

A LIKELIHOOD-BASED SCORING METHOD FOR PEPTIDE IDENTIFICATION USING MASS SPECTROMETRY

BY QUNHUA LI^{*}, JIMMY K. ENG[†] AND MATTHEW STEPHENS[‡]

Penn State University^{}, University of Washington[†] and University of
Chicago[‡]*

Mass spectrometry provides a high-throughput approach to identify proteins in biological samples. A key step in the analysis of mass spectrometry data is to identify the peptide sequence that, most probably, gave rise to each observed spectrum. This is often tackled using database search: each observed spectrum is compared against a large number of theoretical “expected” spectra predicted from candidate peptide sequences in a database, and the best match is identified using some heuristic scoring criterion. Here we provide a more principled, likelihood-based, scoring criterion for this problem. Specifically, we introduce a probabilistic model that allows one to assess, for each theoretical spectrum, the probability that it would produce the observed spectrum. This probabilistic model takes account of peak locations *and* intensities, in both observed and theoretical spectra, which enables incorporation of detailed knowledge of chemical plausibility in peptide identification. Beside placing peptide scoring on a sounder theoretical footing, the likelihood-based score also has important practical benefits: it provides natural measures for assessing the uncertainty of each identification, and in comparisons on benchmark data it produced more accurate peptide identifications than other methods, including SEQUEST. Although we focus here on peptide identification, our scoring rule could easily be integrated into any downstream analyses that require peptide-spectrum match scores.

1. Introduction. Tandem mass spectrometry (MS/MS) provides a high-throughput approach to identify proteins in biological samples. In a typical MS/MS experiment, proteins in the sample are first broken into short sequences, called peptides, and the resulting mixture of peptides is subjected to mass spectrometry, which fragments peptides and generates tandem mass spectra that contain fragmentation peaks characteristic of their generating peptides (Coon et al., 2005; Kinter and Sherman, 2000). A variety of computational methods are then used to process the mass spectra, with, typically, the ultimate goal being to identify which proteins and/or peptides are

Keywords and phrases: generative model, maximum likelihood, peptide identification, proteomics

present in the mixture, and to provide some measure of confidence in these identifications.

The computational pipelines used for processing these kinds of data can vary considerably, even for analyses that share the same ultimate goal. However, one element that plays an important role in the vast majority of these pipelines is the need to “score” how well each observed spectrum matches a number of candidate generating peptides. Despite the fact that such scoring procedures play a key role in all kinds of downstream analyses, existing scoring procedures are generally fairly simple and ad hoc. In this paper we develop a more statistically rigorous, likelihood-based, approach to peptide-spectrum scoring, which, in the examples considered later, performs better than existing scoring rules (e.g. the Xcorr score in SEQUEST) in discriminating between the true generating peptide and other candidate peptides. This scoring rule could be integrated easily into any downstream analyses that require peptide-spectrum match scores, including decoy database search strategies (Elias and Gygi, 2007) and formal statistical modeling approaches for protein identification (Gerster et al., 2010; Li et al., 2010; Nesvizhskii et al., 2003; Shen et al., 2008).

In brief, our scoring approach, in common with many existing approaches, has two steps: first, for a given candidate peptide sequence, we generate a theoretical “expected” spectra; second, the observed spectrum is compared with this theoretical spectrum. Most existing algorithms (reviewed in Hernandez et al. (2006); Sadygov et al. (2004)) use simple approaches in both these steps. Specifically, they typically use coarse theoretical spectra containing only predicted locations (not intensities) of spectral peaks derived from a few major chemical fragmentation pathways, and score similarity primarily by the matching of peak locations (again ignoring peak intensities) using ad hoc rules. The resulting peptide identification procedures are generally rather inaccurate (typically 70-90% of the top-scoring peptide identifications are incorrect (Keller et al., 2002; Nesvizhskii and Aebersold, 2004)).

In comparison, our approach attempts to be more sophisticated in both of these steps. For the first step we make use of the improved theoretical prediction algorithm from Zhang (2004) (see also Klammer et al. (2008)), which incorporates gas phase chemistry mechanisms of peptides into a kinetic model for peptide fragmentation, to generate detailed theoretical predicted spectra for any given peptide. These detailed theoretical spectra contain predicted locations and intensities of peaks from both major and minor fragmentation pathways, all of which may help improve accuracy of peptide identification. Indeed such features are commonly used in manual annotation to validate

chemical plausibility of putative peptide identifications (Sun et al., 2007). For the second step we develop a novel likelihood-based scoring rule for this problem. This likelihood-based approach is based on a probabilistic model for differences between the theoretical and observed spectrum, in both peak intensities and locations, and allows for the fact that predicted low-intensity peaks from minor pathways are more often absent from observed spectra than are predicted high-intensity peaks from major pathways. It is in this second step that our work differs from all existing scoring algorithms, including the few that do make use of complex predicted spectra (Yen et al., 2011; Zhang, 2004), which by comparison use simple, ad hoc, measures of similarity to compare observed and theoretical spectra.

