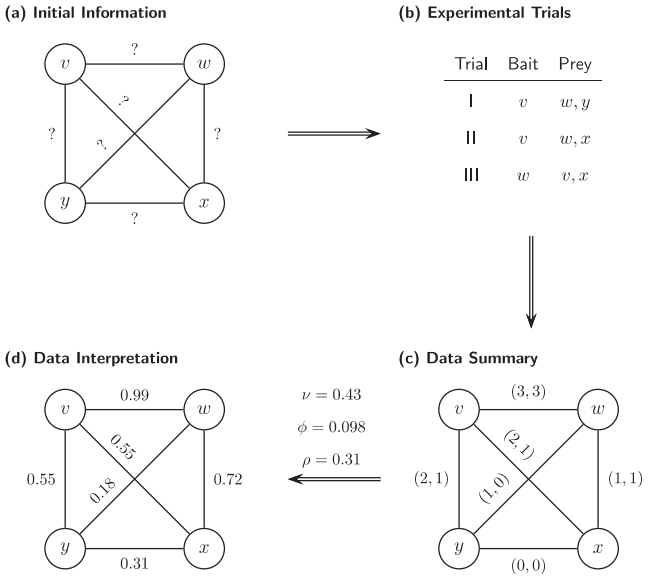

Fig. 1. Illustration of our statistical approach with a hypothetical dataset. (**a**) Two proteins are connected by an edge if they are part of the same protein complex. The panel illustrates the initial lack of confidence in any such association. (**b**) The results from three separate hypothetical experimental trials in which protein v was used twice as a bait protein and protein w was used once. (**c**) Representation of the experimental data from the trials in (b) through their (t, s) values. The symbol t represents the number of opportunities (trials) we had to observe an association while s represents the number of times such an association was experimentally observed. (**d**) The posterior probability of each possible protein–protein association based on the data summarized in (b) and a hypothetical false positive error rate, ϕ, false negative error rate, ν, and the prior probability of an association, ρ. The specific values shown for these error rates are arbitrarily chosen and merely serve to illustrate the principle behind our approach.

MODEL AND RESULTS

A hypothetical dataset

From a sampling perspective, each experiment using one bait protein provides an opportunity or 'trial' to gather information on which proteins associate with that bait protein. For example, imagine a scenario involving only four proteins v, w, x and y (Fig. 1). If we were to conduct an experiment using protein v as a bait, we can view this experiment as an opportunity to observe a possible association between protein v and the proteins w, x and y. In repeating this experiment, we would have a second chance to observe associations of v with w, x and y. A third experiment, now using protein w as a bait, can be viewed as a third opportunity to observe a possible association between proteins v and w, as well as a first opportunity to observe a possible association between protein w and proteins x or y. At the end of these three experiments, we have had three trials for observing an association between v and w, two trials for observing an association between v and x, one trial for observing an association between w and x and no trials for observing an association between x and y. We define t as the number of trials we have for observing an association between two particular proteins. For example, in

the above scenario, t is equal to 3, 2, 1 and 0 for the protein pairs (v, w), (v, x), (w, x) and (x, y), respectively.

However, just because we have the opportunity to observe an experimental association does not mean we will necessarily do so. Consequently, we define s ('success') as the number of experimental observations that two proteins associate ($0 \leq s \leq t$). Imagine that we carry out the above three experiments and that we get the following results: our first experiment using protein v as a bait identifies proteins w and y, our second experiment using v identifies w and x, and our single experiment using w identified v and x (Fig. 1b). In Figure 1c we illustrate how these experimental results can be summarized as a set of (t, s) values for each possible association. As we shall show, the number of trials t to observe an association and the number of times s such an association is experimentally observed will affect our confidence that such an association truly occurs.

A statistical model for analyzing protein association data

As we have just shown, we can categorize experimental information on any particular protein–protein association by the number of experimental trials and successes in a dataset, i.e. by t and s, respectively. Our goal in this section is to build a statistical model using Bayes' law that allows us to interpret this information in a quantitative manner. After building this statistical model, we will first illustrate its use by applying it to the above hypothetical dataset. Then, we will apply it to the high-throughput datasets mentioned above.

We begin by defining two alternative and complementary hypotheses. The first hypothesis, H_1, is that two proteins i and j associate by occurring within the same protein complex. The second hypothesis, H_2, is that proteins i and j do not associate, i.e. they do not occur within the same protein complex. We emphasize that the protein complex data we analyze does not permit inference of direct physical contact between proteins, which motivates our definition of association as adherence to the same complex.

The principal idea behind Bayesian statistics is to improve an estimate of the probability that a hypothesis is correct by weighting information from an experiment with a prior probability, i.e. a probability that the hypothesis is correct in the absence of any such data. This improved estimate is generally referred to as a posterior probability. The goal of our statistical approach is to calculate the posterior probability of H_1, $\Pr(H_1)$, based on available experimental data. [Because H_1 and H_2 are complementary hypotheses, by definition $\Pr(H_1) = 1 - \Pr(H_2)$.] We reach this goal in two steps. First, we calculate the probability of observing s successes in t trials under each of the hypotheses H_1 and H_2. We then use Bayes' law to calculate the posterior probability for H_1.

In order to calculate the probability of observing s successes under H_1 and H_2, we need to define two terms: the false negative error rate ν and the false positive error rate ϕ. Each of these rates is specific to a particular experimental technique and represents the random errors associated with such a technique. Non-random errors, i.e. systematic errors which repeatedly occur due to the inherent nature of an experimental technique, can also occur. However, at this point we will ignore such errors because, as we shall later show, they appear to be much lower than the random error rates ν and ϕ.

The false negative error rate, ν, is equal to the probability that, for any given trial, we will fail to observe an association between two proteins that occur within the same complex. Conversely, the false positive error rate, ϕ, is equal to the probability that, for any given trial, we will observe an association between two proteins that do not occur within the same complex. Using our hypothetical example, if proteins v, w and x associate with one another to form a single complex, then our first experiment with v had one false negative observation because an association between v and x was not observed. This experiment also had one false positive observation because an association between v and y was observed. In contrast, our second and third experiments had no false negative or false positive observations.

