PROTEINS: Structure, Function, and Bioinformatics 64:436–443 (2006)

A Statistical Framework to Discover True Associations From Multiprotein Complex Pull-Down Proteomics Data Sets

Changyu Shen,[1] Lang Li,[1] and Jake Yue Chen[2,3]*

[1]*Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana*
[2]*School of Informatics, Indiana University, Indianapolis, Indiana*
[3]*Department of Computer and Information Science, Purdue University School of Science, Indianapolis, Indiana*

ABSTRACT **Experimental processes to collect and process proteomics data are increasingly complex, and the computational methods to assess the quality and significance of these data remain unsophisticated. These challenges have led to many biological oversights and computational misconceptions. We developed an empirical Bayes model to analyze multiprotein complex (MPC) proteomics data derived from peptide mass spectrometry detections of purified protein complex pull-down experiments. Using our model and two yeast proteomics data sets, we estimated that there should be an average of about 20 true associations per MPC, almost 10 times as high as was previously estimated. For data sets generated to mimic a real proteome, our model achieved on average 80% sensitivity in detecting true associations, as compared with the 3% sensitivity in previous work, while maintaining a comparable false discovery rate of 0.3%. Cross-examination of our results with protein complexes confirmed by various experimental techniques demonstrates that many true associations that cannot be identified by previous approach are identified by our method. Proteins 2006;64:436–443.**

© 2006 Wiley-Liss, Inc.

Key words: binomial Bernoulli (BB); complete binomial Bernoulli (CBB); high-throughput; protein–protein interaction; multiprotein complex (MPC); Y2H; yeast two-hybrid

INTRODUCTION

Proteomics studies in post-genome eras are crucial to the understandings of hidden links between genetic predispositions and phenotypes of an organism. For the past two decades, researchers have made significant progress in collecting and analyzing genome sequences from various organisms.[1,2] To gain a "holistic" view of how particular genetic information plays out in living cells, however, requires researchers to continue to invest in collecting, analyzing, and integrating new types of high-throughput experimental data, e.g., global gene/protein expressions and molecular (protein–DNA, protein–protein) interactions. Proteomics provides researchers with the opportunity to observe the posttranscriptional states (presence/absence) of hundreds of gene products—proteins. Therefore, it is possible to deduce a minimal set of "protein biomarkers" as indicators of certain diseases' early prognosis. Interaction-based proteomics, however, provides biologists with molecular binding information between proteins. This information enables computational scientists to build computer models of protein complexes and molecular pathways, which further enables biomedical researchers to explain and find cures to complex human diseases. However, dealing with interaction-based proteomics data is significantly more challenging than common genomics tasks, because brute-force analysis and visualization methods cannot reveal novel insights into biological pathways because of inherent experimental data noise/inconsistency, and complexity of the problem.[3]

In this work, we are interested in the study of interaction-based proteomics data, inspired primarily by the recent progress of high-throughput system-scale protein–protein interaction mapping projects. These projects can be classified into four broad techniques: (i) yeast two-hybrid (Y2H) methods, which seek to measure direct physical interaction among protein pairs in mated yeast hybrid strains;[4–7] (ii) multiprotein complex (MPC) pull-down experimental methods, coupled with a series of protein complex purification, separation, and identification methods often involving liquid chromatography and peptide mass spectrometry techniques;[8,9] (iii) genetic interactions methods, for example, synthetic lethality, which aims to identify closely related proteins in parallel pathways by testing whether cells would die when introduced with double mutations;[10] and (iv) computational protein pairing methods, which assign protein pairs either when there is conserved gene coevolution patterns found in different genomes, or when there is conserved mRNA coexpression patterns under a variety of controlled stimulatory conditions.[11] Other approaches also exist.[12–16] Note that only Y2H and MPC methods provide direct evidence of physical protein–protein interactions. Recent statistical methods on analyzing Y2H data include Chen[17] and Liu et al.[18] However,

*Correspondence to: Jake Yue Chen, Indiana University School of Informatics, 535 W. Michigan Street, IT #493, Indianapolis, IN 46202. E-mail: jakechen@iupui.edu

Received 4 January 2006; Accepted 3 March 2006

Published online 16 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20994

© 2006 WILEY-LISS, INC.

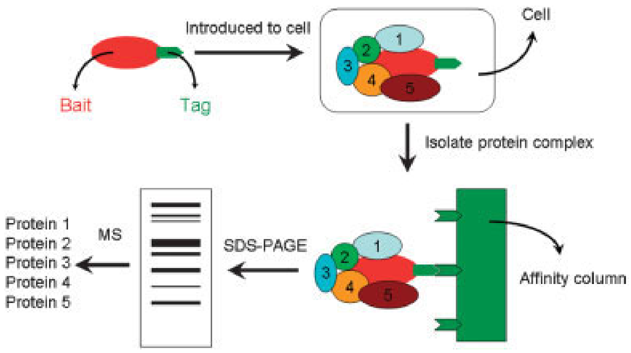

Fig. 1. MPC pull-down experiment.

fewer quantitative analysis tools have been proposed for data yielded from the MPC technique. In this article, we concentrate on the development of statistical analysis of interaction-based proteomics data generated from this technique.