As we demonstrate on examples later, likelihood-based scoring has some important practical benefits: it provides natural measures for assessing the uncertainty of each identification, and, on benchmark data we consider here, ultimately improves the accuracy of identification. In addition it has the attraction of putting peptide scoring on a sounder theoretical statistical footing.

The structure of the paper is as follows. We first describe the generation of theoretical spectra (Section 2.1) and the procedure we use to preprocess both observed and theoretical spectra (Section 2.2). Section 2.3 describes our probabilistic model and the methods we use to estimate parameters in this model and score peptide sequences. In Section 3, we check the effectiveness of these methods on simulated data. In Section 4, we illustrate the methods using a publicly available benchmark dataset. Section 5 concludes and discusses future work.

2. Methods and models.

2.1. *Refined theoretical spectra and its use in peptide identification.* We use the chemical model from Zhang (2004) to predict the theoretical spectra for peptide sequences. This model generates refined theoretical spectra containing both locations and intensities of the peaks from comprehensive pathways. For convenience of producing our own pipeline we coded this prediction algorithm in Java. Our implementation produced similar results to Zhang’s software for the examples shown in Zhang (2004), although there were some quantitative differences (peak heights did not always agree). In as much as these differences could reflect deficiencies of our implementation, we note that correcting these deficiencies would be expected to yield further improvements in performance compared with those reported below. Our implementation is available on request from the first author.

Though there is still marked deviation between theoretical prediction and

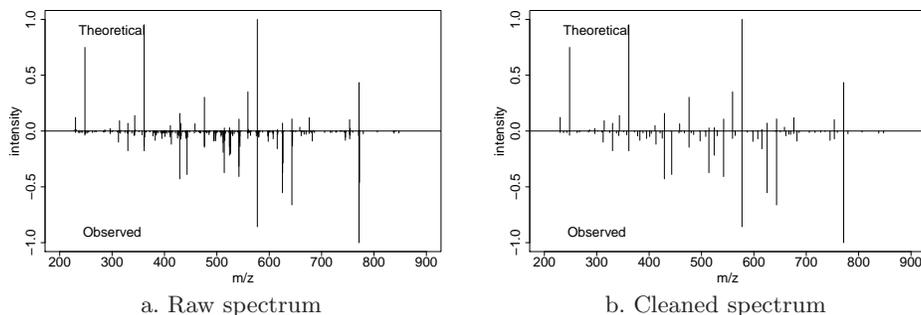


FIG 1. Observed and theoretical spectra of a charge 1+ peptide sequence LVTDLTK. In each plot, top panel is the theoretical spectrum predicted using our implementation based on Zhang (2004), and bottom panel is the observed spectrum. Each spectrum is rescaled by dividing by its highest peak for better visualization. a. Raw spectrum. b. Cleaned spectrum from our preprocessing procedure.

observed spectra (e.g. Figure 1a), the more refined predicted spectra from this model increase the detail with which one can assess similarity of observed and theoretical spectra.

2.2. *Preprocessing.* Observed tandem mass spectra (specifically, those produced by LCQ or LTQ instruments) usually contain a large number of clustered peaks and low-intensity peaks, and have highly variable peak intensities (Figure 1a). These factors pose challenges for developing statistical models. For example, clustered peaks often represent variants from the same fragmentation product (e.g. isotopic peaks) and so are not independent. Preprocessing has been reported to be important for the accuracy of peptide identification (Sun et al., 2007).