If we assume that the random experimental errors are independent of each other, then the probability of observing s associations out of t trials follows a binomial distribution. Under H_1, the two proteins i and j occur within the same protein complex. The probability that we will successfully observe an association between i and j is equal to $1 - \nu$. Thus if H_1 is true, the probability of observing s associations out of t trials for proteins i and j is,

$$\Pr(s|H_1, t, \nu) = \binom{t}{s} \nu^{t-s} (1-\nu)^s. \quad (1)$$

In contrast, under the complementary hypothesis H_2, the two proteins i and j do not occur within the same protein complex. The probability that we will observe a false association between i and j under H_2 is equal to ϕ. Thus, if H_2 is true, the probability of observing s associations out of t trials is

$$\Pr(s|H_2, t, \phi) = \binom{t}{s} (1-\phi)^{t-s} \phi^s. \quad (2)$$

Because our false positive error rate is defined from a sampling perspective of the prey population, it should be noted that it differs from the false positive error rates estimated by other researchers (e.g. Mrowka *et al*., 2001; Edwards *et al*., 2002). These researchers define a false positive error rate, FP, as the probability that an observed association does not actually occur. In contrast, our false positive error rate, ϕ, measures the probability that any given prey protein will be erroneously purified in an experimental trial. The false positive rate FP is useful for interpreting results of a specific experimental trial in a high-throughput dataset. However, it is not only a function of the false positive sampling error rate, ϕ, but also the false negative sampling error rate, ν. We will discuss more explicitly how ν and ϕ relate to FP in the Results section.

Equations (1) and (2) allow us to calculate the probability of s observations out of t trials under the assumption that either H_1 or H_2 is true, respectively. Bayes' law allows us to combine these probabilities in order to calculate the posterior probability that H_1 is true. The use of Bayes' law requires that we introduce the prior probabilities that either H_1 or H_2 are true. Prior probabilities represent our knowledge of a system before the incorporation of additional information such as our values of t and s (Hilborn and Mangel, 1997). Let ρ denote the prior probability for H_1, i.e. ρ is equal to the probability that two proteins selected at random from a proteome are found within the same protein complex. It follows that the prior probability for H_2 is simply $1-\rho$. Applying Bayes' law to our system, it follows that the posterior probability of H_1 given t experimental trials and s experimental observations is,

$$\Pr(H_1|t,s,\nu,\phi,\rho) = \frac{\Pr(s|H_1,t,\nu)\rho}{\Pr(s|H_1,t,\nu)\rho + \Pr(s|H_2,t,\phi)(1-\rho)}. \quad (3)$$

Bayes' law, as applied in Equation (3), simply states that the posterior probability of H_1 is equal to the probability of observing the data, (t,s), under the hypothesis H_1 weighted by the prior probability of H_1, divided by the total probability of observing (t,s). In the absence of any experimental data (i.e. $t=0$), the right-hand side of Equation (3) simplifies to the prior probability for H_1, ρ. As the number of trials, t, increases, the posterior probability of an association diverges from the prior probability, ρ.

Figure 1d illustrates the application of this approach to our hypothetical example. Given a false positive error rate a false negative error rate, and a prior probability, ρ, that two proteins occur within the same complex, Equations (1) and (2) serve to calculate the posterior probability for any of the possible associations in our hypothetical example. Furthermore, we can represent all the possible protein–protein associations using a complete (fully connected) graph whose nodes correspond to proteins, and whose edges have weights that correspond to the probability that two proteins are part of the same complex (Fig. 1d). The properties of this weighted graph can either be analyzed directly, or one can convert the graph into an unweighted graph by choosing a probability threshold and including only edges with weights above this threshold.

For applications to real data, it is of course important to estimate model parameters such as the false positive and negative error rates. Before showing how to do this, we note that Bayes' law allows us to incorporate information from multiple data sources. Indeed, Equation (3) can be generalized to incorporate information on protein interactions or associations from different experimental techniques, or global network properties such as clustering coefficients (Goldberg and Roth, 2003). For example, if $\vec{t}$ and $\vec{s}$ represent sets of trials t and observations s generated with different experimental techniques and $\vec{\nu}$ and $\vec{\phi}$ represent vectors of mean false negative and false positive error rates associated with each of these techniques, then it follows that,

$$\Pr(H_1|\vec{t},\vec{s},\vec{\nu},\vec{\phi}) = \frac{\Pr(\vec{s}|H_1,\vec{t},\vec{\nu})\rho}{\Pr(\vec{s}|H_1,\vec{t},\vec{\nu})\rho + \Pr(\vec{s}|H_2,\vec{t},\vec{\phi})(1-\rho)}, \quad (4)$$

where,

$$\Pr(\vec{s}|H_1,\vec{t},\vec{\nu}) = \prod_{i=1}^{n} \Pr(s_i|H_1,t_i,\nu_i),$$

$$\Pr(\vec{s}|H_2,\vec{t},\vec{\phi}) = \prod_{i=1}^{n} \Pr(s_i|H_2,t_i,\phi_i).$$

In our current analysis we have assumed that the underlying processes follow a binomial model as in Equations (1) and (2). Other datasets, such as the two hybrid datasets, may require more complex models since they give information on direct protein–protein interactions rather than protein associations within a complex. Furthermore, Equation (4) could be generalized further to allow for other forms of data, such as mRNA expression correlation coefficients or functional data.

Applying our model

In this section we move from a hypothetical dataset to the TAP and HMS-PCI datasets created by Gavin *et al*. (2002) and Ho *et al*. (2002), respectively. The false positive error rate ν, false negative error rate ϕ and prior association probability ρ associated with each of these datasets are unknown. Thus, we must first estimate these parameters. As we will show now, it is possible to do so by studying the distribution of (t,s) values in a high-throughput dataset.

