Discover True Association Rates In Multi-protein Complex Proteomics Data Sets

Changyu Shen
Division of Biostatistics
Department of Medicine
School of Medicine
Indiana University
Indianapolis, IN 46202,USA
chashen@iupui.edu

Lang Li
Division of Biostatistics
Department of Medicine
School of Medicine
Indiana University
Indianapolis, IN 46202,USA
lali@iupui.edu

Jake Yue Chen*
[1]School of Informatics
Indiana University
[2]Department of Computer and Information Science,
Purdue University School of Science
Indianapolis, IN 46202,USA
jakechen@iupui.edu

Abstract

Experimental processes to collect and process proteomics data are increasingly complex, while 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 a complete empirical Bayes model to analyze multi-protein complex (MPC) proteomics data derived from peptide mass spectrometry detections of purified protein complex pull-down experiments. Our model considers not only bait-prey associations, but also prey-prey associations missed in previous work. Using our model and a yeast MPC proteomics data set, we estimated that there should be an average of 28 true associations per MPC, almost ten 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%.

1. 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 post-transcriptional 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, on the other hand, provides biologists with molecular binding information between proteins. This information enables computational scientists to build computer models of protein complexes and molecular pathways, which 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 due to 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 categorized into four broad categories of experimental 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) **multi-protein complex (MPC)** experimental methods, coupled with a series of protein complex purification, separation, and identification methods often involving liquid

* To whom correspondence should be sent

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 co-evolution patterns found in different genomes, or when there is conserved mRNA co-expression patterns under a variety of controlled stimulatory conditions [11]. Other approaches also exist, including [12-14]. Note that only Y2H and MPC methods provide direct evidence of physical protein-protein interactions. In recent related work, we presented several ideas on how to assess and analyze Y2H data using frequentists' statistical methods [15]. In this paper, we want to concentrate on the study of interaction-based proteomics data from MPC methods.

The general strategy of MPC can be described as follows. First, a pre-selected protein (called "**bait**") is modified to have a "tag" peptide inserted into the protein's 3'-terminus using DNA recombination. The DNA vector containing encoded tagged bait proteins is subsequently introduced into target expression cells. Next, the bait proteins are profusely expressed in target expression cells, harvested, purified, and affixed to solid-state surfaces through protein tags. These bait proteins are used to "pull down", or associate by protein affinity, all the protein extracts from cell lysates eluting through the solid-state media. The transient protein complexes formed from this "pull down of all proteins" are therefore called Multi-protein Complexes (**MPC**, as abbreviated previously). Since each MPC may contain hundreds of associated proteins (called "**preys**"), it needs to go through careful protein separation procedures such as 2-D gel electrophoresis or liquid chromatography until each separated aliquot contains much smaller number of possible types of proteins. Finally, peptide Mass Spectrometers (**MS**) are used to determine the peptide constitutes in each aliquot. Bioinformatics data analysis tools, such as SEQUEST and MASCOT, are available to identify the proteins that these peptides come from.

It is not surprising to note that a complex method such as MPC could be subject to many sources of experimental errors. This has presented itself a huge challenge in the practical use of MPC proteomics data for subsequent biological pathway studies. For example, system errors could be introduced if samples are contaminated; random errors are also unavoidable, since the quality of final prey protein identifications are subject to accurate collection and interpretation of MS peaks. As observed in [16] and [17], 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 a "degree of overlap" method, in which a newly collected MPC 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 [16, 18]. Sprinzak *et al.* [19] used cellular co-localization 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 no 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. We will describe how to develop such an effective model using previously missed information. Surprisingly, using our model and a yeast protein interaction data set from [8], we found that the previous estimate of 2-3 "true protein associations" per MPC experiment (a "**trial**") by Gilchrist *et al*. [20] were off by almost 10 times—our estimate came at approximately 28 "true protein associations". In Section 2, we will first provide some background on Gilchrist *et al*.'s model, followed by concept introduction and model details. We then apply our model to one high-throughput data and validate advantages of our approach in Section 3. Finally, we conclude this paper with a discussion.