Here we use a novel preprocessing procedure that attempts to distill the spectra down to the primary signals, normalizes the peak intensities on all theoretical and observed spectra to a comparable scale, and stabilizes peak intensities. In brief, the procedure distills the spectra down to the primary signals by clustering neighboring peaks and pooling near-by peaks into a single representative peak (Figure 1). Peak intensities in each cleaned spectrum are then normalized by dividing by the 90th percentile of the intensities of the peaks on the spectrum, to put the peaks from different spectra on a comparable scale. Finally, the normalized intensities are transformed by raising to 1/4 power to stabilize the highly variable intensities. We apply the same procedure to both theoretical and observed spectra before scoring. The preprocessing steps are described in detail in Table S1 in Supplementary

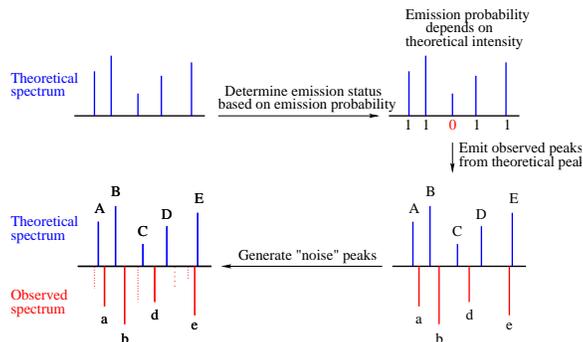


FIG 2. The probabilistic model for generating a random observed spectrum from a given theoretical spectrum, described in Section 2.3.

materials.

2.3. *A probabilistic model.* We now outline the probabilistic model that is the central contribution of this paper. Let $\mathbf{T} = (T_1, \dots, T_n)$ be a predicted theoretical spectrum with n spectral peaks, where $T_i = (X_i^t, Y_i^t)$ denotes the location (X_i^t) and intensity (Y_i^t) for the i th peak, and $X_{i_1}^t \neq X_{i_2}^t$ if $i_1 \neq i_2$. Similarly, let $\mathbf{O} = (O_1, \dots, O_m)$ be an observed spectrum with m spectral peaks, where $O_j = (X_j^o, Y_j^o)$ and $X_{j_1}^o \neq X_{j_2}^o$ if $j_1 \neq j_2$. We find it convenient to assume that the peaks in \mathbf{O} and \mathbf{T} are arbitrarily ordered (that is, they are randomly labelled $1, \dots, n$ in \mathbf{T} and $1, \dots, m$ in \mathbf{O}), rather than being ordered by location for example.

If \mathbf{T} and \mathbf{O} are generated from the same peptide sequence, then we view \mathbf{O} as a distorted (i.e. noisy) realization of \mathbf{T} . Our aim here is to define a probability model $p_\theta(\mathbf{O}|\mathbf{T})$, depending on a set of parameters, θ , described in detail below, that captures this distortion. (Actually in specifying the probability model we condition on the number of peaks, m , in the observed spectrum, so $p_\theta(\mathbf{O}|\mathbf{T})$ should read $p_\theta(\mathbf{O}|\mathbf{T}, m)$, but for notational simplicity we omit the explicit conditioning on m .)

To specify this model, consider generating a random “observed” spectrum \mathbf{O} from $p_\theta(\mathbf{O}|\mathbf{T})$ as follows (Figure 2):

1. Each theoretical peak either does or does not “emit” an observed peak, independently for the n theoretical peaks. We use p_i to denote the probability that the i th theoretical peak, which with slight abuse of notation we abbreviate as T_i , emits a peak. We allow p_i to depend on the intensity of T_i , so $p_i = g(Y_i^t, \theta)$ for some function g , defined below. If T_i emits a peak then the location of the emitted peak is randomly

sampled from a truncated normal distribution centered at X_i^t , and its intensity is randomly sampled from some distribution $f_1(\cdot; \theta)$.

2. Assume that the previous step produces k emitted peaks, where $k \leq n \leq m$. (Typically, including every case we needed to consider in practical applications presented here, $n \leq m$.) Now generate $m - k$ additional “noise” peaks, so that the total number of peaks is m . These noise peaks have locations independently randomly sampled from a uniform distribution across the whole observable m/z range, and intensities independently sampled from a distribution $f_0(\cdot; \theta)$. Note that, despite their name, these noise peaks may represent either measurement noise, or genuine peaks that were simply not included in \mathbf{T} due to limitations of the theoretical prediction model.
3. Randomly label the observed peaks $1, \dots, m$, uniformly on all possible labelings.

The above process is flexible enough to capture several important properties of real data. For example, by letting p_i depend on the intensity of T_i , it captures the fact that high-intensity theoretical peaks are more likely to have matching observed peaks. And by allowing f_1 to be stochastically larger than f_0 , it can take account of the fact that observed peaks that match a theoretical peak will tend to have higher intensities than other observed peaks (Figure 1). Here we assume that g is a logistic function, and use histogram-like density estimates (i.e. piecewise constant densities) for f_0 and f_1 , with parameters of these functions being estimated from data as described below.