Estimating model parameters

We begin our estimation by noting that Equations (1)–(3) allow us to calculate the probability of observing s experimental associations given t observations and the parameters, ν, ϕ and ρ. From these equations it follows that the likelihood $\mathcal{L}$ of observing a set of parameter values ν, ϕ and ρ given a single (t,s) value is given by

$$\mathcal{L}(\nu,\phi,\rho|t,s) = (1-\nu)^s \nu^{t-s}\rho + \phi^s(1-\phi)^{t-s}(1-\rho). \quad (5)$$

Any one observed (t,s) value does not contain very much information on the parameters, ν, ϕ and ρ. However, because high-throughput datasets contain many experimental trials, the amount of information contained in the distribution of (t,s) values for an entire dataset can be appreciable. We find it convenient to tabulate this distribution in a matrix Z, whose entries $z_{t,s}$ correspond to the number of times a particular pair of values (t,s) occurred in a high-throughput experiment. Assuming independence between associations, the total likelihood, $\mathcal{L}$, of a set of values ν, ϕ and ρ given this matrix Z

can be calculated from Equation (5) as

$$\mathcal{L}(\nu, \phi, \rho|\mathrm{Z}) = \prod_{t=1}^{t_{\max}} \prod_{s=0}^{t} \left[(1-\nu)^s \nu^{t-s} \rho + \phi^s (1-\phi)^{t-s} (1-\rho)\right]^{z_{t,s}}, \tag{6}$$

where $t_{\max}$ is the maximum number of times any one association has been used in the high-throughput experiment. By finding the parameter values that maximize $\mathcal{L}$ of Equation (6) for a given set of experimental data (t, s), we can obtain maximum likelihood estimates of ν, ϕ and ρ.

In order to arrive at such estimates, it is necessary to know approximately how many different prey proteins a particular experimental technique can detect. For our likelihood function in Equation (6), the values of $z_{t,s}$ for $s = 0$ will be a function of both the total number of trials and the number of different detectable prey protein. For instance, if an experimental technique can only detect interactions between 1000 of all 6000 yeast proteins, then each bait protein experiment effectively conducts a trial for 1000 possible associations. In this case, a single bait experiment that detects three interacting proteins generates three instances of $(t, s) = (1, 1)$ and 997 instances of $(t, s) = (1, 0)$. If, however, an experimental technique can potentially detect 2000 prey proteins, then the same single bait experiment would generate three instances of $(t, s) = (1, 1)$ and 1997 instances of $(t, s) = (1, 0)$. In the Appendix, we use a randomization approach to estimate the number of detectable prey proteins. Our results indicate that, on average, the TAP and HMS-PCI datasets sample from a population of approximately 2500 prey proteins. A computational analysis (data not shown) of yeast codon usage bias, and experimental analyses of protein expression data (Gygi *et al.*, 2000) suggest a reason for this low prey population size: mass-spectroscopy techniques that rely on previous electrophoretic separation are poor at detecting proteins at low abundances. Such proteins constitute a large fraction of the yeast proteome.

Having estimated the number of detectable prey proteins allows us to maximize Equation (6), and thus to obtain maximum likelihood estimates of $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$ without having to refer to a protein reference set. Because both the TAP and HMS-PCI datasets can detect similar types of protein–protein associations, and because our estimates (Appendix A) for the number of prey proteins they can detect are similar, we constrained the interaction prior ρ to be the same for both high-throughput datasets. A stand-alone software program which estimates these parameters given a Z matrix and the number of detectable prey proteins is available at http://www.unm.edu/~compbio/software/DatasetAssess.

In addition to increasing our confidence that an association does or does not exist between two proteins, our ability to combine datasets also increases the number of protein–protein associations on which we have any information. For example, of the 2.45×10^6 possible protein–protein associations detectable with these techniques, the TAP and HMS-PCI datasets provide information on approximately 25 and 20% of them, respectively. In contrast, the combined dataset provides information on 39% of all possible protein–protein associations, thus illustrating a great advantage of being able to combine information from separate datasets.

Table 1. Maximum-likelihood estimates of false negative error rate, ν, false positive error rate, ϕ and global association prior, ρ for the TAP (Gavin *et al.*, 2002) and HMS-PCI (Ho *et al.*, 2002) datasets

Dataset	Parameter $\hat{\nu}$	$\hat{\phi}$	$\hat{\rho}$
TAP	0.346*	1.07×10^{-3}*	1.88×10^{-3}
HMS-PCI	0.539*	1.30×10^{-3}*	1.88×10^{-3}

The asterisk (*) indicates parameters which differ between datasets at the $P \ll 0.0001$ level using likelihood ratio tests. As explained in the text, we assumed both datasets had the same prior probability, $\hat{\rho}$.

Comparing the global performance of TAP and HMS-PCI

Table 1 contains maximum likelihood estimates of $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$ as well as the results of a likelihood ratio test (Hilborn and Mangel, 1997) comparing these parameters between datasets. The estimated false negative rates ν are high for both datasets, but the TAP approach is significantly better at detecting true protein–protein association than HMS-PCI. Specifically, our estimates for ν imply that each TAP experimental trial will miss one out of three true protein–protein associations ($\hat{\nu} = 0.346$), whereas the HMS-PCI data will miss one out of two true protein–protein associations ($\hat{\nu} = 0.539$). In addition to having a lower false negative error rate, the TAP dataset also has a significantly lower estimated false positive error rate than the HMS-PCI dataset ($\hat{\phi} = 1.07 \times 10^{-3}$ versus 1.30×10^{-3}). Although the false positive error rates are more than two orders of magnitude less than the false negative error rates, the number of false positive errors in an experimental trial scales with the total number of detectable prey proteins. Specifically, if each bait experiment samples from an estimated population of 2500 experimentally detectable prey proteins, on average we would expect to see $1.07 \times 10^{-3} \times 2500 = 2.66$ false positive protein interactions in each TAP experiment, and $1.30 \times 10^{-3} \times 2500 = 3.25$ false positive interactions in each HMS-PCI experimental trial. Taken together with the association prior, ρ, of 1.88×10^{-3}, these estimates imply that we would expect to see 3.08 and 2.17 true positive interactors for the average TAP and HMS-PCI experimental trial, respectively.

As mentioned previously, our false positive error rate is defined from a sampling perspective of the prey population and differs from the FP false positive error rate used by other researchers (e.g. Mrowka *et al.*, 2001; Edwards *et al.*, 2002). We can estimate FP from our parameter estimates dividing the expected number of observed false positive associations

by the sum of the expected number of observed true and false positive associations, i.e.

$$\mathrm{FP} = \frac{\phi(1-\rho)}{\phi(1-\rho) + (1-\nu)\rho}. \quad (7)$$

Note that the FP is independent of the size of the protein population detectable with a particular technique. Using Equation (7) yields FP values of 0.46 and 0.60 for the TAP and HMS-PCI datasets, respectively. These values are consistent with those estimated by Edwards *et al*. (2002) and the general findings of von Mering *et al*. (2002).

