research articles Journal of proteome research

Statistical Validation of Peptide Identifications in Large-Scale Proteomics Using the Target-Decoy Database Search Strategy and Flexible Mixture Modeling

Hyungwon Choi

Departments of Pathology and Biostatistics, University of Michigan, Ann Arbor, Michigan

Debashis Ghosh

Department of Statistics and the Huck Institute for Life Sciences, Pennsylvania State University, University Park, Pennsylvania

Alexey I. Nesvizhskii*

Department of Pathology and Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, Michigan

Received October 22, 2007

Reliable statistical validation of peptide and protein identifications is a top priority in large-scale mass spectrometry based proteomics. PeptideProphet is one of the computational tools commonly used for assessing the statistical confidence in peptide assignments to tandem mass spectra obtained using database search programs such as SEQUEST, MASCOT, or X! TANDEM. We present two flexible methods, the variable component mixture model and the semiparametric mixture model, that remove the restrictive parametric assumptions in the mixture modeling approach of PeptideProphet. Using a control protein mixture data set generated on an linear ion trap Fourier transform (LTQ-FT) mass spectrometer, we demonstrate that both methods improve parametric models in terms of the accuracy of probability estimates and the power to detect correct identifications controlling the false discovery rate to the same degree. The statistical approaches presented here require that the data set contain a sufficient number of decoy (known to be incorrect) peptide identifications, which can be obtained using the target-decoy database search strategy.

Keywords: mass spectrometry • peptide identification • protein sequence database searching • statistical validation • semisupervised modeling • decoy sequences

Introduction

Mass spectrometry has become the method of choice for high-throughput protein identification and quantification in most large-scale studies.[1] Technological advances in this area brought new opportunities for protein analysis, including protein quantification, characterization of post-translational modifications, and protein–protein interactions. In tandem mass spectrometry (MS/MS) based proteomics, sample proteins are first enzymatically digested into shorter fragments, or peptides, and the peptide mixtures are separated by reverse-phase capillary liquid chromatography (LC) or other separation techniques. The separated peptides are ionized and fragmented in the mass spectrometer to produce signature MS/MS spectra. The computational analysis is then carried out to identify the peptides that generated the spectra and to infer the identities of proteins present in the original sample.[2]

The most commonly used method for peptide identification from MS/MS spectra is database searching. In this approach, experimental MS/MS spectra are queried against theoretically derived spectra predicted for peptides contained in a protein sequence database. A number of database search tools are available such as, e.g., SEQUEST,[3] MASCOT,[4] and X! TANDEM[5] (see, e.g., ref 2 for a recent review). Database peptides assigned to experimental spectra are filtered based on search scores generated by the search algorithm, and then filtered data are reported as a list of proteins present in the sample. However, not all peptide assignments are correct. Incorrect identifications result from many reasons, e.g., the peptide sequence is not in the search database or low-quality MS/MS spectra are used for the database search. Regardless of the source of false identification, it is important to be able to assess the statistical significance of individual peptide identifications as well as the composite error rates associated with filtering the data using various thresholds.[6]

* To whom correspondence should be addressed. E-mail:nesvi@med.umich.edu.

10.1021/pr7006818 CCC: $40.75

Published on Web 12/14/2007

Several approaches have been developed for assessing the confidence in peptide identifications. These can be roughly divided into single-spectrum and global (whole data set) modeling approaches. The most commonly used single-spectrum statistical measure is the expectation value, which refers to the expected number of peptides with scores equal to or better than the observed search score under the assumption that the peptide was assigned to the experimental spectrum by random chance.[7,8] The expectation values are less dependent on the details of the scoring method used to compare experimental and theoretical spectra, which gives a clearer interpretation of goodness of match across different instrument platforms and search algorithms. However, the conversion of a raw search score into an expectation value does not control the overall identification error rates, since its construction does not specifically involve steps for multiple testing correction. Thus, when dealing with large-scale data sets of peptide assignments to MS/MS spectra, additional analysis has to be carried out that would allow filtering the data with a desired false discovery rate (FDR).

One of the global statistical approaches, implemented in the commonly used computational tool PeptideProphet, is to model the observed distribution of scores (raw search scores or expectation values) and auxiliary properties (e.g., peptide molecular weight measurement from the first stage of the MS analysis and information from the protein digestion and peptide separation steps) as realizations from a mixture distribution of scores representing correct and incorrect peptide assignments.[9,10] In this method, maximum likelihood estimation of distribution parameters using the expectation−maximization (EM) algorithm[11] leads to mixture deconvolution, and the posterior probability of correct identification is calculated for each assignment from the deconvoluted mixture distribution by Bayes rule. Using this scale-free, univariate probability score, one can call peptide assignments correct if posterior probability is above a certain threshold. This posterior probability is directly related to local false discovery rate (fdr) discussed in Efron et al.[12] and Newton et al.,[13] and thus specifying the minimum probability threshold automatically controls global FDR to a desired degree.