The probability $p_\theta(\mathbf{O}|\mathbf{T})$ captures how probable it is that a peptide with theoretical spectrum \mathbf{T} would have resulted in the observed spectrum \mathbf{O} , and is thus a suitable scoring function for comparing different candidate \mathbf{T} s to identify the peptide that created \mathbf{O} . However, although $p_\theta(\mathbf{O}|\mathbf{T})$, described above, is very easy to simulate from, it is tricky to compute for any given \mathbf{T} , because we do not observe which theoretical peak, if any, “emitted” each observed peak. So computing $p_\theta(\mathbf{O}|\mathbf{T})$ involves a computationally-intensive sum over all possibilities.

To formalize this model, let \mathbf{e} denote the unobserved “emission configuration” which identifies which theoretical peaks emitted which observed peaks. Each emission configuration \mathbf{e} determines an emission function e^t , with $e^t(i) = j$, ($j = 1, \dots, m$), if T_i emits O_j , and $e^t(i) = 0$ if T_i does not emit any observed peak. Similarly, \mathbf{e} also determines another emission function e^o , with $e^o(j) = i$, ($i = 1, \dots, n$), if O_j is emitted from T_i , and $e^o(j) = 0$ if O_j is a noise peak. Note that $e^t(\cdot)$ and $e^o(\cdot)$ contain the same information.

Now, we can write $p_\theta(\mathbf{O}|\mathbf{T})$ as a sum over all possible values of \mathbf{e} :

$$(2.1) \quad p_\theta(\mathbf{O} | \mathbf{T}) = \sum_{\mathbf{e}} [p_\theta(\mathbf{O} | \mathbf{T}, \mathbf{e})p_\theta(\mathbf{e}|\mathbf{T})].$$

Here

$$(2.2) \quad p_\theta(\mathbf{e}|\mathbf{T}) = \frac{(m-k)!}{m!} \prod_{\{i:e^t(i)>0\}} g(Y_i^t; \theta) \prod_{\{i:e^t(i)=0\}} (1 - g(Y_i^t; \theta)),$$

where $k \equiv |\{i : e^t(i) > 0\}| = |\{j : e^o(j) > 0\}|$ is the number of emission peaks; and

$$(2.3) \quad p_\theta(\mathbf{O} | \mathbf{T}, \mathbf{e}) = \left(\frac{1}{r}\right)^{m-k} \prod_{\{j:e^o(j)=0\}} f_0(Y_j^o) \\ \times \prod_{\{j:e^o(j)>0\}} [N_T(X_j^o; X_j^t, \sigma^2, w) f_1(Y_j^o)]$$

where r is the length of the m/z range of the uniform distribution on noise peaks, and $N_T(x; \mu, \sigma^2, w)$ denotes the truncated normal distribution, with mean μ , variance σ^2 , truncated at distance w from the mean (here we assume that $w > 0$ is a known constant reflecting the precision of the instrument).

Each term in the sum (2.1) is easy to compute. However, the number of terms is sufficiently large to create computational challenges (even when we take account of the fact that each emitted observed peak must be within $\pm w$ of the corresponding theoretical peak, which does substantially reduce the number of terms, and provides the primary motivation for using a truncated normal distribution rather than a non-truncated normal). To reduce the computation, we replace the likelihood (2.1) with the complete data likelihood under the most probable configuration, i.e.

$$(2.4) \quad \hat{L}(\theta; \mathbf{O}, \mathbf{T}) := \max_{\mathbf{e}} [p_\theta(\mathbf{O} | \mathbf{T}, \mathbf{e})p_\theta(\mathbf{e} | \mathbf{T})].$$

The procedure for searching the most probable configuration and estimating parameters is described in detail in Section 2.5. In general the complete data likelihood (2.4) will be a good approximation to the likelihood (2.1) only if the sum in the latter is dominated by its single biggest term, which will not always be the case. Nonetheless, simulations (Section 3) and empirical evaluation presented below, demonstrate that use of (2.4) produces good performance in practice.