The general strategy of MPC experiment can be described as follows. First, a preselected protein (called "bait") is modified to have a "tag" peptide inserted into the protein's C-terminus using DNA recombination. The DNA vector containing encoded tagged bait protein is subsequently introduced into target expression cells. Next, the bait proteins are profusely expressed in target expression cells and allowed to form complexes with other proteins via protein affinity. The target cells are then disrupted and the protein complexes are harvested, purified, and affixed to solid-state surfaces through protein tags. The transient protein complexes formed from this "pull down of all proteins" are therefore called multiprotein complexes (MPCs). Because each MPC may contain hundreds of associated proteins (called "preys"), it needs to go through careful protein separation procedures such as two-dimensional gel electrophoresis or liquid chromatography until each separated aliquot contains much smaller number of possible types of proteins. Finally, mass spectrometers are used to determine the protein constitutes in each aliquot. A cartoon version of the experiment is shown in Figure 1.

It is not surprising to note that a complex method such as MPC pull-down experiment could be subject to many sources of experimental errors,[19] which has presented a huge challenge in data analysis for subsequent biological pathway studies. For example, system errors could be introduced if samples are contaminated; random errors are also unavoidable, because the quality of final prey protein identifications are subject to accurate collection and interpretation of mass spectrometry peaks. As observed in Edwards et al.[20] and von Mering et al.,[21] errors produced from several high-throughput MPC proteomics projects remain high, or at least uncertain. However, the only available general practice to assess the quality of this type of data is to resort to a "degree of overlap" method, in which a newly collected proteomics data set is compared with another existing experimental data and/or curated protein interaction records to seek agreement between data sets for the identification of interacting proteins.[20,22] Sprinzak et al.[23] used cellular colocalization and annotation term co-occurrence of interacting proteins to assess true-positive interactions from various experiments. However, this type of assessment has been questioned because high-throughput protein interaction data sets may bring together novel proteins whose functions are previously presumed to be unrelated.[7] To our best knowledge, there has been little reported success in setting up a complete quantitative model that can help biologists answer the following question: How do we assess and discover true protein interactions from noisy proteomics data sets?

The main thesis of this work is to introduce an effective statistical framework to gauge the random errors found in MPC proteomics data sets. Specifically, we are interested in making inference on whether or not protein pairs are within the same complex (not necessarily direct physical contact). From now on, we will use the term "association" and "interaction" interchangeably to refer to the concept of "being in the same complex" unless otherwise noted. As shown later, we incorporate into our framework information embedded within the MPC data sets that was ignored by Gilchrist et al.[24] Inclusion of such information results in (i) appropriate definition of the proportion of protein pairs within a proteome that interact with each other and (ii) substantial improvement in the identification of truly interacting protein pairs whereas controlling the false-positive rate at a comparable level.

The remainder of this article is organized as follows. In Materials and Methods, we briefly review some background on the area of analyzing MPC data, followed by introduction of concepts used to construct our statistical framework and model details. In Results, we present analyses of two real data sets, studies based on simulated data sets and a cross-examination study. We complete this article with some concluding remarks in the Discussion section.

MATERIALS AND METHODS

Background

Gilchrist et al.[24] recently described a novel method to estimate "global association prior" (ρ), the percentage of interacting protein pairs among an implicitly defined group of protein pairs. The investigators incorporated the observed association between the single bait protein and all the prey proteins "pulled down" by the bait protein in each MPC experiment into the construction of a binomial-Bernoulli model (the "BB" model). They applied empirical Bayes approach to estimate the global association prior for two yeast MPC protein interaction data sets by Gavin et al.[8] and Ho et al.[9] In the study, the authors concluded that both experiments have a $\rho = 1.88 \times 10^{-3}$. For the data by Gavin et al., the total number of protein pairs under consideration is 6.8×10^{5}, which suggests 1,278 true interacting pairs among the 533 MPCs based on the estimated global association prior. Accordingly, there would have been on average only 2.4 (1,278/533) interactions per MPC. This result seems to be inconsistent with most of today's available MPC experimental results, which are often known to associate several dozens or even hundreds of proteins in a complex. Yarmush and Jayaraman[25] also

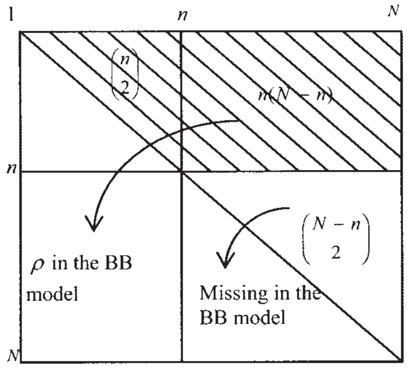

Fig. 2. Illustration of the space of interacting protein pairs. Proteins 1 to n serve as bait proteins and proteins $n + 1$ to N only appear in the preys. The shadowed area refers to the total protein pairs under the consideration of Gilchrist et al.,[24] upon which ρ is defined.