Posterior probabilities of protein associations

The above error rate estimates not only allow a crude global comparison of high-throughput protein complex data, they can also be used to extract more fine-grained information on the likelihood of observing individual interactions, information that depends on the number of trials and observations (t, s) for any particular protein–protein association. To begin with, we note that there are multiple differences among the TAP and HMS-PCI datasets that the above global error analysis does not reveal. For instance, although the TAP dataset consists of fewer experimental trials than the HMS-PCI dataset, the TAP dataset has greater breadth in that it used 588 different bait proteins as opposed to the 490 different bait proteins of the HMS-PCI dataset. Conversely, the HMS-PCI dataset has greater depth in that approximately 33% of the bait proteins were used more than once. Thus even though the error rates associated with the HMS-PCI dataset is greater than the TAP dataset, these additional trials can give us greater confidence in the presence or absence of specific individual associations. Furthermore, we can use Equation (4) to combine the information from both datasets so that we may take advantage of both the breadth of the TAP dataset with the depth of the HMS-PCI dataset. This combined dataset consists of 1325 experimental trials with 984 different prey proteins.

Using Equation (3) and our maximum likelihood estimates (MLEs) of the false negative error rate, $\hat{\nu}$, false positive error rate, $\hat{\phi}$ and interaction prior, $\hat{\rho}$, we created tables of posterior probability values for the TAP and HMS-PCI datasets (Tables 2 and 3). Because these tables are organized by the number of experimental trials to observe an association, t, and the number of times, s, an association was observed, we will refer to them as t–s posterior probability tables. We also calculated a t–s posterior probability table from Equation (4) for the (t, s) data from the combined TAP and HMS-PCI datasets (Table 4).

Examining the t–s posterior probability tables (Tables 2–4), we find that the probability that an observed association truly occurs changes with the number of experimental trials, t, and observations, s, in a straightforward manner. For example, with only one experimental trial and an observed association, i.e. $(t, s) = (1, 1)$, we can only have moderate confidence that the association really exists [$\Pr(H_1) = 0.537$ in Table 2]. This lack of strong confidence is due to the fact that both the false positive error rate, ϕ, and the association prior, ρ, are of similar magnitude. If we carry out two trials and observe an association with a particular prey protein in one trial, and no association in the other trial, the posterior probability that the observed association truly occurs is even lower [$\Pr(H_1) = 0.286$ in Table 2]. However, if two trials detect the same association ($s = 2$), then the likelihood that this association truly occurs attains a value of $\Pr(H_1) = 0.999$. In general, with only one experimental trial, it is unclear whether an observed association reflects a true association or a false positive error. However, for any given number of trials, t, the likelihood that an association truly occurs

Table 2. The t–s posterior probability table for the TAP (Gavin *et al*., 2002) dataset

Trials, t	Observations, s		
	0	1	2
0	0.00188	—	—
1	0.000652	0.537	—
2	0.000226	0.286	0.999

Each cell in the table represents the posterior probability that two proteins associate given t trials and s experimental observations of an association. Posterior probability values were generated using Equation (3) and the MLEs of the false negative error rate, false positive error rate and the association prior, $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$, respectively.

Table 3. The t–s posterior probability table for the HMS-PCI (Ho *et al*., 2002) dataset

Trials, t	Observations, s				
	0	1	2	3	≥4
0	0.00188	—	—	—	—
1	0.00102	0.401	—	—	—
2	0.000549	0.265	0.996	—	—
3	0.000296	0.163	0.992	1.00	—
4	0.00016	0.0951	0.986	1.00	1.00
5	8.62e−05	0.0537	0.974	1.00	1.00
6	4.65e−05	0.0297	0.953	1.00	1.00
7	2.51e−05	0.0162	0.916	1.00	1.00
8	1.35e−05	0.00883	0.854	1.00	1.00
9	7.3e−06	0.00478	0.76	1.00	1.00
10	3.94e−06	0.00258	0.63	0.999	1.00
11	2.13e−06	0.0014	0.479	0.998	1.00
12	1.15e−06	0.000754	0.332	0.997	1.00
13	6.19e−07	0.000407	0.211	0.994	1.00
14	3.34e−07	0.00022	0.126	0.99	1.00

Each cell in the table represents the posterior probability that two proteins associate, given t trials and s experimental observations of an association. Posterior probability values were generated using Equation (3) and the MLEs of the false negative error rate false positive error rate and the association prior, $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$, respectively. This dataset differs from the TAP dataset in that multiple experiments were carried out with the same bait protein.

Table 4. A subset of the t–s posterior probability table based on both the TAP (Gavin *et al*., 2002) and HMS-PCI (Ho *et al*., 2002) datasets

HMS-PCI trials, t	HMS-PCI successes, s				
	0	1	2	3	≥4
(a) TAP $(t,s) = (2,0)$					
0	0.000226	—	—	—	—
1	0.000122	0.0742	—	—	—
2	6.57e−05	0.0414	0.966	—	—
3	3.55e−05	0.0228	0.939	1.00	—
4	1.91e−05	0.0124	0.892	1.00	1.00
5	1.03e−05	0.00674	0.817	1.00	1.00
6	5.57e−06	0.00365	0.707	0.999	1.00
7	3e−06	0.00197	0.565	0.999	1.00
8	1.62e−06	0.00106	0.412	0.998	1.00
9	8.74e−07	0.000575	0.274	0.996	1.00
10	4.72e−07	0.00031	0.169	0.993	1.00
11	2.54e−07	0.000167	0.0992	0.986	1.00
12	1.37e−07	9.03e−05	0.0561	0.975	1.00
13	7.41e−08	4.87e−05	0.0311	0.955	1.00
14	4e−08	2.63e−05	0.017	0.919	1.00
(b) TAP $(t,s) = (2,1)$					
	0	1	2	≥3	
0	0.286	—	—	—	
1	0.178	0.993	—	—	
2	0.105	0.987	1.00	—	
3	0.0592	0.976	1.00	1.00	
4	0.0329	0.957	1.00	1.00	
5	0.018	0.923	1.00	1.00	
6	0.00979	0.867	1.00	1.00	
7	0.00531	0.778	1.00	1.00	
8	0.00287	0.654	0.999	1.00	
9	0.00155	0.505	0.999	1.00	
10	0.000837	0.355	0.997	1.00	
11	0.000452	0.229	0.995	1.00	
12	0.000244	0.138	0.991	1.00	
13	0.000132	0.0796	0.983	1.00	
14	7.1e−05	0.0446	0.968	1.00	
(c) TAP $(t,s) = (2,2)$					
	0	1	≥2		
0	0.999	—	—		
1	0.997	1.00	—		
2	0.995	1.00	1.00		
3	0.991	1.00	1.00		
4	0.984	1.00	1.00		
5	0.97	1.00	1.00		
6	0.946	1.00	1.00		
7	0.905	1.00	1.00		
8	0.836	1.00	1.00		
9	0.734	0.999	1.00		
10	0.598	0.999	1.00		
11	0.445	0.998	1.00		
12	0.302	0.997	1.00		
13	0.189	0.994	1.00		
14	0.112	0.988	1.00		