Recently, this approach was extended by incorporating information from decoy assignments in the mixture estimation algorithm.[10] Decoys are peptide identifications that are known to be incorrect. They can be obtained using the target-decoy database search strategy, in which the protein database for the organism of interest (target database) is appended with a decoy database (e.g., reversed, randomized, or shuffled sequences from the target database). For a recent review, see ref 14. The distribution information from decoys is exploited by allowing their scores to contribute to the estimation of negative distribution only. The decoy distribution effectively yields a stable reference distribution of incorrect assignments, resulting in more reliable control of false discovery rates across many challenging examples. In this method, the likelihood to be maximized (in the M-step of the EM algorithm) is decomposed as a product of the following two distributions: a two-class (e.g., Gumbel/Normal in MASCOT, Gamma/Normal in SEQUEST) mixture distribution for peptide assignments from the original search database, and a fixed distribution (e.g., Gamma in SEQUEST, Gumbel in MASCOT, TANDEM) for decoy assignments. In the E-step, the probability of correct identification is then calculated for target assignments only, and the mixture proportion is also estimated using target assignments. The use of a decoy peptide makes the classification problem semisupervised, in the sense that the known class labels for decoy assignments are used but the class labels for all other assignments are not known a priori. In what follows, we call the parametric mixture model the unsupervised parametric model if the database search was performed without decoys and the semisupervised parametric model otherwise. Several data analysis examples using the semisupervised parametric model are presented in ref 10, where the impact of the information from the decoys on the accuracy of statistical modeling was assessed in comparison with the unsupervised parametric model.

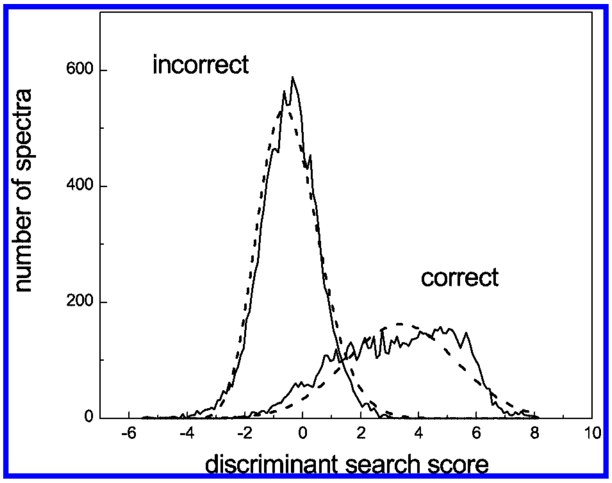

Figure 1. Histogram of SEQUEST discriminant search score *S* plotted separately for incorrect and correct peptide assignments to doubly charged spectra (solid lines). Also shown are the distributions learned by the semisupervised parametric model (dashes). See Data section for a description of the data set.

In the parametric methods, however, the reliability of validation depends on proper parametric specification of the probability model for the search scores. One of the limitations of PeptideProphet is that it is difficult to find a default probability model (i.e., shape of the score distribution) since the appropriate choice depends on the scoring algorithm and, to some degree, on the choice of various database search parameters. As a result, adoption of the computational approach to each new database search tool requires analysis of the database search score distributions and investigation of various score transformations. Even then, deviations from the parametric assumptions can sometimes be observed in practice, in which case computed probabilities may be less accurate. One such example, a SEQUEST search of a data set of MS/MS spectra generated using a linear ion trap Fourier transform (LTQ-FT) mass spectrometer, is shown in Figure 1.

To improve the accuracy of statistical validation and simplify the application of the method to model the output of new search tools, we describe two flexible mixture models that avoid restrictive model specification in the estimation of mixture distribution. In the first, variable component approach, parametric specification of continuous mixture components is replaced by a mixture of Gaussian mixtures of an unknown number of subcomponents. The second method is an iterative weighted kernel density estimation, where the parameters of the positive distributions and the mixture proportion are estimated using a semiparametric density estimation method similar to that of Robin et al.[15] Both methods require decoy peptides since they tend to prevent the identifiability problem in the variable component approach, and the negative distribution cannot be obtained at all without decoys in the semiparametric model. Using the data set shown in Figure 1, we demonstrate that the information from decoy peptides significantly contributes to more reliable estimation of the negative distribution (score distribution of incorrect assignments), that relaxing parametric model assumption allows accurate model fits even in the most challenging cases, and that the power to detect true positives is improved compared to the parametric model given a fixed cut point for posterior probability of correct identification.

