Probabilistic and Likelihood-Based Methods for Protein Identification from MS/MS Data

Ryan Gill and Susmita Datta

1 Introduction

The main goal of proteomic studies is to detect biomarker proteins for the early detection of cancer. Tandem mass spectrometry (MS/MS) plays a significant role in the discovery of these biomarker proteins. The first step of an experiment using tandem mass spectrometry (MS/MS) is the digestion of a mixture of proteins by an enzyme, often trypsin. Each of the proteins is separated into peptides which are subsequently ionized. Then selected peptide ions are fragmented in the gas phase, and the mass-to-charge ratios and abundances or intensities of the small fragmented ions are recorded in an MS/MS spectra.

Technological improvements have led to a greater abundance of tandem mass spectrometry data as well as an increase in the size of generated data sets [1]. Consequently, it is not feasible to manually attempt to identify the peptides present in the sample, and hence software tools are needed to perform this task. SEQUEST [2] is a popular software tool which uses the precursor ion mass for each observed spectrum to find candidate peptides from a database of protein sequences which are sufficiently close in mass to the spectrum. Each observed spectrum is preprocessed by finding the highest intensity peaks in each of a set of pre-specified bins and normalizing those values to obtain an observed n-dimensional vector u. The theoretical spectrum, denoted by v, is also computed for each of the candidate peptides, and each theoretical spectrum is then preprocessed the same way that the

R. Gill (✉)
Department of Mathematics, University of Louisville, Louisville, KY 40292, USA
e-mail: ryan.gill@louisville.edu

S. Datta
Professor, Department of Biostatistics, University of Florida, Gainesville, FL, USA
e-mail: susmita.datta@ufl.edu

S. Datta, B.J.A. Mertens (eds.), *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-45809-0_4

observed spectrum was preprocessed. Then, each observed spectrum is compared with each theoretical spectrum in its candidate list by a preliminary score S_p based on the number of predicted fragment ions that match ions in the spectrum and their abundances as well as the number of predicted sequence ions. Finally, a further score

$$Xcorr = R_0 - \frac{1}{151}\sum_{\tau=-75}^{75} R_\tau$$

is computed for each of the top 500 candidate spectra where $R_\tau = \sum u_i v_{i+\tau}$ is the discrete cross-correlation with lag τ. Here R_0 is the scalar dot product between the observed and theoretical spectra. As described in [2] and [3], *Xcorr* gives a measure of spatial similarity to assess the coherence of the observed and each theoretical spectra by not only computing R_0 but also by including a correction factor to account for background correlation between the observed and theoretical spectra by using offset values. The highest *Xcorr* score is reported by SEQUEST as a peptide-spectrum match (PSM). SEQUEST is a commercial program, but there are alternate implementations of the original SEQUEST algorithm such as Crux [3] and Tide [4] which are freely available and which also reportedly lead to drastic increase in speed compared with the original SEQUEST algorithm. Software for other database search algorithms such as X!TANDEM [5], Mascot [6], MS-Tag [7], and MS-GF [8–10] are also available.

In spite of the developments of the above-mentioned search algorithms, there still remain uncertainties associated with the peptide and protein identifications. Experimental errors and lack of adequate search algorithms can, sometimes, lead to highly erroneous peptide and protein identifications from a tandem mass spectrometry experiment; in fact, without proper filtering, it is possible that 80–90 % of identified proteins may not be correct [11, 12]. The situation becomes more complicated in the presence of "degenerate" peptides. A peptide is referred to as "degenerate" if it is generated by multiple proteins. Degenerate peptides create additional challenges for protein identification because even if the peptide identification were known to be correct with no uncertainty, the identity of the protein that generated it is not clearly determined. A typical situation of degeneracy is explained through Fig. 1 (adapted from figures in [13–15]).

Figure 1 summarizes the steps in an MS/MS experiment for three proteins P_1, P_2, and P_3 in red, green, and blue, respectively; each of these proteins generates two peptides: P_1 generates p_1 and p_2, P_2 generates p_3 and p_4, and P_3 generates p_2 and p_5. Since p_2 is generated by two distinct proteins P_1 and P_3, it is referred to as a degenerate peptide. Note that only some peptide ions are selected for fragmentation; some might be selected multiple times like peptide p_1 in Fig. 1, while others might not be selected at all, like peptides p_3 and p_5. Of course, errors can occur during peptide identification; in Fig. 1, peptide p_4 is misidentified as p_x. This also leads to the incorrect conclusion that protein P_x is present in the sample, and the incorrect decision that P_2 is not present because of the misidentification of peptide p_4 and the fact that p_3 is not sampled. Degeneracy of peptides can also be an issue at the database search stage as the example illustrates with protein P_y; if an algorithm