Each cell in a table represents the posterior probability that two proteins associate given t trials and s associations for the HMS-PCI dataset. Posterior probability values were generated using Equation (3) and the MLEs of the false negative error rate, false positive error rate and the association prior, $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$, respectively. (a) when $(t,s) = (2,0)$ in the TAP dataset, (b) when $(t,s) = (2,1)$ in the TAP dataset and (c) when $(t,s) = (2,2)$ in the TAP dataset.

increases dramatically as the number of observed associations s increases. This is less obvious from Table 2, because the TAP data does not contain more than $t = 2$ trials for any protein. It is more strikingly demonstrated by Table 3, because the HMS-PCI data contains up to $t = 14$ trials for some associations. Table 3 also shows the confidence that any one association truly occurs decreases for a constant number of association observations, s, in an increasing number of trials, t. For instance, for low and moderate values of t (i.e. $t \leq 7$) two or more observations of an association ($s \geq 2$) leads to a posterior association probability greater than 0.9. However, if t increases to eight trials the posterior association probability decreases to 0.854 for $s = 2$. For $t = 10$ it further decreases to 0.63. The reason for this decrease in the posterior probability is that the chance of observing the same false positive association multiple times increases with the number of trials.

Incorporating data from different datasets has the same qualitative effect as conducting additional trials. For example, if a particular association has (t,s) values equal to $(2,0)$ and $(6,2)$ for the TAP and HMS-PCI data, respectively, the posterior probability that this association truly exists is 0.707. A similar value of 0.854 for this posterior probability would result if the TAP data contained no trials for this association but the HMS-PCI dataset contained eight trials and two observations, i.e. $(t,s) = (0,0)$ and $(8,2)$, respectively. The discrepancies between these values comes from the differences in error rates between the datasets. The complete tables for the combined dataset as well as a database of trial, success and posterior probability values are available at http://www.unm.edu/~compbio/software/DatasetAssess.

Model validation

For evaluating the predictive ability of our posterior probabilities, we used the MIPS Complex Catalog as a reference set. The MIPS Complex Catalog (http://mips.gsf.de/proj/yeast/catalogues/complexes/) is a hand-curated database created in 1998 in which the composition of protein complexes was confirmed by a variety of experimental techniques but does not include information from high-throughput datasets. The information in this catalog is at the protein complex level and not the level of direct protein–protein interactions, thus allowing us to compare our posterior probability values to the actual frequency at which an association between two proteins with a given (t,s) value occurs according to this database.

Overall, the MIPS Complex Catalog contains 1045 unique open reading frames (ORFs) which implies more than 5×10^5 possible pairwise protein–protein associations. Out of these 5×10^5 possible associations, only 8711 are documented as actually occurring in the MIPS Complex Catalog. We can classify each of the 5×10^5 possible protein–protein associations by the corresponding number of trials, t, and observations, s, in the TAP and HMS-PCI data. Doing so

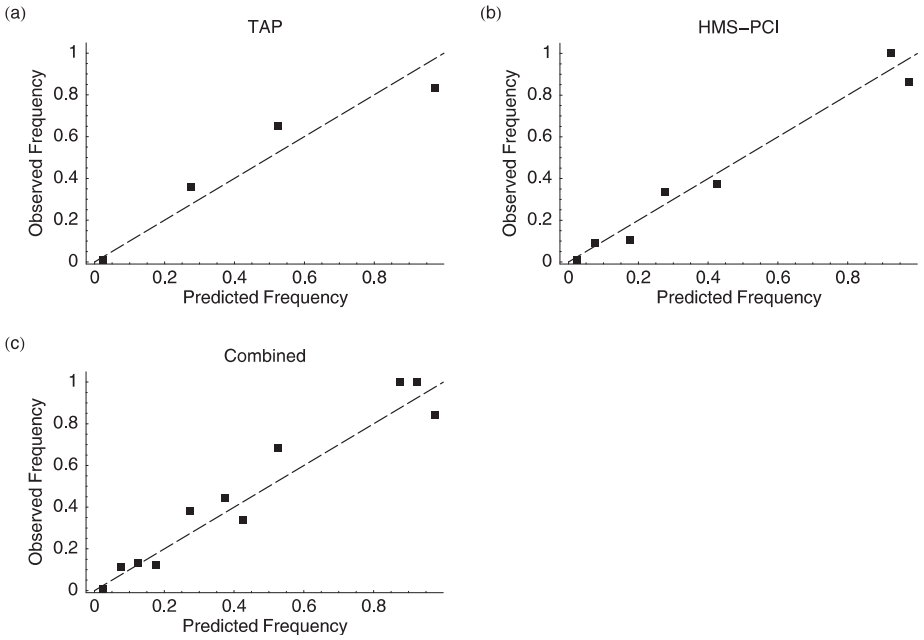

Fig. 2. Observed frequency of protein–protein associations in the MIPS dataset versus the expected frequency for the (**a**) TAP (Gavin *et al*., 2002), (**b**) HMS-PCI (Ho *et al*., 2002) and (**c**) combined datasets. Expected frequency values are based on Equations (3) and (4) and the corresponding $\hat{\nu}$, $\hat{\phi}$ and $\hat{\rho}$, values. The dashed line illustrates the expected 1 : 1 correlation between observed and expected frequencies. Data has been binned in 5% intervals.

allows us to calculate the frequency at which a true association occurs for a given set of (t, s) values in a given dataset. For example, in the TAP dataset, there are 707 possible protein–protein associations where $(t, s) = (1, 1)$. Of these 707 possible associations, 459 correspond to a true association according to the MIPS Complex Catalog. Therefore, the observed frequency of a true protein–protein association given $(t, s) = (1, 1)$ is 0.649 for the TAP data. The corresponding posterior probability value in Table 2 is 0.537. The results of applying this approach to all values of (t, s) are summarized in Figure 2, which illustrates that despite having used no information from the MIPS Complex Catalog, we can use our statistical model to predict the probability of a true association in this reference dataset to a remarkable degree. For example, fitting a linear function to the weighted data explains more than 90% of the variance and results in a highly significant model fit ($P < 0.001$) in all three cases with slopes ranging from 0.92 to 0.95.