Methods

Data. MS/MS spectra used in this work were generated using a linear ion trap Fourier transform (LTQ-FT) instrument using a mixture of purified proteins.[16] Prior to MS/MS analysis, proteins were digested with trypsin and separated using reversed-phase LC coupled online to the mass spectrometer. The spectra (mixture 3, 10 LC-MS/MS runs in total) were searched using SEQUEST against the sequences of proteins known to be present in the sample (target database) appended with a much larger database of reversed protein sequences extracted from the Human IPI database (decoys). The search was conducted allowing peptides that are not tryptic at no more than one of the termini (partially constrained mode), with two missed cleavages or less, using 3 Da monoisotopic mass tolerance and no modifications. Note that the large 3 Da mass window was used to illustrate the ability of the methods to accurately capture the distribution of mass accuracy scores and may not be optimal. The resulting data set contained a total of 21 821 peptide assignments. Among them, 8323 were assignments to sequences of the proteins known to be in the sample, and 13 498 were matches to sequences from the decoy database. Because the size of the decoy database was much larger than the size of the target database, all assignments to target sequences in this data set were assumed to be correct.

Discriminant Features. Database search tools compare each MS/MS spectrum against all candidate peptides in the searched protein sequences database that satisfy a certain set of user-defined criteria (precursor mass tolerance, digestion constraint, etc.).[17] The highest scoring peptide is assigned to each experimental spectrum based on database search score. In addition to the database search score S (in this work, SEQUEST discriminant search score[9,10]), further useful information may be available regarding the quality of peptide assignments depending on the search options, serving as independent discriminating features for peptide identification. A more detailed discussion of various discriminant features can be found in ref 10. A peptide assigned to an experimental spectrum is more likely to be correct if it conforms to the specificity of the enzyme used to digest the proteins. In the case of trypsin digestion, the confidence in assigned peptides that have less than two tryptic termini, and/or missed internal cleavage sites, should be downweighted. These sequence properties are quantified using the number of tryptic termini parameter, NTT, and the number of missed cleavages, NMC, respectively. Furthermore, the molecular weight of the assigned peptide can be computed theoretically and compared to the observed precursor ion mass (information available from the first stage of mass spectrometry, MS^1 spectrum). A large difference between the two weights, denoted here by dM, reduces the confidence in the peptide assignment. While only these four scores are used in this work given the nature of the data, this framework can host other information on the features of peptide matches as long as it is legitimate to assume conditional independence among the information variables given identification class.

Probability Model. Given the four scores, search score S and auxiliary information NTT, NMC, and dM (collectively designated as E), the task is to model the joint distribution of these scores as a two-component mixture, where the components represent the score distributions of incorrect and correct identification classes. On the basis of the deconvoluted mixture, the posterior probability of correct identification for individual peptide assignments can be determined, leading to a classification rule that calls peptide assignments correct if probability is above the threshold probability p_T, where $1 - p_T$ corresponds to the local false discovery rate (fdr).

The joint distribution of scores, $f(S, E)$ is specified as previously described.[9,10] For a peptide assignment to the ith spectrum in the data set with the scores (S_i, E_i)

$$f(S_i, E_i) = \pi_0 f_0(S_i, E_i) + \pi_1 f_1(S_i, E_i) \tag{1}$$

where $\pi_0 + \pi_1 = 1$ and f_k is the joint distribution of the scores in incorrect identifications if $k = 0$ and correct identifications otherwise. It is also assumed that

$$f_0(S,\ E) = f_0^S(S) f_0^{\mathrm{NTT}}(\mathrm{NTT}) f_0^{\mathrm{NMC}}(\mathrm{NMC}) f_0^{\mathrm{dM}}(\mathrm{dM}) \tag{2}$$

and the same for f_1, where $f_k^x(x)$ is the marginal distribution of individual score x in each identification class k. The additional key assumption of the parametric mixture model was that the distributions $(f_0^S(S), f_1^S(S))$ are modeled using a certain fixed shape, e.g., (Gamma, Normal) or (Gumbel, Normal). This limitation, as well as the requirement for discretization of the mass accuracy variable dM, will be removed in the models described below. The assumption that the distributions of individual scores are mutually independent conditional on the identification class remains unchanged. Thus, it follows that all the scores interact with one another through the overall proportion of incorrect and correct identifications (π_0, π_1) only. In practice, this assumption is not severely violated in typical data sets. As in ref 9, peptide assignments to spectra of different charge state are modeled separately.