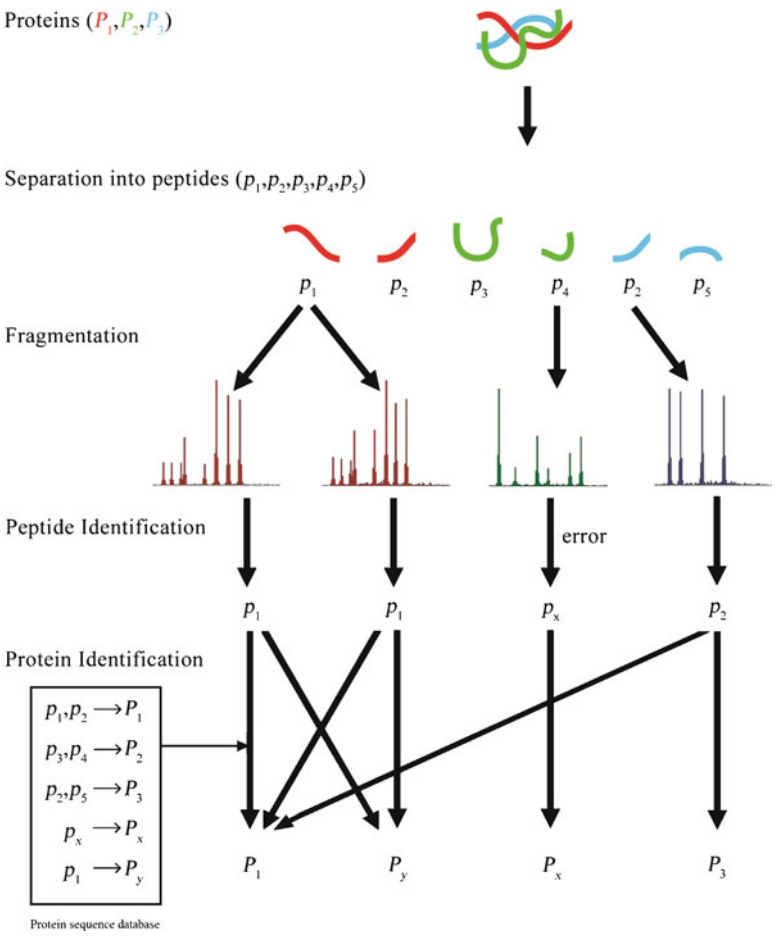

Fig. 1 The steps of an MS/MS experiment for the identification of peptides and proteins by database search methods. This figure is adapted from several sources [13–15]

includes proteins for which its peptides are present, then protein P_y is incorrectly identified as present since its peptide p_1 is correctly identified. The proteins P_1 and P_3 are identified, though there may be concerns about these conclusions since other proteins generate each of the peptides used to identify the proteins.

In recent years, several efforts have been made for providing confidences to the peptide and protein identification. This paper first describes PeptideProphet and ProteinProphet, among the first and still among the most popular probabilistic two-step methods for peptide and protein identification. Then we discuss a couple of likelihood-based one-step methods (hierarchical statistical model (HSM) and nested mixture model (NMM)) which attempt to improve on some of the weaknesses of two-step procedures. Finally, we compare the methods reviewed in this paper in the Discussion section.

2 Two-Step Process

In this section, we describe a highly regarded two-step process. In this process, peptides are identified first using database search scores such as those provided by SEQUEST through an algorithm called PeptideProphet. Then the resulting identified peptides along with their confidences can be used to attempt to determine the proteins which are present in the sample using an algorithm called ProteinProphet. A brief description of each of these two algorithms is given below.

2.1 PeptideProphet

Suppose that $x_1, \ldots, x_s$ are database search scores for a peptide. In [16], four SEQUEST search scores are used:

1. $Xcorr' = \begin{cases} \frac{\ln(Xcorr)}{\ln(N_L)} & \text{if } L < L_c \\ \frac{\ln(Xcorr)}{\ln(N_C)} & \text{if } L \geq L_c \end{cases}$, where L is the number of amino acids in the peptide, N_L is the expected number of fragment ions for a peptide with L amino acids, L_c is a threshold beyond which *Xcorr* does not depend on the length, and N_C is the expected number of fragment ions for a peptide with more than L_C amino acids.
2. ΔC_n, the relative difference between the two highest *Xcorr* scores.
3. ln(SpRank), the natural logarithm of the rank of the preliminary score S_p.
4. d_M, the absolute difference of the masses of the precursor ions for the spectrum and the assigned peptide.