Comparing global protein network statistics among different data sources

We can represent all possible protein–protein associations in a graph whose nodes correspond to proteins and whose edges have weights that correspond to the probability that two proteins are part of the same complex (Fig. 1d). As described below, we converted the weighted graphs that result from our analysis of protein complexes into unweighted graphs, in order to compare their global structure with those of protein interaction networks derived from other interaction data. While one loses information when converting the weighted posterior probability graph into an unweighted graph, unweighted graphs are the most common means of representing association networks and, consequently, one can draw on a wide variety of tools for their analysis (Harary, 1969; Bollobás, 1985). In addition, previous analyses of proteomic datasets all utilize unweighted graphs, providing a baseline for comparison. It is important to note that in graphs based on binary, two-hybrid interaction data (Wagner, 2001), an edge between two proteins represents a direct interaction. In contrast, in the protein complex data we analyze here, an edge represents an association between two proteins on the level of the complex. As a consequence, we do not expect a congruence between the statistics for these two different types of graphs.

We converted our weighted graphs into unweighted graphs by retaining all edges with posterior probability values greater than or equal to 0.5 and by disregarding all edges with a posterior value less than 0.5. A cutoff value of 0.5 may seem arbitrary, but this choice is motivated by the fact that it minimizes the total number of mismatch errors between the simple graph we construct and the graph one can construct based on the information in the MIPS Complex Catalog discussed above (Fig. 3). For each unweighted graph we calculated the number of connected components, the average degree of a node, as well as the distribution of node degrees. For the nodes in the largest components in each graph we

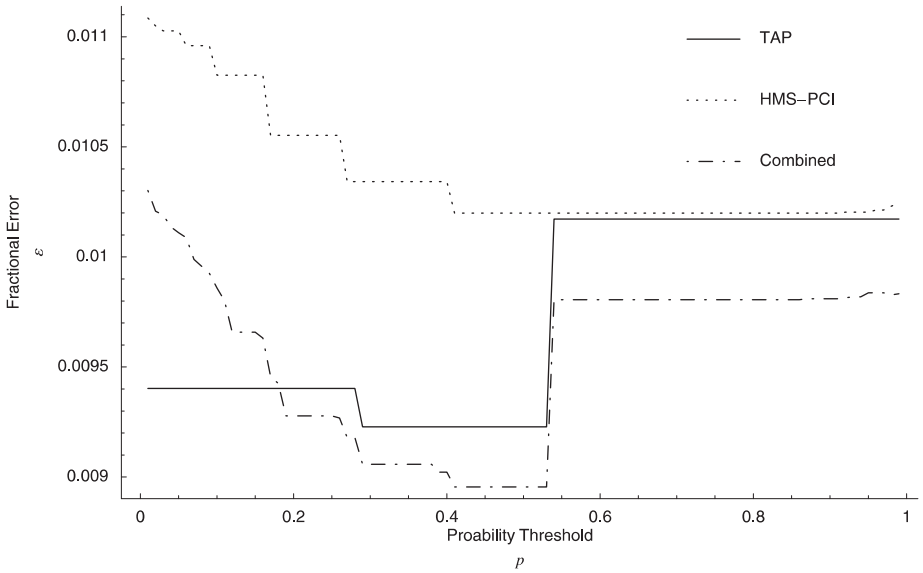

Fig. 3. The fractional error for an unweighted protein interaction subgraph whose nodes are all contained in the MIPS Complex Catalog, as a function of the probability threshold value for retaining edges between nodes. Errors were computed by comparing the subgraph to a reference graph of known protein–protein associations derived from the MIPS Complex Catalog. An error is defined as a discrepancy between the experimentally derived subgraph and this reference graph. The fractional error is equal to the total number of errors divided by the total number of possible associations in the graph.

Table 5. Statistics for unweighted graphs derived from TAP, HMS-PCI and combined datasets

	TAP	HMS-PCI	Combined
All components			
Nodes	1322	350	1371
Components	51	66	77
Mean degree	3.99 (4.31)	1.78 (1.85)	3.48 (3.79)
Degree exponent	1.45	2.26	1.54
Largest component			
Nodes	1169	86	1136
Mean clustering coefficient	0.11 (0.23)	0.01 (0.05)	0.10 (0.26)
Mean path length	5.63 (0.86)	5.88 (1.65)	5.87 (1.05)

Isolated, i.e. disconnected, nodes were ignored for these calculations. Values in parentheses represent SD.

calculated the mean clustering coefficient and the characteristic path length. Finally, we compared the clustering coefficients for the entire graph to the node degree to look for evidence of hierarchical structure in the graph (Ravasz *et al*., 2002). The results are summarized in Table 5 and representative distributions are presented in Figure 4.

For the TAP and combined dataset, the average protein degree was greater than that observed earlier for binary protein interactions (Wagner, 2001). Such a higher value is expected given that we are studying complex level associations which include direct and indirect interactions, rather than just direct interactions. For all three datasets the degree distribution can be approximated by a power function where the frequency of a protein of degree y is proportional to d^{-a} (Fig. 4a). Similar to other studies, we do find that the tail of the distribution falls off faster than expected under a power law. The exponents a are within the range found in several previous studies that first suggested a power-law degree distribution in protein interaction networks (Jeong *et al*., 2001; Wagner, 2001; Ho *et al*., 2002).

The average clustering coefficient in the largest graph components were orders of magnitude greater and the characteristic path lengths were considerably smaller than those observed earlier (Wagner, 2001). Again, these differences are not surprising. For example, one would expect a lower path length between proteins in our graphs because each protein has, on average, more associations than direct interactions. Finally, we note that the clustering coefficient of a protein appears to decline inversely with protein degree (Fig. 4b). Ravasz *et al*. (2002) argue that such a relationship is evidence of a graph's hierarchical structure.