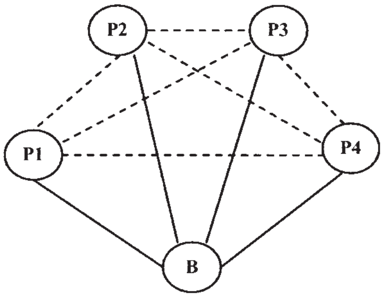

Fig. 3. Schematic drawing of a hypothetical MPC pull-down trial using a bait shown as the node labeled "B." Preys P1, P2, P3, and P4 are also shown as labeled nodes. Solid lines connecting the bait and preys are type I associations and dotted lines connecting preys are type II associations.

suggested that a bait protein in an MPC experiment is generally associated with more than 50 proteins in yeast. What could have gone wrong?

We believe that the fundamental cause of a plausible underestimate of true positives for each MPC experiment lies in the fact that much protein–protein interaction information among prey proteins was unaccounted for in the BB model. Such a simple "spoke" (bait as the "center of spoke") data representation scheme of MPC proteomics pull-down data is a common practice, and can be found in such analysis as in Bader and Hogue.[22] However, a "spoke" model does not take into account interactions among prey proteins, which leads to miscalculated percentage of interacting protein pairs. To elucidate this, suppose that there are totally N proteins within a proteome, among which n proteins are selected as the bait proteins. We are interested in estimating the true association rate among the $N(N - 1)/2$ protein pairs and identifying those protein pairs. What Gilchrist et al.[24] tried to estimate is the true association rate among the $n(n - 1)/2 + n(N - n)$ protein pairs, ignoring those potentially associated pairs within the $N - n$ proteins that do not serve as baits. In Figure 2, the shadowed area is the total protein pairs under Gilchrist et al.'s consideration and their true association rate (ρ) is defined for this population. Hence, it does not reflect the intended true association rate, which should be defined for the shadowed and the blank area above the diagonal line in Figure 2. For the same reason, their model does not provide the posterior probability of having an association for the $(N - n)(N - n - 1)/2$ protein pairs in the blank area.

Concepts

To include the prey–prey associations, we construct a statistical model using missed information in BB to improve sensitivity. First, similar to the concept of "global association prior," we introduce the concepts of true association and true association rate of a proteome:

True association: two proteins have a true association if they are located in the same protein complex within a biological system;

True association rate (ρ): the probability that two proteins randomly selected from a proteome have a true association.

True association cannot be directly observed in any proteomics experiments; however, the association of proteins in MPC experiments can be observed. In general, there are two types of "observable" associations from these experiments: bait–prey association (type I) and prey–prey association (type II). Specifically,

Bait–prey association (type I): proteins A and B have type I association if and only if both proteins are observed in the same MPC trial, and one of the proteins is the bait protein;

Prey–prey association (type II): proteins A and B have type II association if and only if both proteins are observed in the same MPC trial, and both proteins are prey proteins.

In type I association, we are concerned primarily with protein–protein interactions between the bait protein and prey proteins in the MPC. Obviously, such interactions provide direct evidence of true associations between the bait and the preys. In type II association, we are concerned primarily with the protein–protein interactions among prey proteins in the MPC, which provide indirect, yet important, information of true association status among prey proteins. Intuitively, we would expect two truly associated proteins to behave in a concordant manner when a third protein serves as the bait (both in the preys or none in the preys). Hence, type II association fills in the technology inadequacy that not every protein is used as bait. We illustrate these two concepts in Figure 3 (note that the notion "bait–bait association" does not apply in this setting). It is clear that the BB model only includes type I association and loses a fairly large number of protein–protein interactions embedded within the type II association. As shown in later sections, our model accounts for all the information embedded within type I and II associations to infer the likelihood of true association for every pair of proteins in a proteome that might play different roles (bait or prey) at different MPC trials.

The imperfection of any experimental technique is attributable to the random errors associated with it. We first define the following two error terms:

Type I false-positive rate (r_0): probability that two proteins have a type I association given that they do *not* have a true association;

Type I false-negative rate (s_0): probability that two proteins do *not* have a type I association given that they have a true association.

Hence, r_0 and s_0 describe the error when going from the true association status to type I association status. To connect the true association status with type II association, we further define two extra terms:

Type II false-positive rate (r_1): probability that two proteins have a type II association, given that they do *not* have a true association;

Type II false-negative rate (s_1): probability that two proteins do *not* have a type II association, given that they have a true association.

To summarize the definitions of these parameters in a more rigorous manner, suppose that A, B, and C are three proteins randomly selected from a proteome. We have:

$$r_0 = \Pr[\text{A and B have a type I association}|\text{A and B do } \textit{not} \text{ have a true association; A is the bait}],$$

$$s_0 = \Pr[\text{A and B do } \textit{not} \text{ have a type I association}|\text{A and B have a true association; A is the bait}],$$

$$r_1 = \Pr[\text{A and B have a type II association}|\text{A and B do } \textit{not} \text{ have a true association; C is the bait}],$$

$$s_1 = \Pr[\text{A and B do } \textit{not} \text{ have a type II association}|\text{A and B have a true association; C is the bait}].$$

With the parameters just introduced, we construct an empirical Bayes model, the **complete binomial-Bernoulli** model or **CBB**, to account for the information embedded within both type I and type II associations, which enables us to estimate the true association rate for the total $N(N - 1)/2$ protein pairs and calculate the posterior probability of any two specific proteins having a true association.