2. A statistical framework

2.1. Background

Gilchrist *et al.* [20] 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, **TAP** [8] and **HMS-PCI** [9]. In the study, the authors concluded that both

experiments have a $\rho = 1.88 \times 10^{-3}$. For the TAP data set, the total number of protein pairs under consideration is 6.8×10^5 (see **Section 3**), which suggests 1278 true interacting pairs among the 533 multi-protein complexes based on the estimated global association prior. Therefore, there "should be" on average 2.4 (1278/533) true positive interactions per MPC. This result seems alarmingly low! If it were true, this estimate could invalidate the majority of today's MPC experiments, which often are known to contain dozens up to hundreds of proteins in our experience. Yarmush *et al*. [21] also pointed out recently that a bait protein in an MPC experiment is generally associated with more than 50 proteins in yeast.

What could have gone wrong? As we soon describe, we believe the fundamental cause of such a low estimate of true positives for each MPC experiment lies in the fact that a lot of protein-protein interaction information among prey proteins was ignored by 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 [18]. This over-simplification translates into a highly restrictive (likely incorrect) biological assumption, which states that all prey proteins must be directly associated with the bait protein (no secondary/indirect associations among prey proteins are allowed). For the rest of the paper, we will explain these concepts in detail, and present a complete empirical Bayes model that includes all protein-protein interaction information embedded within each MPC experiment.

2.2. Concepts

Similar to the concept of "global association prior" (ρ), we introduce the concepts of true association and true association rate (ρ) of a *proteome* (the total collection of proteins in a given cell):

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

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

True association cannot be 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 concern 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 concern 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). As shown later, type II association fills in the technology inadequacy that not every protein is used as bait. We illustrate these two concepts in Figure 1. As shown in next section, the BB model only includes type I association and loses a fairly large number of protein-protein interactions embedded within the type II association.

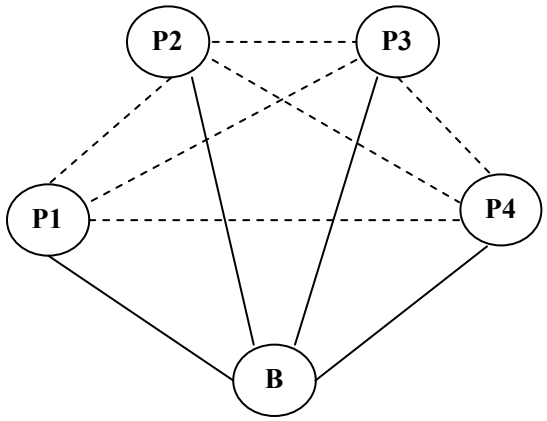

Figure 1. Schematic drawing of a hypothetical multi-protein complex pull-down trial using a bait shown as the node labeled "B". Preys P1, P2, P3, and P4 are also shown as labeled nodes. Solid lines connecting the bait and preys are type I associations and dotted lines connecting preys are type II associations

The imperfection of any experimental technique is due 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[A and B has a type I association | A and B do NOT have a true association]

s_0 = Pr[A and B do NOT have a type I association | A and B have a true association]

r_1 = Pr[A and B have a type II association | A and B do NOT have a true association; C is the bait]

s_1 = Pr[A and B do NOT have a type II association | A and B have a true association; C is the bait].

2.2. A complete empirical Bayes model

The ρ in the BB model is solely based on the type I association and consequently fails to represent the true association rate among all protein pairs within a proteome. 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 $N(N\text{-}1)/2$ protein pairs and identifying those protein pairs. What Gilchrist *et al.* tried to estimate is the true association rate among the $n(n\text{-}1)/2+n(N\text{-}n)$ 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 a true association for the $(N\text{-}n)(N\text{-}n\text{-}1)/2$ protein pairs in the blank area.

Now we introduce a complete empirical Bayes model to account for the information embedded within both type I and type II associations, which enables us to define the true association rate for the total $N(N\text{-}1)/2$ protein pairs and estimate the posterior probability of any two specific proteins having a true association. It is called the **Complete Binomial-Bernoulli** model (CBB) in the sense that it models the association of all possible protein pairs. Parameters in the CBB are estimated by the Expectation-Maximization (EM) algorithm [22], which automatically provides the posterior probability of two proteins having a true association at convergence.