DISCUSSION

The approach we presented here rests on a representation of high-throughput protein association data in the form of (t, s) values, where t indicates the number of trials or opportunities to observe an association between a bait and a prey protein, and s indicates the number of times the association was actually observed. We showed how to use this information, together

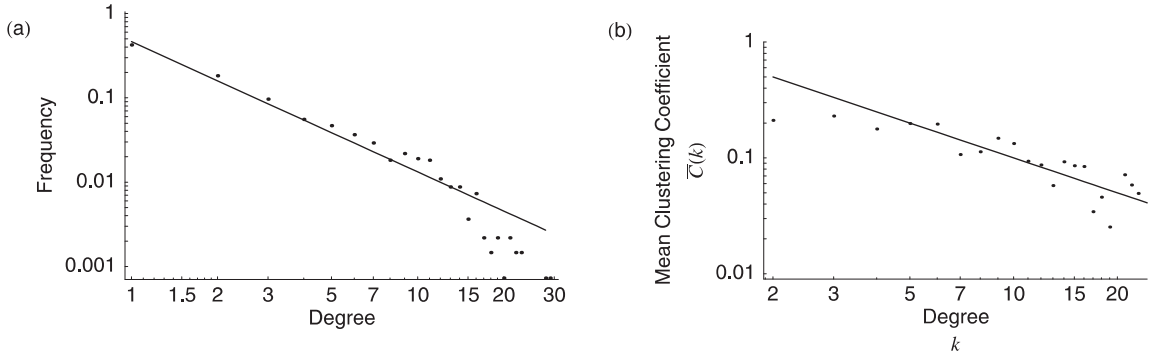

Fig. 4. (**a**) Degree distribution and (**b**) mean clustering coefficient, $\bar{C}(k)$, versus degree, k, for data combined from the TAP and HMS-PCI experiments (Gavin *et al*., 2002; Ho *et al*., 2002). The solid line in (a) represents expected frequencies of proteins with degree y, based on a power law distribution where $\Pr(d) \propto d^{-\hat{a}}$. $\hat{a}$ is the MLE for the exponent of the distribution. The solid line in (b) represents the expected relationship $\bar{C}(k) \propto k^{-1}$ given hierarchical structuring.

with a binomial model of protein sampling, Bayes' law, and a maximum-likelihood estimation of error probabilities, to calculate a posterior probability that any two proteins are associated within the same protein complex. We applied this approach to the publicly available TAP and HMS-PCI high-throughput datasets (Gavin *et al*., 2002; Ho *et al*., 2002). In applying our approach we were also able to estimate independently the random false positive and false negative error rates for each dataset. In spite of the high random error rates in both datasets, our approach permits combining information from multiple experiments within and between different high-throughput datasets. This results in identification of protein associations with high statistical confidence. We validated our approach by showing that the protein association posterior probabilities calculated for any (t, s) value are remarkably similar to the observed association probabilities for proteins in MIPS Complex Catalog, a reference set of manually curated protein complexes. When representing all moderate to high-confidence protein associations as an unweighted graph, we find the graph topology is consistent with that emerging from other studies of protein networks (Jeong *et al*., 2001; Wagner, 2001; Ho *et al*., 2002).

In order to evaluate the quality of proteomic data, previous researchers have used reference sets of 'known' interactions (e.g. Edwards *et al*., 2002; von Mering *et al*., 2002). Our approach does not require the use of such reference sets which is a great advantage, because such reference sets have various shortcomings. First, they include, by their very nature, a small number of protein interactions. Second, they may be biased towards particular types of intensely studied protein complexes. Third, despite manual curation of protein interactions, they may contain false positive errors as well. Our maximum-likelihood estimation of experimental error rates permits examination of the self-consistency of data within a dataset. In other words, information on the random error rates of an experimental technique is contained within a high-throughput dataset itself.

Our approach does not allow us to detect systematic errors and it would be difficult to do so for the given available data. To illustrate the nature of systematic errors, imagine that it was possible to control laboratory conditions such that each replicate experiment using the same bait protein identifies the same prey proteins. Using our likelihood technique, we would estimate the random false positive and false negative error rates to be zero. This, however, would not mean our dataset was error free. For example, the required chemical modifications of a bait protein may change its conformation. They may thus interfere with some of the bait's native protein interactions and they may generate other, spurious interactions. Put differently, such modifications could generate false negative and false positive errors that occur in every experiment testing the same association.

While we cannot currently estimate systematic error rates, these rates are likely to be much smaller than random error rates for the available TAP and HMS-PCI data. This is because the interaction probabilities predicted with our approach are in very good agreement with those derived from the MIPS Complex Catalog, where protein interactions are derived from a wide variety of techniques. If this agreement was poor, then it would suggest that systematic errors may be an important source of error in the protein complex purification techniques used. Nonetheless, it will be important to distinguish between these types of errors in the future and it is possible to expand our statistical framework to take such errors into account.

Further, we note that our ability to estimate random error rates suggests that a similar approach may allow the estimation of systematic error rates. This would require high-throughput data on protein associations derived from different experimental techniques that are subject to different kinds of systematic errors. For example, imagine that we have three large datasets each of which uses the same bait protein in several replicate trials. When two of the datasets indicate a very high probability for a particular bait–prey association, while the third indicates a low probability, the third observation

is best explained as a false negative error. A maximum-likelihood approach similar to the one we used here may then help to estimate the systematic error rates for each of the datasets. However, the importance of using maximally diverse experimental techniques must be emphasized. For example, it is likely that the TAP and HMS-PCI approach to detecting protein interactions generate similar systematic false negative errors because both techniques modify bait proteins in a similar manner.

Several recent studies complementary to ours attempt to validate protein interaction data. Some of these studies also use a Bayesian approach. For example, Edwards *et al*. (2002) showed that although the error rates of most high-throughput datasets are large, they are of similar magnitude as the error rates of smaller-scale experiments. Using data from small-scale experiments, they calculated an odds-ratio of protein–protein interactions from a number of different data sources. However, this approach did not incorporate error rates associated with different experiments and techniques. In another Bayesian analysis, Goldberg and Roth (2003) use global properties of protein interaction networks, such as their high 'clustering coefficients' (Watts and Strogatz, 1998) to increase statistical confidence in individual protein interactions. Non-Bayesian approaches include one by Saito *et al*. (2002, 2003), which is conceptually similar to the one taken by Goldberg and Roth (2003). It uses local network topology to come up with 'interaction generality' metrics. Such metrics are not as immediately interpretable as Bayesian based posterior probabilities.