Linear discriminant analysis can be used to combine these search scores into a single score

$$F(x_1, \ldots, x_s) = c_0 + \sum_{i=1}^{s} c_i x_i$$

where the weights $c_0, c_1, \ldots, c_s$ are selected to optimize the probability of a correct classification of the peptide assignments as being correct or incorrect based on training data where it is known which peptide assignments are correct. See Chapter 4 of [17] for details on linear discriminant analysis as well as other linear and nonlinear methods for classifying categorical data.

This score is used to help compute the probability that a peptide score is correct. Let $F|+$ and $F|-$ be the distributions of the discriminant scores for correct and incorrect peptide assignments, respectively. The number of tryptic termini (NTT) also provides useful information regarding the probability that the peptide score is correct as discussed in [16], so let $G|+$ and $G|-$ denote the NTT distributions for

correct and incorrect peptide assignments, respectively. Then the probability that the peptide assignment is correct given the discriminant scores can be computed using Bayes' formula

$$\begin{aligned} p\left(+\middle| F, G\right) &= \frac{p\left(F, G\middle|+\right) p\left(+\right)}{p\left(F, G\middle|+\right) p\left(+\right) + p\left(F, G\middle|-\right) p\left(-\right)} \\ &= \frac{p\left(F\middle|+\right) p\left(G\middle|+\right) p\left(+\right)}{p\left(F\middle|+\right) p\left(G\middle|+\right) p\left(+\right) + p\left(F\middle|-\right) p\left(G\middle|-\right) p\left(-\right)} \end{aligned} \tag{1}$$

where $p\left(+\right)$ and $p\left(-\right)$ are prior probabilities of correct and incorrect peptide assignments, respectively. The prior probabilities used in [16] are the observed proportions in the training data, and the conditional distributions of the discriminant scores for the correct and incorrect assignments are modeled by a Gaussian distribution with estimated mean μ and variance σ^2 and by shifted Gamma distribution with estimated shape, scale, and location parameters, respectively. In (1), it is assumed that the discriminant scores and NTT distributions are independent when conditioned on the peptide assignment status; empirical evidence is provided in [16] to support this assumption.

An alternative to directly using the observed proportions from the training data as the prior probabilities in (1) is to use a mixture model which simultaneously estimates the prior and conditional probabilities based on a two-step iterative process using the expectation-maximization (EM) algorithm [18]. Let N be the number of spectra in the data set. Starting with initial estimates of $p\left(+\right)$, $p\left(-\right)$, $p\left(F\middle|+\right)$, $p\left(F\middle|-\right)$, $p\left(G\middle|+\right)$, and $p\left(G\middle|-\right)$, the first step of each iteration of the EM algorithm computes estimates of $p\left(+\middle| F,\ G\right)$ based on (1). The second step of each iteration updates the estimates of $p\left(+\right)$, $p\left(-\right)$, $p\left(F\middle|+\right)$, $p\left(F\middle|-\right)$, $p\left(G\middle|+\right)$, and $p\left(G\middle|-\right)$ under the assumption that the contribution of each of the N spectra to the distribution of correct/incorrect peptide assignments is proportional to the current computed probability that it is correctly/incorrectly assigned. Specifically, the prior probabilities are updated by the formulas

$$p\left(+\right) = \sum_{i=1}^{N} p\left(+\middle| F_i,\ G_i\right)$$

and $p\left(-\right) = 1 - p\left(+\right)$ where F_i and G_i refer to the respective values for the ith spectrum. The parameters of the conditional distributions are computed using estimates reweighted using weights proportional to the probability of a correct/incorrect peptide assignment conditioned on the spectrum; for the Gaussian distribution

which models the distribution of the discriminant scores for positive peptide assignments, the estimated parameters are $\mu = \sum_{i=1}^{N} p\left(+\middle| F_i,\ G_i\right) F_i / \left(Np\left(+\right)\right)$ and $\sigma^2 = \sum_{i=1}^{N} p\left(+\middle| F_i,\ G_i\right) \left(F_i - \mu\right)^2 / \left(Np\left(+\right)\right)$.

2.2 *ProteinProphet*

Once estimates of the probabilities of the peptide assignments are obtained, then the goal is to estimate the probability that a protein is present in the sample. The probabilities of peptide assignments need not be made using PeptideProphet, or even a database search method, but it is only reasonable to expect that the estimates of the probability of the presence of the protein might be good if the estimates of the probabilities for the peptide assignments are good. Temporarily ignore the possibility of degenerate peptides for which there are multiple corresponding proteins. Let D_i^j includes peptide information for the jth assignment to the ith peptide—such as discriminant scores and number of tryptic termini as used in (1). If the independence of events for peptide assignments is assumed, then the formula

$$\widetilde{P} = 1 - \prod_i \prod_j \left(1 - p\left(+\middle| D_i^j\right)\right)$$

gives the probability that at least one peptide assignment corresponding to the protein is correct.