Model 1: Variable Component Approach. The key improvement that this work aims to achieve is that there is no need to specify a particular family of distributions for the marginal distribution of continuous scores in each identification class. One way to achieve this is to perform Bayesian estimation on the variable component mixture model. Integrating eq 1 over all possible values of (E) and modeling $f_k^S(\bullet)$ as a mixture of an unknown number of Gaussian subcomponents for both $k = 0$ and $k = 1$, it is assumed that $\{S_i\}_{i=1}^n$ are iid (independent and identically distributed) draws from

$$\boxed{\begin{aligned} f^S(S) &= \pi_0 f_0^S(S) + \pi_1 f_1^S(S) \\ &\equiv \underbrace{\pi_0 \left(\sum_{j=1}^{k_0} w_j^0 \underbrace{\phi(S|\mu_j^0, \sigma_j^0)}_{\text{subcomponent}} \right)}_{\text{negative component}} + \underbrace{\pi_1 \left(\sum_{j=1}^{k_1} w_j^1 \underbrace{\phi(S|\mu_j^1, \sigma_j^1)}_{\text{subcomponent}} \right)}_{\text{positive component}} \end{aligned}} \tag{3}$$

where (π_0, π_1) are mixture weights of incorrect and correct identifications from eq 2, respectively; triplet (w, μ, σ^2) denote mixture weights, means, and variances for subcomponents; and ϕ denotes Gaussian density. This specification allows a smooth model fit not bound by a particular shape restriction. We call ϕ's *subcomponents* in the sense that a collection of ϕ's comprises a mixture component representing the score distribution of either correct or incorrect identifications. Most importantly, the numbers of subcomponents in each class (k_0 and k_1) are assumed to be random, and the model is therefore called the variable component model.

We also assume the mass accuracy scores $\{\mathrm{dM}_i\}_{i=1}^n$ are iid draws from the distribution of the form in eq 3, where the mixture weights (π_0, π_1) remain the same from the integration of eq 1 over all possible values of (S, NTT, and NMC). Modeling dM with a continuous distribution is an important difference from parametric models, where dM is discretized into a certain number of bins and modeled as a mixture of two multinomial distributions.

The other two scores (NTT, NMC) are modeled as realizations from a mixture of two multinomial trials of size 1. That is, $\{\mathrm{NTT}_i\}_{i=1}^n$ are iid draws from the discrete mixture

$$\begin{aligned} f^{\mathrm{NTT}}(\mathrm{NTT}) &= \pi_0 f_0^{\mathrm{NTT}}(\mathrm{NTT}) + \pi_1 f_1^{\mathrm{NTT}}(\mathrm{NTT}) \\ &\equiv \pi_0 M(\mathrm{NTT};\ p_0,\ p_1,\ p_2) + \pi_1 M(\mathrm{NTT};\ q_0,\ q_1,\ q_2) \end{aligned} \tag{4}$$

where M denotes the multinomial distribution. The same form of a marginal mixture distribution is defined for NMC.

Model estimation is performed using reversible jump Markov chain Monte Carlo (MCMC) devised for the variable component mixture model.[18,19] Although MCMC is computationally heavy relative to the parametric model estimation, the variable component model thus acquired tends to be more robust since the sampling algorithm explores a wide range of possible models that could have generated the data. See Supplemental Methods for a more detailed account of the prior elicitation and the sampling algorithm used here.

Model 2: Semiparametric Approach. As an alternative to the variable component approach, we also examine a semiparametric density estimation method similar to that of Robin et al.[15] In this approach, decoy peptides play a more important role since the negative distributions for all four scores (S, NTT, NMC, dM) are obtained solely from decoy peptides. Thus we assume that the negative distribution is known (estimated nonparametrically from decoys) and iterate EM-like steps with fixed negative distribution. In the iterative algorithm, the conditional probability of each observation belonging to the positive component is calculated in the E-step and the positive distribution is re-estimated by weighted kernel estimation with weights being the aforementioned conditional probability for each observation.

In this approach, the first step is to estimate the negative distributions from the decoy peptides. By applying kernel density estimation with a Parzen window, we obtain $f_0^S(\bullet)$ and $f_0^{\mathrm{dM}}(\bullet)$ as follows. First, an equally spaced, dense grid of 100−500 points is fixed on the domain of continuous discriminant search scores S, and the Gaussian kernel density estimate is calculated at each grid

$$f_0^S(S|h_S) = \frac{1}{n_0 h_S}\sum_{i=1}^{n_0} K\left(\frac{S-S_i}{h_S}\right) \tag{5}$$

where K is the Gaussian density function and n_0 is the number of decoy assignments. Note that in this part of the estimation only the decoy peptides are used. Then the same is applied to mass accuracy dM. The bandwidths (h_S, h_{dM}) are selected only once in the beginning of the algorithm and fixed throughout. They are estimated by 5-fold cross-validation with log-likelihood maximization criterion. $f_0^{\mathrm{NTT}}(\bullet)$ and $f_0^{\mathrm{NMC}}(\bullet)$ are easily calculated by sample proportions.