Perhaps the simplest non-Bayesian approach to integrate information within one dataset is exemplified by the commonly used 'core' dataset of Ito *et al*. (2001), which consists of interactions observed at least three times in a larger high-throughput experiment. Similarly, von Mering *et al*. (2002) combined interaction data consistently observed in a number of experiments, including TAP and HMS-PCI datasets, as well as protein interaction data generated with the two-hybrid technique (Uetz *et al*., 2000; Ito *et al*., 2001). Such approaches can greatly reduce the number of false positive errors due to random sources. However, they do so at the cost of increasing the number of false negative errors. In contrast, Bayesian approaches like ours permit combining data from different sources so that we may increase our confidence that an association does or does not occur increase while, simultaneously, increasing the number of associations examined.

In this study we focused on high-throughput protein complex purification techniques. Such techniques only indicate whether two proteins are part of the same protein complex. This stands in contrast to other techniques, most notable among them the yeast-two-hybrid assay (Fields and Song, 1989), which identify direct interactions among proteins. Information from both types of techniques, which can be integrated by extending our approach, can provide more detailed information on protein interactions and associations than any one technique by itself. For instance, if two proteins are shown to be part of a purified complex, they may be in physical contact if a two-hybrid assay indicates their direct interaction. Conversely, if two proteins are shown to be part of the same complex, but an assay of direct interactions does not indicate their association, then the proteins may not be adjacent in the complex. The amount of support that each type of observation provides will be a function of the quality and amount of experimental data available.

One of the advantages of the probabilistic approach pursued by us and others (Edwards *et al*., 2002; Goldberg and Roth, 2003) is that it moves away from the 'all or nothing' interpretation of interactions in molecular networks. Instead, it attaches a probability or statistical confidence to every possible protein interaction. Although the resulting network of interactions is more difficult to analyze than a simpler, unweighted network, it also contains a much greater amount of information. Another advantage of the Bayesian framework is that it lends itself naturally to integrating data from various different experimental sources into a cohesive and explicitly quantitative framework. Such integration will become increasingly important as more and more sources of functional genomic information become available.

ACKNOWLEDGEMENTS

M.G. would like to thank the G. Conant and A. Evangelisti for their helpful discussions and assistance with this work and acknowledges support through NIH grant GM63882 to A.W. The authors would also like to thank the helpful suggestions and comments from three anonymous reviewers. L.S. acknowledges support through NSF grant DMS 0104290. A.W. would like to thank the Santa Fe Institute for its continued support.

REFERENCES

Allen,J., Davey,H.M., Broadhurst,D., Heald,J.K., Rowland,J.J., Oliver,S.G. and Kell,D.B. (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.*, **21**, 692–696.

Bollobás,B. (1985) *Random Graphs*. Academic Press, Orlando.

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

Edwards,A.M., Kus,B., Jansen,R., Greenbaum,D., Greenblatt,J. and Gerstein,M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**, 529–536.

Fields,S. and Song,O.K. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

Gavin,A., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J., Michon,A., Cruciat,C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci.*, **100**, 4372–4376.

Gygi,S.P., Corthals,G.L., Zhang,Y., Rochon,Y. and Aebersold,R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci.*, **97**, 9390–9395.

Harary,F. (1969) *Graph Theory*. Addison-Wesley, Reading, Massachusetts.

Hilborn,R. and Mangel,M. (1997) *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G., Moore,L., Adams,S., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

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.*, **98**, 4569–4574.

Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Mrowka,R., Patzak,A. and Herzel,H. (2001) Is there a bias in proteomic research. *Genome Res.*, **11**, 1971–1973.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barbasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Saito,R., Suzuki,H. and Hayashizaki,Y. (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res.*, **30**, 1163–1168.

Saito,R., Suzuki,H. and Hayashizaki,Y. (2003) Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics*, **19**, 756–763.

Steinmetz,L.M., Scharfe,C., Deutschbauer,A.M., Mokranjac,D., Herman,Z.S., Jones,T., Chu,A.M., Giaever,G., Prokisch,H., Oefner,P.J. and Davis,R.W. (2002) Systematic screen for human disease genes in yeast. *Nat. Genet.*, **31**, 400–404.

Tong,A.H.Y., Drees,B., Nardelli,G., Bader,G.D., Brannetti,B., Castagnoli,L., Evangelista,M., Ferracuti,S., Nelson,B., Paoluzi,S. *et al*. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, **18**, 1283–1292.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

APPENDIX

Estimating the number of detectable prey proteins

In order to estimate the number of prey proteins that can be observed experimentally with any given technique, we take a statistical approach that asks how the number of unique detected prey proteins changes with each experimental trial. Because the number of proteins in the proteome is finite, the number of experimentally detectable prey proteins is also finite. It follows that the total number of observed prey proteins will approach an asymptotic value as the number of experimental trials becomes very large. This asymptotic value corresponds to the size of the population of prey proteins which can be observed experimentally (the prey population size, for short).

In order to estimate this asymptotic value, we first randomized the order of experimental trials in a high-throughput experimental dataset. We then calculated the total number of unique prey proteins, F, observed after z trials. We repeated this randomization procedure 300 times, averaging the number of unique prey proteins observed after each trial to calculate $\bar{F}(z)$. We found that the following Hill function provides an excellent fit to $\bar{F}(z)$,

$$f(z) = k_1 \frac{z^{k_2}}{k_3 + z^{k_2}}.$$

In this function, $f(z)$ represents the number of unique prey proteins observed in z experimental trials. $f(z)$ is a general univariate function of three parameters, k_1, k_2 and k_3, of which k_1 is the asymptotic value for $f(z)$ as z goes to infinity. The estimated value of k_1 thus represents our estimate of the prey population size. For both the TAP and HMS-PCI datasets we were able to find extremely close fits (explaining more than 99% of the error) for both datasets of $f(z)$ to the mean randomized data $\bar{F}(z)$. We did so by minimizing the mean squared distance between the data and $f(z)$. This procedure yielded estimates of the prey protein population size as 2212 ($\pm$10.8 SE) and 2782 ($\pm$10.7 SE) for the TAP and HMS-PCI datasets, respectively.