The CBB Model

We use N to denote the number of proteins in a proteome. Practically, N is defined to be the union of the bait proteins and their preys. To write out the likelihood function, we first label the N proteins by numbers 1, 2 … N such that the first n proteins correspond to the baits. For $k = 1, 2,\ldots,n$; $i, j = 1, 2,\ldots, N$ $(i \neq j)$ and $t = 1, 2\ldots, n_k$ (n_k is the number of trials with bait k), define

$$Z_{ij} = Z_{ji} = \begin{cases} 1; \text{ if proteins } i \text{ and } j \text{ have a true association} \\ 0; \text{ otherwise} \end{cases}$$

$$Y_{ij}^{kt} = Y_{ji}^{kt} = \begin{cases} 1; \text{ if proteins } i \text{ and } j \text{ have a type I association for the } t^{\text{th}} \text{ trial of bait } k = i \text{ (or } j) \\ 1; \text{ if proteins } i \text{ and } j \text{ have a type II association for the } t^{\text{th}} \text{ trial of bait } k \neq i,j \\ 0; \text{ otherwise} \end{cases}.$$

Moreover, let $Y^{kt} = (Y_{ij}^{kt}; i < j,i,j = 1,2\ldots N)$ be a binary vector composed of the status of the $(N - 1)$ type I associations and $(N- 2)/2$ type II associations for the t^{th} trial of bait $Y = (Y^{kt}$; k = 1,2. . .n, t = 1,2,. . .n$_\text{k}$, be the observed data, and $Z = (Z_{\text{ij}}$, i < j, ij = 1,2,. . .N be the binary vector composed of the true association status of the $N(N - 1)/2$ pairs of proteins. Then the probability of observing Y^{kt} given Z can be written as:

$$L_{kt}(Y^{kt}|Z,\theta) = \prod_{j\neq k}^{N}\{(1 - s_0)^{Z_{k1j}}r_0{}^{1-Z_{k1j}}\}^{Y_{1j}^{kt}}\{s_0^{Z_{kj}}(1 - r_0)^{1-Z_{k1j}}\}^{1-Y_{k1j}^{kt}} \quad (1)$$

$$\prod_{i<j,i\neq kj\neq k}^{N} \{(1 - s_1)^{Z_{ij}}r_1^{1-Z_{ij}}\}^{Y_{ij}^{kt}}\{s_1^{Z_{ij}}(1 - r_1)^{1-Z_{ij}}\}^{1-Y_{ij}^{kt}}$$

where $\theta = (\rho,r_0,s_0,r_1,s_1)$. Note that the first product in Eq. (1) involves type I associations and the second product involves type II associations. The model and corresponding likelihood function of Z can be written as:

$$L(Z|\theta) = \rho\sum_{i<j}^{N}Z_{ij}(1-\rho)\sum_{i<j}^{N}(1-Z_{ij}). \quad (2)$$

Conditional on Z, the outcomes from each trial with a particular bait protein can be treated as independent. Hence, the model for Y and Z is:

$$L(Y,Z|\theta) = L(Z|\theta)L(Y|Z,\theta) = L(Z|\theta)\prod_{k=1}^{n}\prod_{t=1}^{n_k}L_{kt}(Y^{kt}|Z,\theta)$$

and the corresponding log-likelihood function is:

$$l(\theta;Y,Z) = \ln[L(Z|\theta)] + \sum_{k=1}^{n}\sum_{t=1}^{n_k}\ln[L_{kt}(Y^{kt}|Z,\theta)]. \quad (3)$$

Because we do not observe Z, it is treated as missing data in our EM algorithm.[26] This algorithm is composed of two steps: the expectation step (E) and the maximization step (M). During the E step of the m^{th} iteration, Z is updated by the conditional expectation given the estimate of θ from the last iteration [$\theta^{(\text{m}-1)}$] and Y, that is, $Z^{(m)} = E[Z|Y,\theta^{(m-1)}]$; then in the M step, we find $\theta^{(m)}$ that maximizes $l[\theta;Y,Z^{(m)}]$. This procedure is repeated until convergence (see Appendix for details). The advantage of this algorithm in our case is that we can obtain a closed form solution during the M step, which greatly enhances the computation speed. Another bonus is that we automatically obtain the probability of two proteins having a true association given Y, or, $\Pr[Z_{ij} = 1|Y]$.