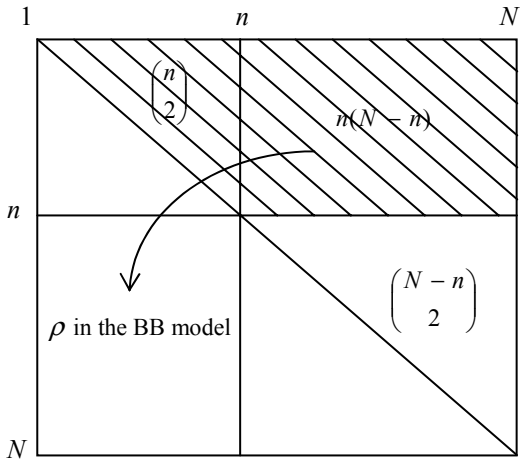

Figure 2. Illustration of the space of interacting protein pairs. Proteins 1 to n serve as bait proteins and proteins n+1 to N only appear in the preys. The shadowed area refers to the total protein pairs under Gilchrist *et al.*'s consideration, upon which ρ is defined.

We will use ρ to denote the true association rate, which is equal to the probability that two proteins randomly selected from a proteome of N proteins have a true association. Practically, N is defined to be the union of the bait proteins and their preys. To write out the likelihood function, we need some extra notations. 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,…,n; i, j=1, 2,…, N ($i \neq j$) and t=1, 2…, n_k (n_k is the number of trials with bait k), define

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

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

Moreover, let $Y^{kt} = (Y_{ij}^{kt}; i < j, i, j = 1,2...N)$ be a binary vector composed of the status of the (N-1) type I associations and (N-1)(N-2)/2 type II associations for the t^{th} trial of bait k, $Y = (Y^{kt}; k = 1,2...n, t = 1,2...n_k)$ be the observed data, and $Z = (Z_{ij}, i < j, i, j = 1,2...N)$ 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

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

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

where $\theta = (\rho, r_0, s_0, r_1, s_1)$. Note that the first product in (1) involves type I associations and the second product involves type II associations. The log-likelihood function is

$$\begin{aligned} & l_{kt}(\theta; Y^{kt} \mid Z) \\ &= \ln(L_{kt}(Y^{kt} \mid Z, \theta)) \\ &= \sum_{j \neq k}^{N} Y_{kj}^{kt} [Z_{kj} \ln(1-s_0) + (1-Z_{kj}) \ln r_0] \\ &\quad + (1 - Y_{kj}^{kt})[Z_{kj} \ln s_0 + (1-Z_{kj}) \ln(1-r_0)] \quad (2) \\ &\quad + \sum_{\substack{i<j, i \neq k \\ j \neq k}}^{N} Y_{ij}^{kt} [Z_{ij} \ln(1-s_1) + (1-Z_{ij}) \ln r_1] \\ &\quad + (1 - Y_{ij}^{kt})[Z_{ij} \ln s_1 + (1-Z_{ij}) \ln(1-r_1)]. \end{aligned}$$

The model and corresponding likelihood function of Z can be written as

$$L(Z \mid \theta) = \rho^{\sum_{i<j}^{N} Z_{ij}} (1-\rho)^{\sum_{i<j}^{N} (1-Z_{ij})};$$

$$l(\theta; Z) = \ln \rho \sum_{i<j}^{N} Z_{ij} + \ln(1-\rho) \sum_{i<j}^{N} (1-Z_{ij}).$$

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

$$\begin{aligned} L(Y, Z \mid \theta) &= L(Z \mid \theta) L(Y \mid Z, \theta) \\ &= L(Z \mid \theta) \prod_{k=1}^{n} \prod_{t=1}^{n_k} L_{kt}(Y^{kt} \mid Z, \theta), \end{aligned}$$

and the corresponding log-likelihood function is

$$l(\theta; Y, Z) = l(\theta; Z) + \sum_{k=1}^{n} \sum_{t=1}^{n_k} l_{kt}(\theta; Y^{kt} \mid Z). \quad (3)$$

Since we do not observe Z, it is treated as missing data in our EM algorithm. This algorithm is composed of two steps: the Expectation step (E) and the Maximization (M) step. During the E step of the m^{th} iteration, Z is updated by the conditional expectation given the estimate of θ from last iteration ($\theta^{(m-1)}$) and Y, that is, $Z^{(m)} = E[Z \mid Y, \theta^{(m-1)}]$; then in the M step, we find $\theta^{(m)}$ that maximizes $l(\theta; Y, Z^{(m)})$. This procedures is repeated until convergence. 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, $\Pr[Z_{ij}=1 \mid Y]$.