In the E-step, the probability that spectrum i is correctly assigned is calculated as

$$p_i \equiv P(+|S_i,\ E_i) = \frac{\pi_1 f_1(S_i,\ E_i)}{f(S_i,\ E_i)}. \tag{6}$$

where + stands for the event that an assignment is correct. In the next step, similar to the M-step in parametric models, kernel density estimates are calculated

$$f_1^S(S|h_S) = \frac{\Sigma_{i=1}^n p_i K\left(\frac{S-S_i}{h_S}\right)}{h_S \Sigma_{i=1}^n p_i} \tag{7}$$

at all grid points. The same step is applied to mass accuracy dM. For the other two discrete variables, the proportion of each value weighted by probabilities is calculated for $f_1^{\mathrm{NTT}}(\bullet)$ and $f_1^{\mathrm{NMC}}(\bullet)$. Finally, the mixture proportion π_1 is estimated by the average of the probability of correct identification, i.e.

$$\pi_1 = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{8}$$

where n is the number of all target peptide assignments.

Posterior Class Probability and Classification. The goal of this work is to accurately calculate the posterior probability of correct identification, or class probability, for each peptide assignment. In the semiparametric approach, this probability can be directly calculated using eq 6 at the last step of iteration. In the variable component approach, the same quantity can be acquired by taking the sample average of the posterior output. If we denote the overall model parameters by $\Theta = (k_0, k_1, \pi, w, \mu, \sigma, p, q)$, then the probability that ith peptide is correct is

$$\begin{aligned} P(+|S_i, E_i) &= \int_{\Theta} P(+|\Theta,\ S_i,\ E_i)\,dF(\Theta|S_i,\ E_i) \\ &= \sum_{d} p(d)\int_{\Theta^{(d)}} P(+|\Theta^{(d)},\ S_i,\ E_i)\,dF(\Theta^{(d)},\ S_i,\ E_i) \\ &\approx \frac{1}{I}\sum_{k=1}^{I} 1(+|\Theta_k,\ S_i,\ E_i) \end{aligned} \tag{9}$$

where I is the number of iterations in the Markov chain and d is the varying number of mixture subcomponents in both identification classes.

The resulting classification rule calls correct all assignments with $P(+)$ greater than the threshold probability p_T. The threshold probability can be selected with proper control of error rates if we interpret it as the complement of the local false discovery rate (see, e.g., Efron[20] for a discussion in the context of large-scale hypothesis testing). A classification rule of this form controls the local false discovery rate at 100 × (1 – p_T)%. The control of the local false discovery rate is generally reliable as long as the mixture is properly deconvoluted. Further discussions on the efficiency of false discovery rates based on mixture modeling can be found in Newton et al.[13]

Results and Discussion

The performance of the flexible models described above, as well as the parametric models, was evaluated using a control protein mixture data set generated using an LTQ-FT mass spectrometer (see Data). The accuracy and absolute number of correct peptide identifications at a fixed FDR were calculated across the four models: (1) parametric model with no decoy