RESULTS

Case Studies

We applied the proposed model CBB to the study by Gavin et al.[8] and Ho et al.,[9] in which high-throughput protein complex data sets for yeast *Saccharomyces cerevi-*

TABLE I. Data Summary and Parameter Estimates From the BB and CBB for the High-Throughput Experiments Using TAP[8] and MSPCI[9]

	TAP		MSPCI	
No. of proteins	1,550		1,533	
No. of baits	533		485	
	BB	CBB	BB	CBB
No. of pairs covered	6.8×10^5	1.2×10^6	6.3×10^5	1.2×10^6
ρ	1.9×10^{-3}	1.4×10^{-2}	1.9×10^{-3}	6.1×10^{-3}
r_0	1.1×10^{-3}	5.4×10^{-3}	1.3×10^{-3}	3.8×10^{-3}
(r_1)		(5×10^{-5})		(4×10^{-5})
s_0	0.346	0.588	0.539	0.762
(s_1)		(0.993)		(0.998)
Estimated no. of interacting pairs	1,278	16,560	1,197	7,320
No. of interacting pairs per MPC	2.4	28	2.5	15.1

ρ, true association rate; r_0, type I false-positive rate; s_0, type I false-negative rate; r_1, type II false-positive rate; s_1, type II false-negative rate.

siae were generated by tandem affinity purification (TAP)[8] and mass spectrometric protein complex identification (MSPCI),[9] respectively. We will refer to these two experiments by the specific techniques they used. The data are summarized in Table I. For example, Gavin et al. purified protein assemblies that cover 1,550 proteins. Among the 1,550 proteins, 533 serve as the bait protein once and 1,017 of them only present themselves as preys. Hence, Gilchrist et al. only considered the $533 \times 532/2 + 533 \times 1{,}017 = 6.8 \times 10^5$ protein pairs, which is 57% of the total number of pairs formed by the 1,550 proteins (1.2×10^6) covered by the CBB (see Fig. 2).

In Table I, we also compare the parameter estimates from the BB model and the CBB model. Clearly, the estimate of the true association rate from the CBB (ρ) is much higher than that from the BB, which indicates that a large amount of true associations within the prey proteins that never serve as the bait proteins have been ignored by the BB. For instance, the CBB postulates that there are $1.2 \times 10^6 \times 1.38 \times 10^{-2} = 16{,}560$ true associations based on the TAP data, which suggests that on average about 28 (16,560/533) true associations are identified for each MPC trial. Note that this number is about 10 times as large as that from the BB model (2.4). Thus, substantial true associations are missed in the BB model by ignoring the prey–prey association in each MPC trial. However, the CBB estimate based on MSPCI seems to suggest less interacting pairs (less than half predicted by the TAP data), which implies a potential difference in the nature of the two data sets. One possible reason is that the two data sets cover fairly different proteins: among the ~500 bait proteins from each data set, only 87 appear in both data sets; among the ~1,500 proteins covered from each data set, the overlap is 664 (~40%). Nevertheless, the CBB model in general captures more protein interactions than the BB.

The type I false-positive rate and type I false-negative rate from the CBB are higher than that from the BB. Roughly speaking, the CBB says that for every 1,000 pairs that do *not* have a true association, 4–5 of them will have a type I association when one member of the pair serves as the bait; and for every two pairs that have a true association, at least one of them will *not* have a type I association. Moreover, the CBB estimates the type II false-positive rate and false-negative rate to be at the level of 10^{-5} and 0.99, respectively. Thus, it is extremely unlikely for two proteins that do not have a true association to appear in the same MPC as "preys" of a third protein. However, there is 0.2–0.7% probability (1–0.998 or 1–0.993) for two truly associated proteins to be "fished" by a third protein, mainly attributable to few bait proteins that truly interact with the two proteins of interest. We will come back to this point in the Discussion section.

Simulation Studies

To identify which pairs have a true association, a routine practice is to apply a cutoff point to the posterior probability of true association; and protein-pairs are identified as positive if their posterior probabilities exceed the cutoff point. Then we are interested in what proportion of the true associations is covered in the positive pairs (the sensitivity or SEN) and what proportion of the positive pairs actually does not have true associations (false discovery rate or FDR). SEN and FDR provide measurements on the prediction quality. Usually, one index is improved at the price of the other one by applying different cutoff points. To evaluate the performance of the CBB as compared with the BB, we generate a hypothetical true association map of 1,000 proteins that mimics the cluster structure in *S. cerevisiae* as demonstrated by Gavin et al.[8] Specifically, we set the type I false-positive rate and type I false-negative rate to be 0.5 and 50%; and one-third of the 1,000 proteins (330) are selected as the bait proteins with each one having one trial. All these numbers are close to what we found in the TAP data (see Table I).

In the first study, we generate 200 data sets with the 330 bait proteins randomly selected. The results for BB and CBB using a cutoff of 0.85 for posterior probability are shown in Table II. On average, whereas the CBB maintains a similar FDR as BB, much more truly associated protein pairs are identified by the CBB (82.7 versus 2.7%). Applying smaller cutoff points to BB will not enhance the sensitivity substantially, mainly because of the limited coverage of protein pairs by this approach (Fig. 2).