Instead, ProteinProphet [13] makes the conservative estimate that, for each peptide, the probability of all assignments being incorrect is equal to the minimum (one minus the maximum) of the probabilities of incorrect assignments among all assignments. Hence, the inside product in the formula for $\widetilde{P}$ is replaced by $1 - \max_j p\left(+\middle| D_i^j\right)$ and, ProteinProphet [13] estimates the probability that a protein is present in the sample using the formula

$$P = 1 - \prod_i \left\{1 - \max_j p\left(+\middle| D_i^j\right)\right\}. \tag{2}$$

ProteinProphet's use of only the maximum assignment score for each peptide when estimating the protein probabilities may be overoptimistic since high scores for incorrect peptide identifications may occur by chance particularly when the peptides are assigned more than once.

To adjust for multihit proteins (proteins which correspond to multiple correctly assigned peptides), the estimated probabilities in (1) can be modified by conditioning on another random variable, the estimated number of sibling peptides for

each given peptide; it is seen in [13] that correct peptide assignments tend to correspond to multihit proteins, while incorrect peptide assignments are more likely to occur with proteins for which there are no correct peptide assignments. Some further refinements are also proposed as part of the iProphet multi-level models [19] which update the probabilities computed by PeptideProphet conditioning on number of sibling searches, number of replicate searches, number of sibling experiments, number of sibling ions, and number of sibling modifications via Bayes' theorem. Figure 1 in [19] provides a nice figure illustrating the multi-level approach.

The description of ProteinProphet presented in [13] also provides a method for attempting to handle degenerate peptides using the EM algorithm. Let N_s be the number of proteins that the ith peptide is assigned to, and let P_s be the probability that the sth protein is present. Then Eq. (2) is modified so that

$$P_n = 1 - \prod_i \left\{1 - w_i^n \max_j p\left(+\middle| D_i^j\right)\right\}$$

with weights

$$w_i^n = \frac{P_n}{\sum_{s=1}^{N_s} P_s}$$

which give the probability that the ith peptide corresponds to the nth protein. The algorithm begins by using uniform weights and then proceeds by iteratively updating and recomputing the above equations until the values converge.

Since the intensity measurements in the spectra are subject to noise, incorrect peptide identifications will likely lead to incorrect protein identifications. Moreover, using knowledge of probabilities of the presence of proteins in a sample can affect the probabilities that the peptide identification are correct, and it is clear that appropriate inclusion of feedback in modeling peptides and proteins is critical in making good inferences about each. In the following section, two one-step processes are presented which attempt to simultaneously determine the proteins which are present and the peptides which are correctly identified.

3 One-Step Processes

For the two one-step likelihood-based methods (HSM and NMM) reviewed in this section, it is important to note that HSM handles the possibility of degenerate peptides by assuming that a peptide will be in the sample if at least one of the proteins that generate it is present in the sample. On the other hand, NMM does not account for degeneracy, which can cause problems, particularly when estimating the probabilities that proteins are present in complex high-level organisms.

3.1 Hierarchical Statistical Model

The hierarchical statistical model (HSM) proposed in [14] assumes a parametric multilayer joint distribution of five random vectors Y, V, Z, W, and S representing N proteins with at least one peptide hit and M peptides assigned to at least one spectrum.

In this model, $Y = (Y_1, \ldots, Y_N)$ is a vector of indicators for the presence/absence of the proteins in the sample where $Y_i = 1$ indicates that the ith protein is present in the sample. Letting ρ be the probability that a protein is present in the sample, HSM assumes that $Y_1, \ldots, Y_N$ are independent Bernoulli random variables with probability mass function

$$f(y_i) = \rho^{y_i}(1-\rho)^{1-y_i}$$

for $i = 1, \ldots, N$.

The HSM also considers a vector of independent Bernoulli variables $V = (V_1, \ldots, V_N)$ for each protein indicating whether the number of peptide hits for the protein exceeds a specified threshold h; in particular, $V_i = 1$ indicates that the ith protein has more than h peptide hits. Then the probability mass functions for the Bernoulli random variables can be expressed as

$$f\left(v_i \middle| y_i\right) = {\gamma_1}^{y_i v_i}(1-\gamma_1)^{y_i(1-v_i)}{\gamma_0}^{(1-y_i)v_i}(1-\gamma_0)^{(1-y_i)(1-v_i)}$$

where γ_1 and γ_0 are parameters for the Bernoulli distributions in the cases where the protein is present or absent, respectively.