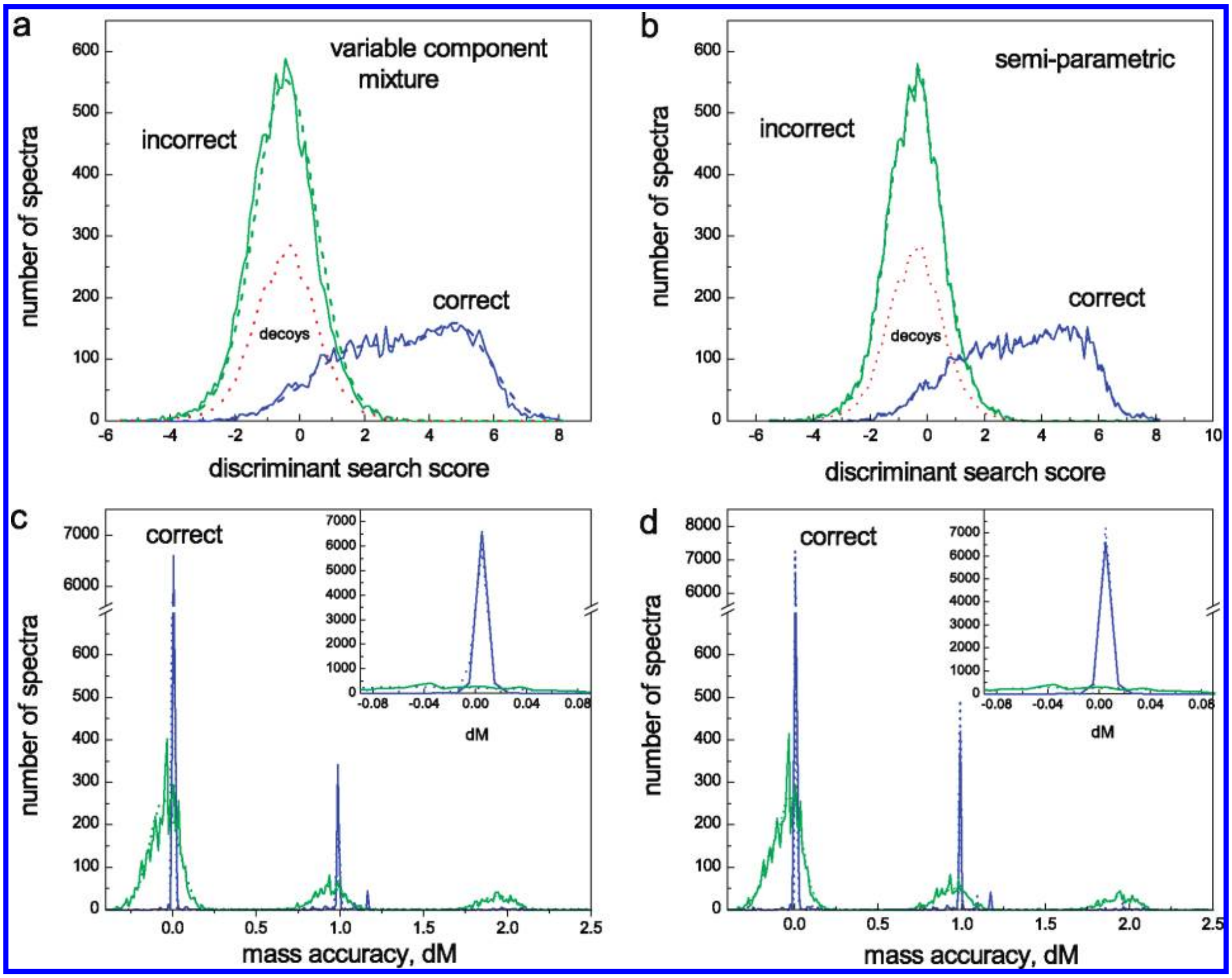

Figure 2. Histogram of SEQUEST discriminant search score *S* and mass accuracy score dM plotted separately for incorrect (solid green) and correct (solid blue) peptide assignments to doubly charged MS/MS spectra. Also shown are the distributions learned by the models (dashes) and the distribution of *S* plotted for decoy peptides used in the models (red dotted line). (a,c) Variable component mixture model. (b,d) Semiparametric model. Observed distributions (solid lines) are identical between (a) and (b) and between (c) and (d), but appear slightly different as plotted since each method may place break points differently. Insets in (c) and (d) show the region of the mass accuracy dM distribution close to 0.

information[9] (unsupervised parametric); (2) parametric model with decoy information[10] (semisupervised parametric); (3) variable component mixture model; (4) semiparametric approach. As discussed above, both flexible models are semisupervised (use decoy information). To allow objective evaluation of the semisupervised models, half of the decoy peptide matches were randomly selected and considered to be unknown (all decoys are treated as unknown in the unsupervised model). These peptides were used to evaluate the accuracy of computed probabilities. The other half of the decoy peptides was used by the models; i.e., they were forced to contribute to the negative distributions only (variable component approach and semisupervised parametric) or were directly used to empirically estimate the negative distribution (semiparametric approach). Both NTT and NMC parameters were used in the analysis; however, the discussion below will focus on the results with respect to the database search score *S* and the mass accuracy dM, mostly because those two are generated from continuous distributions, in which we seek improvements compared to the parametric approach.

The results for the SEQUEST search of MS/MS spectra from doubly charged precursor peptide–ions are shown in Figure 2a,b, which plot the observed histograms of the discriminant database search score *S* separately for correct and incorrect assignments obtained using the variable component mixture and the semiparametric model, respectively. These plots should be compared with Figure 1, which shows the results obtained using the semisupervised parametric algorithm[10] (the unsupervised solution was close, if slightly worse, and not plotted). Clearly, in this data set, the parametric model does not capture the positive distribution well due to the shape restriction of the Gamma/Normal mixture (modeling using other distribution shapes produced similar results). The flexible models show significant improvement in modeling the observed distributions of scores.

The flexible models also allow accurate modeling of the distributions of the mass accuracy score, dM. Since dM is observed on a continuous scale, the ideal strategy is to model it with a continuous distribution. In the parametric model, dM is discretized into a certain number of bins and then modeled with a mixture of two multinomial distributions. These two approaches may in general give similar results when the distribution of dM is sufficiently smooth. In more challenging cases, however, e.g., high mass accuracy data used in this work, the distribution of dM in correct identifications may be so concentrated on a few bins (this is particularly apparent in the case of high mass accuracy data searched with wide mass tolerance, see Figure 2c,d) that the