3. Analysis of MPC proteomics data

3.1. Analysis of a high-throughput data set

We applied the proposed model (CBB) to the study by Gavin *et al*. [8], in which high-throughput protein complex data sets for yeast *Saccharomyces cerevisiae* were generated by tandem affinity purification (TAP). Gavin *et al.* processed 1739 genes and ultimately purified protein assemblies that cover 1550 proteins. Among the 1550 proteins, 533 serve as the bait protein once and 1017 of them only present themselves as preys. Hence, Gilchrist *et al*. only considered the $533 \times 532/2 + 533 \times 1017 = 6.8 \times 10^5$ protein pairs, which is 57% of the total number of pairs formed by the 1550 proteins (1.2×10^6).

In Table 1, we compare the parameter estimates from the BB model and the CBB model just described. Clearly, estimate of the true association rate from the CBB is much higher than that from the BB, which indicates that a large amount of true associations within the 1017 proteins that never serve as the bait proteins have been ignored by the BB. Essentially, the CBB postulates that there are $1.2 \times 10^6 \times 1.38 \times 10^{-2} = 16560$ true associations, which suggests that on average about 28 (16560/533)

true associations are identified for each MPC experimental 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.

Table 1. Parameter estimates from BB [20] and CBB for the high-throughput experiment in [8]; ρ: true association rate, r_0: type I false positive rate, s_0: type I false negative rate

model	ρ	r_0	s_0
BB	1.88×10^{-3}	1.07×10^{-3}	0.346
CBB	1.38×10^{-2}	5.44×10^{-3}	0.588

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 1000 pairs that do NOT have a true association, 5 of them will have a type I association when one member of the pair serves as the bait; and for every 2 pairs that have a true association, 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 5×10^{-5} and 0.993, 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. On the other hand, there is 0.7% probability for two truly associated proteins to be "fished" by a third protein, mainly due to few bait proteins that truly interact with the two proteins of interest (see **Section 4**).

3.2. Statistical validation

To identify which pairs have a true association, a routine practice is to apply a cutoff point to the posterior probability of the Z_{ij} being equal to 1. For example, if we use 0.8 as the cutoff point, 15560 positive pairs (1.3%) will be identified based on the CBB model. Then we are interested in what proportion of the true associations is covered in these 15560 positive pairs (the sensitivity or SEN) and what proportion of these 15560 positive pairs actually do not have true associations (false discovery rate or FDR). SEN and FDR provide measurements on the prediction quality. Clearly, a model with lower FDR and higher SEN is desirable. Usually, one index is improved at the price of the other one by applying different cutoff points. FDR and SEN can be readily estimated when we know the true association status of each protein pair. Therefore, we generate a hypothetical true association map that mimics the cluster structure in *Saccharomyces cerevisiae* as demonstrated by Gavin *et al.* [8]. We will assume that we know the true association status for every pair of protein and compare the FDR and SEN of the CBB with that of the BB. The data generation process can be divided into two steps. First, we generate the data of the true association status of a proteome (the "unobserved" Z data). Second, we generate the MPC data based on data from step 1 (the "observed" Y data).

Step1: Gavin *et al.* categorized proteins in their MPC experiment into 232 complexes, while some complexes also share certain proteins. We first postulate a proteome (2660 proteins) composed of 232 complexes with the same distribution of cluster size as Gavin *et al*'s; yet none of the complexes shares any proteins. Based on this proteome, we generate a proteome of 1000 proteins, which possesses similar cluster sizes and the number of clusters with a fixed size is reduced proportionally. In Table 2, we show the distribution of the cluster size of the 79 clusters formed by the 1000 proteins.

Step 2: we set the type I false positive rate and type I false negative rate to be 0.5% and 50%, which mimic what we found for Gavin *et al.*'s experiment. 330 proteins are randomly selected as the bait proteins with each one having one trial to mimic the real data ($330/1000 \approx 533/1550 \approx 1/3$).