Next, $Z = (Z_1, \ldots, Z_M)$ is a vector of indicators for the presence/absence of the peptides in the sample, and each Z_i is modeled conditionally on Y with specific parameters based on the type and number of cleavages. It is assumed that $Z_j \middle| Y$ follows a Bernoulli distribution with parameters based on the type and number of cleavages contained in a five-dimensional vector of probabilities $\alpha = (\alpha_n, \alpha_s, \alpha_{nn}, \alpha_{ns}, \alpha_{ss})$ where an n in the index of a component of α indicates a non-specific cleavage and an s indicates a specific cleavage (so, for example, if a protein with a constituent peptide that is generated with one non-specific and one specific cleavage, then α_{ns} is the probability $P\left(Z_j = 1 \middle| Y_i = 1\right)$ that the peptide will be present in the sample given that the protein is present in the sample). Letting C_j be the set of proteins that might generate peptide j, the conditional probability mass function for the presence of the jth peptide is

$$f\left(z_j \middle| y\right) = \left(\prod_{i \in C_j}\left(1 - P\left(Z_j = 1 \middle| Y_i = 1\right)\right)^{y_i}\right)^{1-z_j}\left(1 - \prod_{i \in C_j}\left(1 - P\left(Z_j = 1 \middle| Y_i = 1\right)\right)^{y_i}\right)^{z_j}.$$

In the next layer of the HSM model, $W = (W_{11}, \ldots, W_{1T_1}, \cdots, W_{M1}, \ldots, W_{MT_M})$ is a double-indexed vector of indicators of correct assignments of present peptides to a spectrum where $W_{jk} = 1$ indicates that the kth assignment of the jth peptide to a spectrum is correct and T_j is the number of assignments of the jth peptide to a spectrum. Then the conditional probabilities that particular assignments are correct given that the jth peptide is present is assumed to be Bernoulli with probability τ so that the conditional probability mass function of W_{jk} given Z_j is

$$f\left(w_{jk}\middle|z_j\right) = z_j \tau^{w_{jk}} (1-\tau)^{1-w_{jk}}.$$

Finally, $S = (S_{11}, \ldots, S_{1T_1}, \cdots, S_{M1}, \ldots, S_{MT_M})$ is a double-indexed vector of matching scores for each peptide and potential assignment. The HSM also allows the density to be based on an additional factor Q_{jk} and assumes that there are different density functions depending on whether the assignment of the kth assignment of the jth peptide to a spectrum is correct so that

$$f\left(s_{jk}\middle|w_{jk} = w, q_{jk} = q\right) = f_{q,w}\left(s_{jk}; \beta_{qw}\right)$$

for $w = 0, 1$. Combining all of these components of the HSM, the joint density of Y, V, Z, W, and S based on the model is assumed to have the form

$$f(y, z, w, s, v) = \prod_{i=1}^{N} f(y_i) \prod_{i=1}^{N} f\left(v_i\middle|y_i\right) \prod_{i=1}^{M} f\left(z_i\middle|y\right) \prod_{j=1}^{M} \prod_{k=1}^{T_j} f\left(w_{jk}\middle|z_j\right) f\left(s_{jk}\middle|w_{jk}\right).$$

The EM algorithm is used to iteratively update the parameters of the marginal and conditional distributions and model the latent variables Y, Z, and W to attempt to maximize the joint distribution. Finally, the joint distribution is used to obtain the desired outputs: the conditional probabilities $P\left(Z_j = 1\middle|S, V; \hat{\theta}\right)$ that the jth peptide is present for $j = 1, \ldots, M$, and the conditional probabilities $P\left(Y_i = 1\middle|S, V; \hat{\theta}\right)$ that the ith protein is present for $i = 1, \ldots, N$ using the estimated values of the model parameters $\hat{\theta}$.

3.2 Nested Mixture Model

The nested mixture model (NMM) proposed in [21] assumes a mixture model for the joint density of the random variables Y, P, n, and X. Here $Y = (Y_1, \ldots, Y_N)$ is a vector of indicators for the presence/absence of the proteins in the sample where $Y_k = 1$ indicates that the kth protein is present in the sample, $P = (P_{1,1}, \ldots, P_{1,n_1}, \cdots, P_{N,1}, \ldots, P_{N,n_N})$ is a double-indexed vector of indicators of correct assignments of present peptides to a spectrum where $P_{k,i} = 1$ indicates