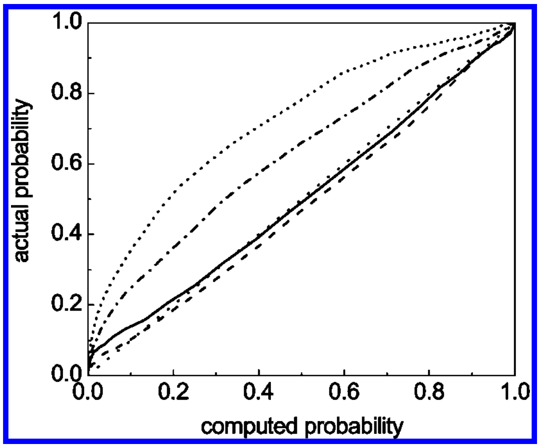

Figure 3. Accuracy of estimated probabilities from the semiparametric (dashes) and variable component (solid) mixture models and from parametric semisupervised (dash dot) and unsupervised (short dashes) models. The results of an ideal model are shown as a 45 degree dotted line.

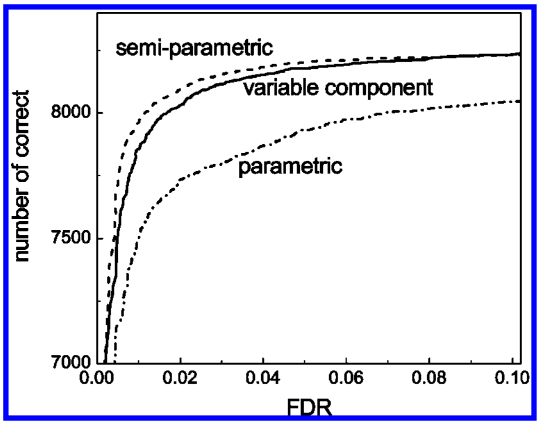

Figure 4. Estimated number of correct peptide identifications as a function of false discovery rate (FDR). Shown are the results of the variable component model (solid), semiparametric model (dashes), and semisupervised parametric model (dash dot). The results from the unsupervised parametric model are similar to those of the semisupervised parametric model and are not shown.

probabilities can be overadjusted for peptides in those bins and underadjusted for peptides elsewhere. In this case, the discrimination by the database search score can easily be dominated by dM. Moreover, when the data set does not include enough correct identification, excessive binning will make the maximum likelihood estimation unstable. The flexible models tend to be free from this behavior since the variable component model is based on smooth distributions and the semiparametric model uses smoothing in the estimation. We illustrate the model fits of dM from the flexible models in Figure 2c,d, which plot the observed distributions of the mass accuracy score among correct and incorrect identifications as well as the model fit obtained using the variable component and semiparametric models. These figures also demonstrate the high mass accuracy (less than 5 ppm) of the LTQ-FT mass spectrometer, which makes dM highly useful for discriminating correct from incorrect identifications.

To quantify the accuracy of calculated class probabilities, peptide assignments were first sorted in decreasing order of estimated probabilities, and the average of the known class labels, i.e., correct (1) and incorrect (0) identifications, was calculated using a sliding window of size 100 (the size of the sliding window had little impact on the shape of the accuracy curve). The average in the sliding window is referred to as actual probability. If actual probability is plotted against the average of model-estimated probabilities within the same window, then a line close to the 45 degree line indicates good agreement between the estimated and actual probabilities. In addition to the estimation accuracy, the concordance between the complement of the probability $1 - p_T$ and the actual local false discovery rate can also be monitored. The actual local false discovery rate can be calculated in a way similar to the way the actual accuracy was computed. For a fixed threshold point p_T, all peptide assignments with estimated probability within a window of δ, say 0.95, are collected, and the proportion of incorrect assignments among them reflects actual local false discovery rate. As the local false discovery rate is essentially complementary to the accuracy, we do not include these plots in the following analysis. Consistent with the goodness of model fit, Figure 3 shows that the accuracy of probability estimates from the semiparametric and the variable component mixture models is superior to that of parametric models (both semisupervised and unsupervised), including the high probability region 0.5 and above, that is of most relevance in practical terms.

The accuracy of the probabilities computed using parametric models can be somewhat improved by implementing an empirical correction, e.g., by bounding the ratio between positive and negative distributions in dM to be within a certain range (e.g., below 10), as shown in the Supplementary Figure S1. With the flexible models, such empirical corrections become unnecessary.

In addition to the improved accuracy of modeling, the flexible mixture models allow better separation between correct and incorrect identifications. The improved sensitivity of filtering the data at a fixed FDR is shown in Figure 4. Both the variable component model and the semiparametric model allow selection of a higher number of correct identifications at all FDRs than the parametric models. This improvement, however, is less significant than the 30% or more improvement[9,10] that is typically observed between the probability-based filtering in general and simple threshold-based methods that are commonly applied in practice.