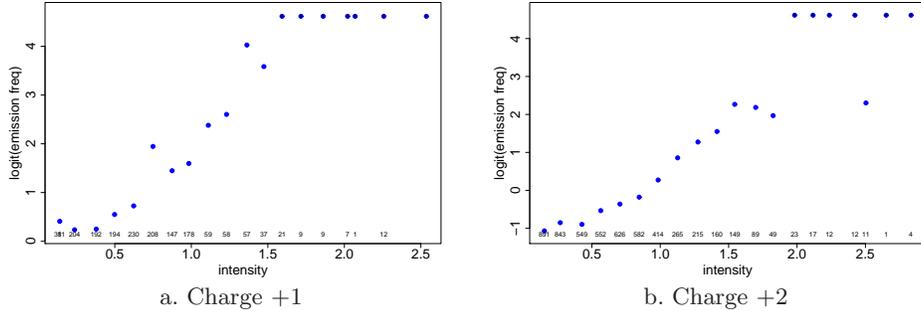


FIG 3. Exploratory data analysis shows an approximate linear trend between theoretical intensities and logit of empirical emission frequencies. The theoretical intensities are pre-processed as described in Section 2.2, and are binned (20 bins) with a fixed binwidth. The emission frequencies are estimated as the proportions of putative matches (i.e. the observed peaks and the theoretical peaks that locate less than 2 Daltons apart) in each bin from training data. To avoid overflow when all peaks in a bin form putative matches, p is bounded at 0.99 at plotting (i.e. Y-axis is bounded at 4.69). The number of observations in each bin is marked at the bottom of the plot.

2.4. *Choice of g , f_0 and f_1 .* As noted above, we assume that g is a logistic function. That is, we allow p_i to depend on y_i^t using a logistic regression:

$$(2.5) \quad \log \frac{p_i}{1 - p_i} = \mu + \beta y_i^t.$$

Here the intercept μ is assumed to be spectrum-specific to take account of the variation across spectra, and the slope β is assumed to be common to all spectra. Exploratory data analysis (e.g. Figure 3) suggests that a logistic form for g is reasonable for our data.

Empirical evaluations on the intensities of peaks on observed spectra show that both f_0 and f_1 are heavy-tailed distributions (Figure 4) where f_1 has a heavier right tail than f_0 . Rather than make specific parametric assumptions regarding f_0 and f_1 , we allow them to take flexible shapes by using piecewise-constant densities. For the data here, we made a specific choice of a piecewise-constant function of 10 bins, where the highest 1% of the intensities of observed peaks in the training set is contained in the 10th bin and the remaining 99% are equally distributed in the remaining 9 bins. The same bin boundaries are used for f_0 and f_1 . These parameters are assumed to be common to all spectra.

2.5. *Parameter estimation, scoring and initialization.* For the applications presented here we used a supervised approach to estimate the param-

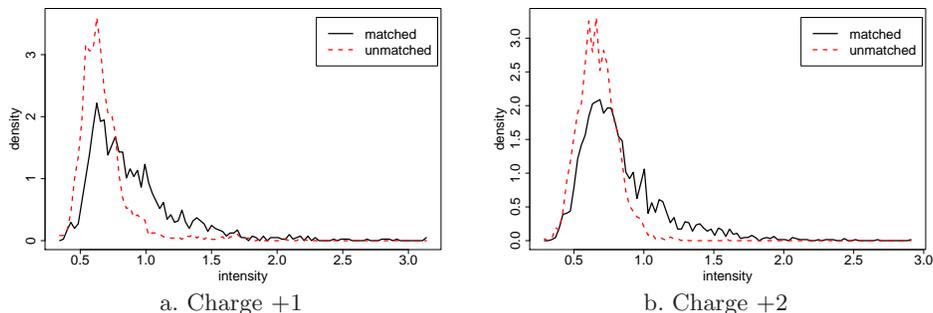


FIG 4. Empirical distributions of intensities for peaks on observed spectra. The emission status of each observed peak is approximated by whether there exists a theoretical peak that is less than $2Da$ apart from the observed peak: Matched: observed peaks within $2Da$ of a theoretical peak; Unmatched: other observed peaks. The observed intensities are normalized and transformed as described in Section 2.2 and are binned (100 bins) with even binwidth for plotting.

eters of our probabilistic model, based on the availability of a training set of N spectrum pairs $(\mathbf{O}_s, \mathbf{T}_s)$, $(s = 1, \dots, N)$, where \mathbf{O}_s and \mathbf{T}_s are known to be generated from the same peptide. However, these parameters could also be estimated in other ways when training data are not available; see Discussion.

We write $\theta = (\theta_0, \mu_1, \dots, \mu_N)$ for the parameters to be estimated, where $\theta_0 = (\sigma^2, \beta, f_0, f_1)$ denotes parameters shared across all spectra pairs in (2.4), and (μ_1, \dots, μ_N) are spectrum-specific intercepts defined in (2.5). Using the training data we estimate $\theta = (\theta_0, \mu_1, \dots, \mu_N)$ by

$$(2.6) \quad \hat{\theta} = \operatorname{argmax}_{\theta} \prod_s \hat{L}(\theta; \mathbf{O}_s, \mathbf{T}_s).$$

When scoring a spectrum \mathbf{T} for an observed spectrum \mathbf{O} , we then compute the score function

$$(2.7) \quad S(\mathbf{T}; \mathbf{O}) := \max_{\mu} \hat{L}(\hat{\theta}_0, \mu; \mathbf{O}, \mathbf{T}),$$

using $\hat{\theta}_0$ estimated from the training stage. The spectra at different charge states are trained and scored separately.