Table 2. Distribution of cluster size within a hypothetical proteome of 1000 proteins

Size	#	Size	#	Size	#	Size	#
1	1	9	2	17	1	38	1
2	16	10	2	18	2	39	1
3	19	11	1	19	1	41	1
4	2	12	2	20	1	76	1
5	2	13	1	22	1	83	1
6	6	14	1	30	1	90	1
7	2	15	2	33	1	93	1
8	3	16	1	35	1		

After we obtain the posterior probability of having a true association from the CBB and BB, a pair of proteins are claimed positive if its posterior probability is greater than 0.85. We repeatedly generate 200 data sets and the results are shown in Table 3, where numbers in the parenthesis are the standard errors (S.E.). On average, while the CBB maintains a similar FDR as BB, much more truly

associated protein pairs are identified by the CBB (82.7% vs. 2.7%). Applying smaller cutoff points to BB will not enhance the sensitivity substantially, mainly due to the limited coverage of protein pairs by this approach (Figure 2).

Table 3. Simulation results based on 200 runs for CBB and BB (cutoff=0.85)

model	FDR (S.E.)	SEN (S.E.)
CBB	0.32% (0.12%)	82.7% (1.4%)
BB	0.25% (0.21%)	2.7% (0.2%)

We conducted another study by fixing the bait proteins (330) and applying various cutoff points (0.1 to 0.99, step width=0.01). The bait proteins are chosen so that each cluster has at least one bait protein and larger clusters have more bait proteins. In Figure 3, we show the graph of SEN versus FDR. 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 greater than 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.

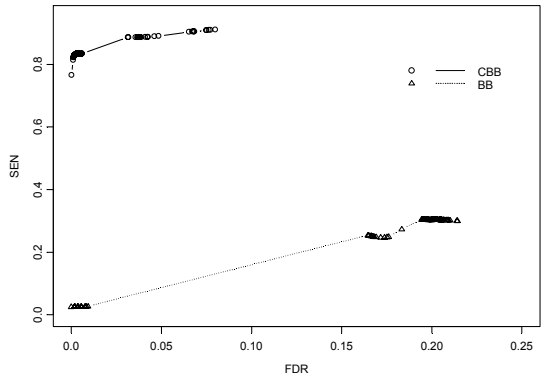

Figure 3. Sensitivity (SEN) versus false discovery rate (FDR) for the CBB and BB in identifying truly associated protein pairs with a fixed set of bait proteins. Different points are obtained by applying different cutoff points (0.1 to 0.99 by 0.01, # of points=90) to the posterior probability of having true association.

4. Discussion

In this work, we developed a complete empirical Bayes model to analyze MPC proteomics data that are prone to random errors. We treat each MPC as an experimental trial to observe bait-prey and prey-prey interactions that are subject to two types of random errors: type I for bait-prey interaction and type II for prey-prey interaction. Maximum likelihood estimation with EM algorithm is utilized to estimate model parameters. We apply our model to one high-throughput data set and obtain an estimate of the number of true positive pairs per MPC that is about 10 times as large as that estimated by Gilchrist *et al.* It clearly demonstrates that a large amount of protein-protein association pairs are recovered by the CBB model by accounting for the prey-prey interactions. Moreover, the validation study further confirms that our model is more powerful in detecting true associations.

The major contributions of our work include: (i) definition of the true association rate of a proteome and its estimate adjusting for random errors; (ii) development of both type I and II associations from an MPC experiment into a statistical model; (iii) assignment of a probability of having true association to each protein pair; and (iv) enhanced sensitivity with fairly small false discovery rate. von Mering e*t al.* [17] pointed out that when assessing the quality of interaction data, coverage and accuracy need to be considered together. Similarly, when analyzing interaction data, a good prediction model should have high sensitivity (coverage) and low FDR (accuracy). Therefore, (iv) is of great significance in terms of prediction quality. We believe both our statistical framework and results will guide future researchers in this domain to extract fruitful knowledge from protein interaction data.

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. On the other hand, if C is not located within the same complex as A and B, then it is likely that C will "fish" neither A nor B, which is not due to experiment error. Hence, although we use the term "type II false negative rate", it does not necessarily reflect experiment error exclusively. Ideally, we can define two distinct s_1 for the scenario when C has a true association with A and B and the scenario when C does not. However, this will seriously complicate the likelihood function that is used to estimate those parameters. Hence we use only one s_1 to indicate an averaged effect, that is, a randomly selected protein C is used as the bait. Since only a small proportion of the whole proteome have true associations with the two proteins of interest, s_1 usually is quite close to 1.