that the ith peptide of the kth protein is correctly identified, n_k is the number of peptide identifications for the kth protein, $n = (n_1, \ldots, n_N)$, and $X = (x_{1,1}, \ldots, x_{1,n_1}, \cdots, x_{N,1}, \ldots, x_{N,n_N})$ is a double-indexed vector of scores for each peptide assignment. Letting π_1^* denote the probability of a protein being present in the model, NMM assumes that $Y_1, \ldots, Y_N$ are independent Bernoulli random variables with probability mass function

$$f(y_k) = \left(\pi_1^*\right)^{y_k} \left(1 - \pi_1^*\right)^{1-y_k}$$

for $k = 1, \ldots, N$. Letting π_1 be the probability that the ith peptide is correctly identified given that the kth protein is present, it is also assumed that the conditional distribution of $P_{k,i}$ given Y_k has probability mass function

$$f\left(p_{k,i} \middle| y_k\right) = \left\{1 - (1 - y_k)\, p_{k,i}\right\} \pi_1{}^{(1-p_{k,i})y_k} (1 - \pi_1)^{p_{k,i}y_k}.$$

Then the conditional distribution of the scores for the kth protein given Y_k is modeled by the mixture distribution

$$g_t\left(x_{k,1}, \ldots, x_{k,n_k}\right) = \prod_{i=1}^{n_k} \sum_{p=0}^{1} f\left(p \middle| y\right) f_p\left(x_{k,i}\right)$$

where f_0 is the probability density function for a Normal random variable with mean μ and variance σ^2 and f_1 is the probability density function for a shifted gamma random variable with shape parameter α, scale parameter β, and shift parameter γ. Finally, [21] assumes that the conditional distribution of n_k given Y_k follows a truncated Poisson distribution with probability mass function

$$h_y(n_k) = \frac{e^{-c_j l_k} \left(c_j l_k\right)^{n_k}}{n_k! \left(1 - e^{-c_j l_k}\right)}$$

for $n_k = 1, 2, \ldots$, where c_j represents the average number of incorrect/correct peptide identification per unit protein length for $j = 0, 1$, respectively. Then combining these components of the NMM, the joint density of Y, P, n and X is assumed to have the form

$$f(y, z, w, s, v) = \prod_{i=1}^{N} f(y_i) \prod_{i=1}^{N} f\left(v_i \middle| y_i\right) \prod_{i=1}^{M} f\left(z_i \middle| y\right) \prod_{j=1}^{M} \prod_{k=1}^{T_j} f\left(w_{jk} \middle| z_j\right) f\left(s_{jk} \middle| w_{jk}\right).$$

Let ψ denote the vector of all model parameters. Then the EM algorithm is used to estimate ψ, and these estimates are used to obtain $P\left(Y_k = 1 \middle| x_{k,1}, \ldots, x_{k,n_k}, n_k\right)$, the probability that the kth protein is present given the scores and number of peptide hits for that protein, and to obtain

$P\left(P_{k,i} = 1 \middle| x_{k,1}, \ldots, x_{k,n_k}, n_k\right)$, the probability that the ith peptide for the kth protein is present given the scores and number of peptide hits for that protein.

4 Discussion

Proper inference from data produced from tandem mass spectrometry experiments regarding proteins present in tissues and fluids can assist in providing important biological information. Several popular probabilistic and likelihood-based methods for protein identification from MS/MS data have been reviewed: the benchmark two-step process of PeptideProphet followed by ProteinProphet and two likelihood-based one-step processes HSM and NMM. It is important to note that there are many other approaches available in the literature. See [22–24] for review of some other two-step methods for peptide and protein identification. There are also several other one-step protein identification procedures proposed in the literature and a few will be discussed here briefly. ProteinFirst [25] is a two-dimensional target decoy method which simultaneously controls the false discovery rates of proteins and peptide-to-spectrum match levels by modifying PSM scores based on the confidence in the protein identification score. A couple of other methods also consider feedback from proteins when determining the peptides that are present. An iterative procedure to compute peptide and protein probabilities simultaneously is considered by [26] which uses the PeptideProphet results as input for confidence concerning the peptides. Alternately, the method in [27] uses a different mechanism for feedback, starting with peptide identification results from a database search; these results are used to obtain a list of proteins which are further used to obtain a peptide adjacency matrix. Then peptide identification probabilities are estimated based on a logistic regression model and subsequently used to update the protein list and adjacency matrix. Another approach proposed in [28] uses a tripartite graph with three layers corresponding to the spectrum, peptide, and protein levels and uses machine learning techniques in a single optimization procedure for protein identification via a Barista model. The number of true proteins identified by this method exceeds that of ProteinProphet for six different data sets in [28] over a wide range of false discovery rate levels. A promising recent full Bayesian approach (BHM) is proposed in [20] that incorporates the fact that proteins which share the same biological pathway may not be independent. Instead, BHM groups the proteins that are functionally related and uses this fact as prior information for protein identification. Moreover, BHM fully handles the degeneracy issue and considers full posterior inference via a Gibbs sampling scheme. Methods of integrating additional information outside the MS/MS experiment have also been considered and are briefly reviewed in [15].