Because the term \hat{L} defined in (2.6) and (2.7) involves both \mathbf{e} and θ , the maximization involves simultaneously maximizing the configuration \mathbf{e} and parameters. Detailed steps are described in Table 4 in Appendix A. As a configuration is updated only when the likelihood is increasing, this

procedure guarantees the likelihood will be nondecreasing throughout the procedures. This procedure usually converged within 30 iterations for the data we tested.

2.6. Uncertainty of identifications. One advantage of our likelihood-based scoring approach is that it leads naturally to an assessment of the confidence that a given peptide generated a given spectrum. Specifically, if we assume that exactly one of the candidates generated the observed spectrum, and all are equally likely *a priori*, then by Bayes theorem the probability that \mathbf{T}_t generated \mathbf{O} is given by

$$(2.8) \quad P(\mathbf{T}_t | \mathbf{O}) = \frac{p(\mathbf{O} | \mathbf{T}_t)}{\sum_{i \in C_o} p(\mathbf{O} | \mathbf{T}_i)},$$

where C_o is the collection of candidate sequences for the observed spectrum \mathbf{O} . In practice we use the scores $S(\mathbf{T}; \mathbf{O})$ in place of $p(\mathbf{O} | \mathbf{T})$ to approximate this expression.

A nice feature of (2.8) is that it weighs the evidence for different candidates directly against each other, rather than against some null hypothesis, which often requires a large number of candidates to obtain reasonable estimates of uncertainty, as in, for example, (Fenyo and Beavis, 2003; Klammer et al., 2009). Thus, for example, if several good candidates provide similarly good matches to the observed spectrum \mathbf{O} , then we could not be confident which candidate generates \mathbf{O} and this uncertainty is appropriately reflected in (2.8). (On the other hand, because equation (2.8) is based on an assumption that exactly one of the candidates produced the observed spectrum, it does not incorporate uncertainty due, for example, to the possibility that the database search entirely missed the correct spectrum. Empirically, however, we have found that the absence of the real sequence from the list of candidates tends not to cause a problem: in such cases confidence measures of all candidates tend to be low.)

3. Simulation. Here we use simulation studies to examine the performance of our approach, particularly to assess the accuracy of estimating parameters using the complete data likelihood (2.4) rather than the actual likelihood (2.1). To do so, we generate observed spectra from theoretical spectra using the probabilistic model described in Section 2.3, then estimate parameters and emission statuses by maximizing the complete data likelihood (2.4) using the estimation procedure in Section 2.5.

In an attempt to generate realistic simulations, we first estimate parameters from a training data of 50 charge +1 and 50 charge +2 spectra (described

TABLE 1

Parameter estimation and accuracy of estimated emission status in simulated data. CE_T is the proportion of misclassified emission labels for peaks on the theoretical spectra after estimation, CE_O is the proportion of misclassified emission labels for peaks on the observed spectra after estimation. Each simulation consists of 50 theoretical spectra and their corresponding observed spectra simulated from the probabilistic model. Mean and standard deviation are computed based on 100 simulations.

	charge +1		charge +2	
	True parameter	Estimated from \hat{L}	True parameter	Estimated from \hat{L}
μ	-1.240	-1.052 (0.119)	-5.060	-5.037 (0.145)
β	2.970	2.929 (0.188)	4.740	4.785 (0.150)
σ	0.390	0.382 (0.008)	0.160	0.157 (0.003)
CE_T	-	0.046 (0.004)	-	0.014 (0.002)
CE_O	-	0.049 (0.004)	-	0.028 (0.003)

in Section 4) for each charge state, using the estimation procedure described in Section 2.5. We then simulate one observed spectrum from each theoretical spectrum in the training set using the estimated parameters, with $0.9n$ noise peaks on each observed spectrum, where n is the number of theoretical peaks. We estimate the parameters and evaluate the accuracy of estimated emission status for each peak on the observed and theoretical spectra. The simulation is repeated 100 times.

The estimated parameters (Table 1) are close to the true values, which indicates the complete data likelihood at the most probable emission configuration provides adequate parameter estimates.