While both flexible models produce similar results, it is important to discuss the differences between them. Despite the flexibility and robustness of the variable component mixture model, its success may be subject to proper prior elicitation for some of its modeling components and also requires a sufficient number of sampling steps from the posterior distribution associated with the reversible jump Markov chain Monte Carlo method and thus increased computational burden. These drawbacks are important weaknesses from the practical point of view if the goal of developing the method is to deliver a fast and automated computational tool free from additional tuning of parameters. If the practical implications of computational burden and parameter tuning are critical, the semiparametric model has an obvious advantage. Much of its computation is spent on bandwidth selection, but the entire computation time is much less than that for the Bayesian estimation procedure and comparable to that of parametric models. Furthermore, the semiparametric approach can be easily implemented in the existing computational tools, e.g., PeptideProphet.

At the same time, it is important to point out that the estimation of the negative distribution in the semiparametric approach becomes completely dependent on the decoy peptides, as in other similar strategies.[21] An important issue that has not been fully addressed yet is whether creating decoys by reversing or randomizing target protein sequences can provide an accurate assessment of the negative distribution, since many of the incorrect matches happen to be made to sequences homologous to the true peptides,[22–24] rather than completely random. In that regard, the variable component model has the same advantage as the parametric methods in that the distribution of incorrect peptide identifications can deviate from the distribution of decoy peptides. Additional discussion on the use of decoys peptides within the mixture modeling approach can be found in ref 10.

Finally, it is important to mention that in most studies researchers are interested in identifying proteins present in the original sample rather than peptides resulting from the proteolytic digest of the sample. Thus, peptide identifications need to be grouped according to their corresponding protein and the statistical assessment performed at the protein level.[23] Improved accuracy of computed probabilities should result in more accurate estimation of protein-level probabilities by computational tools such as ProteinProphet[25] that take peptide probabilities as input. Furthermore, the protein level modeling can benefit from using the information from decoy protein identifications in an analogous way. However, the protein level analysis is further complicated due to the presence of homologous protein sequences. Thus, the distribution of protein level scores computed for decoy protein identifications is unlikely to be an accurate representation of the distribution of incorrect identifications of proteins from the target database. This issue should be carefully examined in future work.

Conclusions

In this paper, we have extended the mixture modeling approach of refs 9 and 10 for validation of peptide identification using a wider class of mixture models. The variable component mixture model with a minimal set of parametric model assumptions and the semiparametric mixture model lead to flexible and robust model-based estimation of the classification error rates associated with peptide identification through database searching. In both alternatives, as a consequence of the precise estimation capability of our method, the mixture modeling provides a reliable gauge for controlling false discovery rates in real data analysis, with the benefit of having a univariate score for goodness of match for each peptide assignment. The particular advantage of the method is its ability to model multiple sources of discriminant information in addition to the database search score itself, which improves the accuracy of probability calculation as well as the statistical power of the method. The implementation of the method in R programming language is available upon request from the authors.

Acknowledgment. This work was supported in part by NIH/NCI Grant CA-126239. We thank Eric Deutsch, Ying Ding, Henry Lam, Johan Lim, and David Shteynberg for helpful discussions.

Supporting Information Available: Supporting Information includes a detailed description of the variable component model and a supplementary figure. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207.

(2) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787–797.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 975–989.

(4) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(5) Craig, R.; Beavis, C. R. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(6) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A. I.; Clauser, K.; Nesvizhskii, A. I. The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics* **2004**, *3*, 531–533.

(7) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768–774.

(8) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.

(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications amde by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.

(10) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254–265.

(11) Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc., Ser. B* **1977**, *39*, 1–38.

(12) Efron, B.; Tibshirani, R.; Storey, J. D.; Tusher, V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1160.

(13) Newton, M. A.; Noueiry, A.; Sarkar, D.; Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **2004**, *5*, 155–176.

(14) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207–214.

(15) Robin, S.; Bar-Hen, A.; Daudin, J.; Pierre, L. A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Stat. Data Anal.* **2007**, *51*, 5483–5493.

(16) Klimek, J.; Eddes, J.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P.; Katz, J.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J.; Aebersold, R.; Martin, D. The Standard Protein Mix Database: A diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7*, 96−103.

(17) Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol. Biol.* **2007**, *367*, 87–119.

(18) Green, P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **1995**, *82*, 711–732.

(19) Richardson, S.; Green, P. On Bayesian analysis of mixtures with an unknown number of components. *J. Royal Stat. Soc., Ser. B* **1997**, *4*, 731–792.

(20) Efron, B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.* **2004**, *99*, 96–104.

(21) Higgs, R. E.; Knierman, M. D.; Freeman, A. B.; Gerbert, L. M.; Patil, S. T.; Hale, J. E. Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. *J. Proteome Res.* **2007**, *6*, 1758–1767.

(22) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76*, 3556–3568.

(23) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data - the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4*, 1419–1440.

(24) Feng, J.; Naiman, D. Q.; Cooper, B. Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* **2007**, *23*, 2210–2217.

(25) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658.

PR7006818