TABLE II. Simulation Results Based on 200 Runs for CBB and BB

Model	FDR (SE)	SEN (SE)
CBB	0.32% (0.12%)	82.7% (1.4%)
BB	0.25% (0.21%)	2.7% (0.2%)

cutoff, 0.85; SE, standard error.

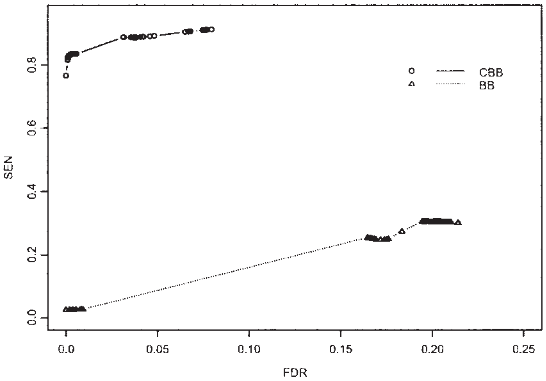

Fig. 4. SEN versus FDR for the CBB and BB in identifying truly associated protein pairs with a fixed set of bait proteins. Different points are obtained by applying different cutoff points (0.1 to 0.99 by 0.01, number of points = 90) to the posterior probability of having true association.

Second, we generate one data set with the 330 bait proteins fixed. The bait proteins are chosen so that each cluster has at least one bait protein and larger clusters have more bait proteins. In Figure 4, we show the graph of SEN versus FDR by applying different cutoff points. The CBB reaches about 90% sensitivity at the price of 10% FDR, whereas the best BB can reach is 30% at the price of a FDR >20%. Hence, it is clear that the CBB has more power to identify truly associated protein pairs than the BB, while in the meantime maintaining a reasonable small FDR.

Cross-Examination

We choose the MIPS Complex Catalog as a reference source to examine the performance of our model. The MIPS Complex Catalog (http://mips.gsf.de/proj/yeast/catalogues/complexes/) is a hand-curated database composed of protein complexes confirmed by various experimental techniques. Every pair of proteins within a complex is defined as having an association. The yeast catalog includes 9,974 associations that cover 1,088 proteins. Then the association rate among the 1,088 proteins is 9,974 × 2/(1,088 × 1,087) = 1.69%, which is very close to the estimate based on the TAP data (1.4%, Table III). We will call the 1,088 proteins and their associations the MP data.

In the cross-examination, we focus on proteins that appear in both the TAP/MSPCI and MP data sets. There are 662 overlapped proteins between TAP and MP and 402 overlapped proteins between MSPCI and MP. We then apply various cut-off points to the posterior probabilities of true associations for the pairs formed by the 662 (or 402) proteins and examine how many of the positive pairs also appear in the MP data. In the comparison, we distinguish two types of protein pairs. We will call a protein pair a “1B” pair if at least one member of the pair ever serves as the bait and a “PP” pair if neither of the members ever serves as the bait. Note that “PP” pairs are not included in the analysis of the BB model. The results are shown in Table III. We use the third row as an example to illustrate how the table should be read. When the cut-off point is 0.9, there are 897 1B pairs identified among the 18,915 1B pairs. Among the 897 pairs, 334, or 37.2% can also be found in the MP. Similarly, there are 2,539 positive PP pairs from the 108,881 PP pairs, among which 1,196 (47.1%) appear in the MP. Totally, 3,436 positive pairs are identified and 1,530 (44.5%) are in the MP. We observe clearly that many PP pairs (those in bold fonts) are validated by the MP, which cannot be discovered by the BB model. In general, the percentage of validation by MP is lower than the posterior probabilities (e.g., 44.5% of pairs with higher than 0.9 posterior probabilities are found in MP). We believe that this is because MIPS’s complex catalog does not yet provide a complete picture of the yeast interaction proteome. We also observe that MSPCI seem to have much less associations than the TAP (e.g., 187 versus 3,436 when cut-off is 0.9). Because there are only 276 overlapped proteins between the 662 and 402 proteins, we believe that the difference is likely to be attributable to the different sub-network structures captured by the two experimental methods.

DISCUSSION

In this work, we developed a statistical model to analyze MPC proteomics data, which are prone to experimental errors. The major contributions of our work include: (i) definition of the true association rate in a given proteome; (ii) integration of both type I and II associations from an MPC experiment into a statistical model that includes all potential protein pairs and accounts for various types of experimental errors. von Mering et al.[21] pointed out that when assessing the quality of interaction data, coverage (e.g., sensitivity) and accuracy (e.g., FDR) need to be considered together. We showed that, compared with the statistical framework that only considers bait–prey associations (type I association), we have the opportunity to discover true protein interactions at significantly higher sensitivity level while controlling FDR at a comparable level. With our assessment results, we believe it is possible to uncover new protein interactions. For example, three positive protein pairs, RPN1 with RPT2, RPN9 and RPN3, are predicted by both TAP and MSPCI data using only prey–prey association information, which cannot be detected by the BB model. All three proteins are parts of the 26S proteasome regulatory subunits; and the association between RPN1 and RPT2 and RPN9 are confirmed by Ferrell et al.[27]