4. Applications.

4.1. *ISB Data.* To illustrate our approach we applied it to a widely-used standard protein mixture, known as the ISB data, for assessing peptide identification (Keller, 2002). This dataset consists of the MS/MS spectra generated from a sample composed of trypsin digest of 18 purified proteins, including 504 charge +1 spectra, 18496 charge +2 spectra, and 18044 charge +3 spectra. The spectra in the dataset have been analyzed using SEQUEST (Eng et al., 1994), a commonly-used software for peptide identification, and a list of 10-11 top-ranked candidates selected by SEQUEST was provided for each spectrum. For a subset of spectra, hand-curation (i.e. manual inspection) has confirmed that the top-ranked peptide assignment from SEQUEST is correct. This subset, which we refer to as the “hand-curated dataset” in what follows, consists of 125 +1 spectra, 1640 +2 spectra, and 1010 +3 spectra. The experimental procedures are described in Keller (2002).

Because we implemented Zhang’s prediction model only for charge +1

and +2 peptides, here we consider only the observed spectra at these charge states, though our scoring method could also be applied to spectra at other charge states (Zhang, 2005). As the resolution of the instrument to generate the spectra in this data set is about 2Da (Wan et al., 2006), we set $w=2\text{Da}$ (e.g. in (2.3) and Table S1 in Supplementary materials).

Because of the computational cost involved in predicting theoretical spectra, the comparison is carried out on only the top 10 candidates selected by SEQUEST rather than all the candidates in the entire database. That is we effectively assess the accuracy of a two-stage procedure that first selects candidates using SEQUEST, and then refines the ranking of the candidates shortlisted by SEQUEST using our likelihood-based score and the similarity index, respectively. Both theoretical spectra and observed spectra are pre-processed using the procedure described earlier (Table S1 in Supplementary materials) prior to scoring with our method and the similarity index.

4.2. Other Methods Compared. We compare the results from our scoring method with the Xcorr score from SEQUEST and a similarity index (I) in Zhang (2004). SEQUEST is one of the most widely-used software for peptide identification. It scores candidate peptides using coarse theoretical spectra and reports multiple scores for each candidate peptide. XCorr score is the main filter score from SEQUEST, defined as $Xcorr = R_0 - \sum_{i=-75}^{i=75} R_i/151$, where R_i is the cross-correlation between the theoretical spectrum and an observed spectrum with lag i (Eng et al., 1994). The similarity index is defined as $I = \frac{\sum_i \sqrt{y_i^o y_i^t}}{\sqrt{\sum_i y_i^o} \sqrt{\sum_i y_i^t}}$, where y_i^o and y_i^t are the peak intensities at (discretized) m/z location i on observed and theoretical spectra, respectively. It was originally proposed in Zhang (2004) for assessing the similarity between a refined theoretical spectrum generated from the kinetic model in Zhang (2004) and its corresponding observed spectrum, and recently was used, in conjunction with other heuristic rules, to validate top-ranked identifications made by common database search algorithms using refined theoretical predictions (Sun et al., 2007; Yu et al., 2010).

4.3. Evaluation on the curated dataset. We first evaluate the performance of our method and the similarity index on the hand-curated dataset. Because, by construction, this dataset includes only spectra that were correctly identified by SEQUEST, we cannot make meaningful comparisons with SEQUEST on these data. For our method, for each charge state, 50 observed spectra are randomly selected as training data, and the remaining spectra are used for testing (resulting in test sets of size 75 for charge +1 and 1590 for charge +2). No training is needed for computing the similar-

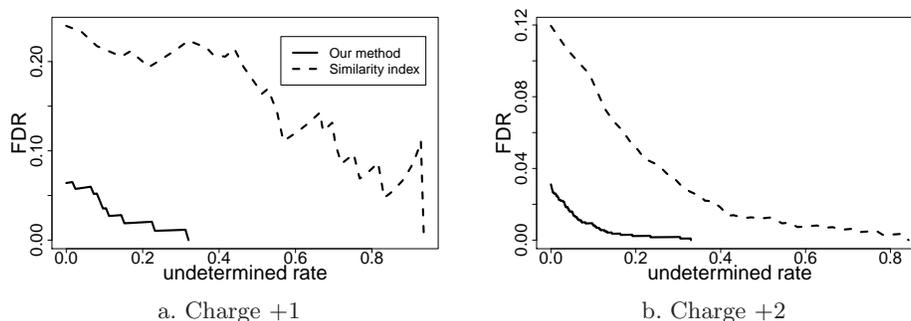


FIG 5. False discovery rate versus undetermined rate for the test data in ISB curated data set. Shown is the comparison between posterior probability from our method (solid line) and the difference in the similarity index between the best and second-best identifications (dashed line).

ity index. It is known that mass spectrometry has difficulty distinguishing several sets of amino acids due to their close or identical mass. Specifically Ile and Leu have identical mass, and Lys is difficult to distinguish from Gln. To allow for these undistinguishable variants, in assessing each method’s performance on this subset, we call an identification correct if the peptide candidate receiving the highest score agrees either with the hand-curated choice or with an undistinguishable variant (i.e. a variant that swaps Ile with Leu and/or Lys with Gln).