Various criteria have been used to evaluate the performance of peptide and protein identification procedures, and the performance of the methods has been analyzed and compared using several data sets in the literature. In [16], a training

dataset from [11] with ESI-MS/MS spectra generated from a control sample with 18 purified proteins was used, and the results of PeptideProphet based on SEQUEST database search scores are thoroughly analyzed. In this application, peptide assignments with known validity were generated in a training dataset using SEQUEST with a database including the sequences of the 18 control proteins and a *Drosophilia* peptide database. Test data was generated using the control peptides and a human peptide database. It is shown in Figure 3 of [16] that the estimated distribution for the discriminant score is very close to the true distribution for the test data. Also, the accuracy of the probability estimates of the peptide assignments for the test data is illustrated in Figure 4 of [16] by comparing the true probability with the computed probability. Finally, a pair of graphs in Figure 5 of [16] illustrated the tradeoff between the fraction of identified peptide assignments which are actually correct (sensitivity) and the fraction of identified peptide assignments which are actually incorrect for various thresholds used to classify the peptide assignments and the relationship between these fractions and the threshold.

Some similar analyses were also performed in [13] using the data from [11] to evaluate the ability of ProteinProphet to make protein identifications. Figures 5 and 6 of [13] compare the true probability with the computed probability for the presence of proteins and the relationships between the sensitivity, error rates, and threshold for declaring a protein to be identified. One important additional consideration in these plots was the comparison of results with or without using the number of sibling peptides; all figures clearly showed that the results were better when this information was included.

Comparisons have been made between PeptideProphet/ProteinProphet, HSM, and NMM by comparing the empirical FDR (false discovery rate) versus the estimated FDR, the sensitivity versus the specificity, and the number of true positive proteins versus the number of false positive proteins. In [14], MS/MS spectra data generated based on standard protein mixture [29] were studied, and peptide and protein identification was performed using HSM, PeptideProphet, and ProteinProphet. The results were evaluated using decoy data from *Shewanella oneidensis*, and the empirical FDR was compared with the estimated FDR for each of the methods in Figure 4 of [14]. It was found that PeptideProphet significantly underestimates the FDR. HSM and ProteinProphet both were slightly optimistic at low values of the empirical FDR, and conservative at high FDR. Receiver operating characteristic curves were also presented in Figure 5 of [14] for these methods to compare the sensitivity with the specificity (fraction of false peptides not identified), and the sensitivity of HSM was best for sufficiently high levels of specificity shown, followed by ProteinProphet and PeptideProphet. Additionally, HSM and ProteinProphet were also compared in [14] for processed MS/MS spectra data from [14] generated from a yeast (*Saccharomyces cerevisiae*) dataset with peptide fragments obtained from a QSTAR mass spectrometer [30]. Database search scores were again obtained using SEQUEST for the true data and decoy data from *Caenorhabditis elegans*. It was found that ProteinProphet selects more proteins at each threshold, and that ProteinProphet was optimistic in estimating the FDR for low values of the empirical FDR, while HSM was always conservative in

estimating the FDR. MS/MS spectra data generated based on standard protein mixture [29] were also analyzed in [21], and peptide identification was performed using PeptideProphet, ProteinProphet, HSM, and NMM with SEQUEST database search scores. It was found that the NMM was much more conservative than the other methods (see Figure 6b of [21]). Furthermore, it was seen that the sensitivity of NMM far exceeded that of PeptideProphet and HSM when the specificity was large (see Figure 6a of [21]).

Additionally in [21], NMM and PeptideProphet/ProteinProphet were compared on another yeast dataset from [31] and scores for the peptide assignments were computed using SEQUEST. The decoy data was created by permuting the target sequences; [21] was unable to run HSM for this dataset because of the large protein group sizes. Probability thresholds were selected based on the number of decoy peptides, and it is seen in Figure 7a of [21] that the number of true peptides identified by NMM clearly exceeded the number identified by PeptideProphet. However, when the threshold was selected to control the number of decoy proteins, ProteinProphet slightly outperformed NMM in terms of the number of true positive proteins for most thresholds.