We plan to extend this 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 information on direct physical interaction of two proteins, mass spectrometry of purified complexes only presents us information on whether or not two proteins are located within the same complex. Therefore, we cannot tell directly from the data whether an observed association indicates a real physical contact, even when there are no random errors. Nevertheless, we can construct a more delicate statistical model that allows us to estimate the likelihood that two proteins interact directly given the data. Certainly, a multi-protein complex is the result of the physical contact of relevant proteins. Suppose two randomly selected proteins have probability p to interact directly and let W_{ij} be the direct interaction indicator of protein i and j. Then Z_{ij} (true association indicator) is entirely determined by the W_{ij}'s. Therefore, we can replace ρ and Z_{ij} in our current model with p and W_{ij}, respectively. However, such a model is much more complicated than the current one and the computation can be very intensive. Hence, it would require further investigation for its theoretical and numerical feasibility.

Acknowledgement

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. We 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, on which the computing in this study was partially conducted.

References

[1] T. Hubbard, *et al.*, "The Ensembl genome database project", *Nucleic Acids Res*, **30**(1), 2002, pp. 38-41.

[2] D. A. Benson, *et al.*, "GenBank", *Nucleic Acids Res*, **30**(1), 2002, pp. 17-20.

[3] J. Y. Chen and A. Sivachenko, "Data mining challenges for protein interactomics studies", *IEEE Magazine in Biology and Medicine*, 2005 (In press).

[4] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions", *Nature*, **340**(6230), 1989, pp. 245-6.

[5] P. Uetz, *et al.*, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae", *Nature*, **403**(6770), 2000, pp. 623-7.

[6] T. Ito, *et al.*, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proc Natl Acad Sci U S A*, **98**(8), 2001, pp. 4569-74.

[7] J. Y. Chen, *et al.*, "Initial large-scale exploration of protein-protein interactions in human brain", presented at Proceedings of IEEE Computational Systems Biology (CSB), Stanford, CA, 2003.

[8] A. C. Gavin, *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes", *Nature*, **415**(6868), 2002, pp. 141-7.

[9] Y. Ho, *et al.*, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry", *Nature*, **415**(6868), 2002, pp. 180-3.

[10] A. H. Tong, *et al.*, "Systematic genetic analysis with ordered arrays of yeast deletion mutants", *Science*, **294**(5550), 2001, pp. 2364-8.

[11] H. Ge, *et al.*, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae", *Nat Genet*, **29**(4), 2001, pp. 482-6.

[12] A. J. Enright, *et al.*, "Protein interaction maps for complete genomes based on gene fusion events", *Nature*, **402**(6757), 1999, pp. 86-90.

[13] E. M. Marcotte, *et al.*, "Detecting protein function and protein-protein interactions from genome sequences", *Science*, **285**(5428), 1999, pp. 751-3.

[14] T. Dandekar, *et al.*, "Conservation of gene order: a fingerprint of proteins that physically interact", *Trends Biochem Sci*, **23**(9), 1998, pp. 324-8.

[15] J. Y. Chen, "High-throughput Protein Interactome Data: Minable or Not?" presented at Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics (BioKDD 2004) at the International Conference on Knowledge Discovery and Data Mining, Seattle, WA, 2004.

[16] A. M. Edwards, *et al.*, "Bridging structural biology and genomics: assessing protein interaction data with known complexes", *Trends Genet*, **18**(10), 2002, pp. 529-36.

[17] C. von Mering, *et al.*, "Comparative assessment of large-scale data sets of protein-protein interactions", *Nature*, **417**(6887), 2002, pp. 399-403.

[18] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks", *BMC Bioinformatics*, **4**(1), 2003, pp. 2.

[19] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein-protein interaction data?" *J Mol Biol*, **327**(5), 2003, pp. 919-23.

[20] M. A. Gilchrist, L. A. Salter, and A. Wagner, "A statistical framework for combining and interpreting proteomic datasets", *Bioinformatics*, **20**, 2004, pp. 689-700.

[21] M. L. Yarmush and A. Jayaraman, "Advances in proteomic technologies", *Annu Rev Biomed Eng*, **4**, 2002, pp. 349-73.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society series B.*, **39**, 1977, pp. 1-38.