Our definition of “type II false-negative rate” needs some additional clarification. Whether or not two truly associated proteins (A and B) are preys of a particular bait (C) depends on the true association status of C and A (and B). If C occurs in the same complex with A and B, then it is very likely that it can “fish” both A and B. However, if C is not located within the same complex as A and B, then it is

TABLE III. Summary of Positive Protein Pairs Defined by Different Cut-Off Points for the Overlapped Proteins Between TAP and MP (662) and Those Between MSPCI and MP (402)

Cut-off	No. of 1B	No. (%) in MP	No. of PP	No. (%) in MP	Total	No. (%) in MP
Overlap of TAP and MP: 662 proteins (662 × 661/2 = 18,915 1B pairs + 108,811 PP pairs)						
0.9	897	334 (37.2)	2,539	**1,196 (47.1)**	3,436	1,530 (44.5)
0.5	1,778	511 (28.7)	4,218	**1,748 (41.4)**	5,996	2,259 (37.7)
0.1	1,778	511 (28.7)	10,007	**2,713 (27.1)**	11,785	3,224 (27.3)
Overlap of MSPCI and MP: 402 proteins (402 × 401/2 = 10,731 1B pairs + 32,385 PP pairs)						
0.9	109	65 (59.6)	78	**52 (66.7)**	187	117 (62.6)
0.5	301	131 (43.5)	250	**126 (50.4)**	551	257 (46.6)
0.1	574	193 (33.6)	1,547	**286 (18.5)**	2,121	479 (22.6)

1B, protein pairs with at least one member ever serving as the bait; PP, protein pairs with no member ever serving as the bait.

likely that C will "fish" neither A nor B, which is not attributable to experiment error. Hence, although we use the term "type II false-negative rate," it actually reflects a "combined" effect of experimental error and presence/absence of proteins that are actually associated with A and B.

In this work, we are primarily focused on the global picture of protein pairs within the same complex. The next level of investigation is direct physical contact, by which proteins form complexes. We plan to extend our model to estimate the probability of any proteins physically interacting with each other given an MPC proteomics data set. Unlike yeast two-hybrid high-throughput data that provide direct information on physical interaction of two proteins, mass spectrometry of purified complexes only presents us "hidden" information on physical interactions by disclosing the result of the physical interactions: protein complexes. Therefore, we plan to construct a more complex statistical model that allows us to estimate the likelihood that two proteins interact physically given the data. We expect the computation burden of such a model to be fairly intensive because of the huge space of physical interaction patterns within a proteome. Nonetheless, the investigation for its theoretical and numerical feasibility will help discover new protein interaction relationships in the new wave of massive protein interaction proteome data to arrive.

ACKNOWLEDGMENTS

This work was supported in part by systems obtained by Indiana University through its relationship with Sun Microsystems Inc. as a Sun Center of Excellence. The authors thank Dr. A. Gavin for kindly providing the TAP raw data, Manjula Aliminati for processing and integrating the raw data into the Oracle 9i relational databases, and Stephanie Burks for maintaining the high-end Sun servers and Oracle 9i servers. The authors also thank Zhiping Wang for his help in extracting the MIPS data sets.

REFERENCES

1. Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res 2002;30(1):38–41.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2002;30(1):17–20.
3. Chen JY, Sivachenko A. Data mining challenges for protein interactomics studies. IEEE Eng Med Biol Mag 2005;24(3):95–102.
4. Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature 1989;340(6230):245–246.
5. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000;403(6770):623–627.
6. 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 USA 2001;98(8):4569–4574.
7. Chen JY, Sivachenko AY, Bell R, Kurschner C, Ota I, Sahasrabudhe S. Initial large-scale exploration of protein-protein interactions in human brain. Proc IEEE Comput Soc Bioinform Conf 2003;2:229–234.
8. Gavin AC, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415(6868):141–147.
9. Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 2002;415(6868):180–183.
10. Tong AH, Evangelista M, Parsons AB, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 2001;294(5550):2364–2368.
11. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet 2001;29(4):482–486.
12. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402(6757):86–90.
13. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285(5428):751–753.
14. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23(9):324–328.
15. Gomez SM, Lo SH, Rzhetsky A. Probabilistic prediction of unknown metabolic and signal-transduction networks. Genetics 2001;159(3):1291–1298.
16. Gomez SM, Noble WS, Rzhetsky A. Learning to predict protein-protein interactions from protein sequences. Bioinformatics 2003;19(15):1875–1881.
17. Chen JY. High-throughput protein interactome data: minable or not? Seattle, WA; 2004. p 18–23.
18. Liu Y, Liu N, Zhao H. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 2005;21(15):3279–3285.
19. Fletcher S, Bowden SE, Marrion NV. False interaction of syntaxin 1A with a Ca(2+)-activated K(+) channel revealed by co-immunoprecipitation and pull-down assays: implications for identification of protein-protein interactions. Neuropharmacology 2003;44(6):817–827.
20. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 2002;18(10):529–536.
21. 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;417(6887):399–403.

22. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4(1):2.
23. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol 2003;327(5):919–923.
24. Gilchrist MA, Salter LA, Wagner A. A statistical framework for combining and interpreting proteomic datasets. Bioinformatics 2004;20:689–700.
25. Yarmush ML, Jayaraman A. Advances in proteomic technologies. Annu Rev Biomed Eng 2002;4:349–373.
26. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 1977;39:1–38.
27. Ferrell K, Wilkinson CR, Dubiel W, Gordon C. Regulatory subunit interactions of the 26S proteasome, a complex problem. Trends Biochem Sci 2000;25(2):83–88.

APPENDIX

Details of EM Algorithm

Let $\theta^{(m-1)}$ be the parameter estimate from iteration $m-1$. At iteration m, define the following quantities:

$$\alpha_0^{(m-1)} = \ln\frac{(1-s_0^{(m-1)})(1-r_0^{(m-1)})}{r_0^{(m-1)}s_0^{(m-1)}},\quad \gamma_0^{(m-1)} = \ln\frac{r_0^{(m-1)}}{(1-r_0^{(m-1)})},\quad \beta_0^{(m-1)} = \ln\frac{s_0^{(m-1)}}{(1-r_0^{(m-1)})}$$

$$\alpha_1^{(m-1)} = \ln\frac{(1-s_1^{(m-1)})(1-r_1^{(m-1)})}{r_1^{(m-1)}s_1^{(m-1)}},\quad \gamma_1^{(m-1)} = \ln\frac{r_1^{(m-1)}}{(1-r_1^{(m-1)})},\quad \beta_1^{(m-1)} = \ln\frac{s_1^{(m-1)}}{(1-r_1^{(m-1)})}$$

$$\lambda^{(m-1)} = \ln\frac{\rho^{(m-1)}}{(1-\rho^{(m-1)})};$$

$$K_{ij} = \begin{cases} \sum_{t=1}^{n_i} Y_{ij}^{it} + \sum_{t=1}^{n_j} Y_{ji}^{jt}, & \text{if } i<j\le n \\ \sum_{t=1}^{n_i} Y_{ij}^{it}, & \text{if } i\le n, j>n \\ 0, & \text{if } n<i<j \end{cases},\qquad Q_{ij} = \begin{cases} n_i + n_j, & \text{if } i<j\le n \\ n_i, & \text{if } i\le n, j>n \\ 0, & \text{if } n<i<j \end{cases}$$

$$T_{ij} = \begin{cases} \sum_{h\ne i,j;h\le n}\sum_{t=1}^{n_h} Y_{ij}^{ht}, & \text{if } i<j\le n \\ \sum_{h\ne i;h\le n}\sum_{t=1}^{n_h} Y_{ij}^{ht}, & \text{if } i\le n, j>n \\ \sum_{h=1}^{n}\sum_{t=1}^{n_h} Y_{ij}^{ht}, & \text{if } n<i<j \end{cases},\qquad S_{ij} = \begin{cases} \sum_{h\ne i,j;h\le n} n_h, & \text{if } i<j\le n \\ \sum_{h\ne i;h\le n} n_h, & \text{if } i\le n, j>n; \\ \sum_{h=1}^{n} n_h, & \text{if } n<i<j \end{cases}$$

$$W_{ij}^{(m-1)} = \alpha_0^{(m-1)}K_{ij} + \beta_0^{(m-1)}Q_{ij} + \alpha_1^{(m-1)}T_{ij} + \beta_1^{(m-1)}S_{ij} + \lambda^{(m-1)}.$$

Then we have
E step:

$$p_{ij}^{(m)} = E[Z_{ij}=1|Y,\theta^{(m-1)}] = Pr[Z_{ij}=1|Y,\theta^{(m-1)}] = \frac{1}{1+e^{-W_{ij}^{(m-1)}}}$$

M step:

$$\rho^{(m)} = \frac{2}{N(N-1)}\sum_{i<j} p_{ij}^{(m)},\quad r_0^{(m)} = \frac{\sum_{k=1}^{n}\sum_{t=1}^{n_k}\sum_{j\ne k} Y_{kj}^{kt} - \sum_{i<j} p_{ij}^{(m)}K_{ij}}{(N-1)\sum_{k=1}^{n} n_k - \sum_{i<j} p_{ij}^{(m)}Q_{ij}},\quad s_0^{(m)} = \frac{\sum_{i<j} p_{ij}^{(m)}(Q_{ij}-K_{ij})}{\sum_{i<j} p_{ij}^{(m)}Q_{ij}}$$

$$r_1^{(m)} = \frac{\sum_{k=1}^{n}\sum_{t=1}^{n_k}\sum_{i,j\ne k} Y_{ij}^{kt} - \sum_{i<j} p_{ij}^{(m)}T_{ij}}{(N-1)(N/2-1)\sum_{k=1}^{n} n_k - \sum_{i<j} p_{ij}^{(m)}S_{ij}},\quad s_1^{(m)} = \frac{\sum_{i<j} p_{ij}^{(m)}(S_{ij}-T_{ij})}{\sum_{i<j} p_{ij}^{(m)}S_{ij}}.$$