NMM and HSM were also compared in [21] for the processed MS/MS spectra data from [14] generated from the yeast dataset. Thresholds were selected based on the number of false positive proteins, and it is seen in Figure 8 of [21] that the number of true detected proteins of NMM exceeded that of HSM for each threshold (fixed number of false positives). The same data was also shown in [15], and the number of true positives for BHM exceeded NMM and HSM for most thresholds. The number of true positives for BHM also exceeded that of NMM and ProteinProphet in an example in [20] and [15] on *Haemophilus influenzae* data from [13] (see Figure 6 of [15]).

References

1. Yates, J. R., Ruse, C. I., & Nakorchevsky, M. (2009). Proteomics by mass spectrometry: Approaches, advances, and applications. *Annual Review of Biomedical Engineering, 11*(1), 49–79.
2. Eng, J. K., McCormack, A. L., & Yates, J. R., III. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry, 5*(11), 976–989.
3. Eng, J. K., Fischer, B., Grossmann, J., & Maccoss, M. J. (2008). A fast SEQUEST cross correlation algorithm. *Journal of Proteome Research, 7*(10), 4598–4602.
4. Diament, B. J., & Noble, W. S. (2011). Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research, 10*(9), 3871–3879.
5. Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466–1467.
6. Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry. *Electrophoresis, 20*(18), 3551–3567.

7. Clauser, K. R., Baker, P., & Burlingame, A. L. (1999). Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry, 71*(14), 2871–2882.
8. Kim, S., Gupta, N., & Pevzner, P. A. (2008). Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *Journal of Proteome Research, 7*(8), 3354–3363.
9. Swaney, D. L., Wenger, C. D., & Coon, J. J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of Proteome Research, 9*(3), 1323–1329.
10. Granholm, V., Kim, S., Navarro, J. C. F., Sjolund, E., Smith, R. D., & Kall, L. (2014). Fast and accurate database searches with MSGF+ Percolator. *Journal of Proteome Research, 13*(2), 890–897.
11. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., & Kolker, E. (2002). Experimental protein mixture for validating tandem mass spectral analysis. *Omics, 6*(2), 207–212.
12. Nesvizhskii, A. I., & Aebersold, R. (2004). Analysis, statistical validation and dissemination of large-scale proteomics data sets generated by tandem MS. *Drug Discovery Today, 9*(4), 173–181.
13. Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry, 75*(17), 4646–4658.
14. Shen, C., Wang, Z., Shankar, G., Zhang, X., & Li, L. (2008). A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics, 24*(2), 202–208.
15. Sikdar, S., Gill, R., & Datta, S. (2015). Improving protein identification from tandem mass spectrometry data by one-step methods and integrating data from other platforms. *Briefings in Bioinformatics, 17*(2), 262–269.
16. Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383–5592.
17. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
18. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38.
19. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., et al. (2011). iProphet: Multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & Cellular Proteomics, 10*(12), 1–15.
20. Mitra, R., Gill, R., Sikdar, S., & Datta, S. (2015). Bayesian hierarchical model for protein identifications. *Under review*.
21. Li, Q., MacCoss, M., & Stephens, M. (2010). A nested mixture model for protein identification using mass spectrometry. *The Annals of Applied Statistics, 4*(2), 962–987.
22. Huang, T., Wang, J., Yu, W., & He, Z. (2012). Protein inference: A review. *Briefings in Bioinformatics, 13*(5), 586–614.
23. Nesvizhskii, A. I., Vitek, O., & Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods, 4*(10), 787–797.
24. Serang, O., & Noble, W. (2012). A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface, 5*(1), 3–20.
25. Bern, M. W., & Kil, Y. J. (2011). Two-dimensional target decoy strategy for shotgun proteomics. *Journal of Proteome Research, 10*(12), 5296–5301.
26. Shi, J., & Wu, F.-X. (2012). A feedback framework for protein inference with peptides identified from tandem mass spectra. *Proteome Science, 10*, 68.
27. Shi, J., Chen, B., & Wu, F.-X. (2013). Unifying protein inference and peptide identification with feedback to update consistency between peptides. *Proteomics, 13*(2), 239–247.
28. Spivak, M., Weston, J., Tomazela, D., Maccoss, M. J., & Noble, W. S. (2012). Direct maximization of protein identifications from tandem mass spectra. *Molecular & Cellular Proteomics, 11*(2), M111.012161.

29. Purvine, S., Picone, A. F., & Kolker, E. (2004). Standard mixtures for proteome studies. *OMICS, 8*(1), 79–92.
30. Elias, J. E., Haas, W., Faherty, B. K., & Gygi, S. P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods, 2*(9), 667–675.
31. Kall, L., Canterbury, J., Weston, J., Noble, M. J., & MacCoss, W. S. (2007). A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods, 4*, 923–925.