Average identification accuracy

Table 2 summarizes identification accuracy for these data. Our model correctly identifies most spectra (96.6% of the spectra in the test set), and performs markedly better than the similarity index (87.6% correct on test set).

High-confidence subsets and calibration of posterior probabilities

Besides providing accurate average performance, two other features of a method are desirable. First, it should provide a meaningful ranking of confidence in different identifications. In particular, it would be helpful if the method were able to identify a subset of high confidence identifications that are highly likely to be correct. Second, it should provide a calibrated assessment of confidence in each individual identification: in our case, one would like the probabilities assigned to individual identifications to be calibrated, so that, for example, of identifications assigned 50% probability of being correct, around half are actually correct.

For these data, our method exhibits both of these desirable properties.

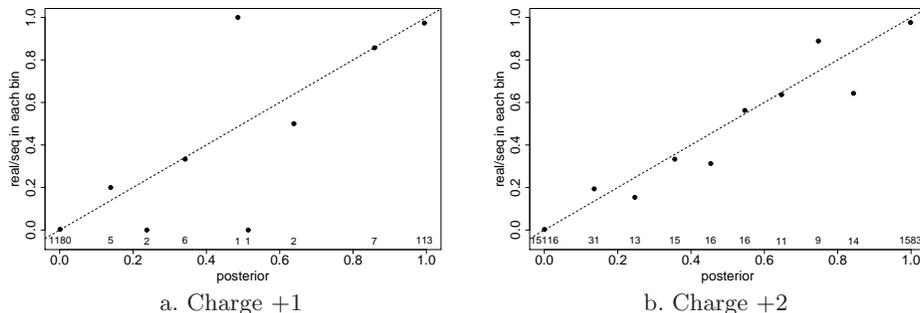


FIG 6. Calibration of posterior probabilities in ISB curated data set. The observations are binned by the assigned probabilities. For each bin, the assigned probabilities (X -axis) are compared with the proportion of identifications that actually correct (Y -axis). The dashed line marks perfect calibration. The number of observations at each point is marked at the bottom of the plot.

First, among identifications scored with the highest confidence by our method ($P(\mathbf{T} \mid \mathbf{O}) \geq 0.99$), 98.7% of spectra are correctly identified (Table 2). Of course, restricting attention to this class of high-confidence calls reduces the overall number of spectra identified: in this case, 93.1% of spectra fall into our high-confidence category, so when we use a calling threshold of 0.99, 6.9% of spectra are “undetermined”. Figure 5 shows the general trade-off between identification accuracy (actually, False Discovery Rate), and undetermined rate, as the calling threshold changes. For comparison, Figure 5 also shows the same trade-off for the similarity index, with confidence in each call measured by the difference in the similarity index between the best and second-best identifications. Although the similarity index is able to provide some meaningful ranking of confidence in each call – as indicated by the lower FDR at more stringent thresholds – the FDR at any given undetermined rate is consistently higher than for our method.

Turning to calibration, Figure 6 shows the calibration of posterior probabilities from our model. To produce this plot, we took all candidate sequences in this dataset (not just the top-ranked sequences) and group them into bins by their posterior probabilities. Within each bin we compare the posterior probabilities with the empirical correct identification rate. The approximate linear trend in Figure 6 shows that, for these data, our method provides reasonably well calibrated probability assessments in both charge states. The ability to produce well-calibrated probabilities of correct identifications is a potential advantage of using likelihood-based scoring rules such as the one we present here, compared with similarity-based scoring rules that do not

TABLE 2

Correct identification rate on the curated ISB dataset. The confident subset consists of testing spectra whose top candidates are highly confident, i.e. $P(\mathbf{T}_{top} | \mathbf{O}) \geq 99\%$.

		Charge +1	Charge +2	All
Likelihood score (S)	train	94.0% (n=50)	100.0% (n=50)	97.0% (n=100)
	test	93.3% (n=75)	96.8% (n=1590)	96.6% (n=1665)
	confident subset	98.3% (n=58)	98.7% (n=1492)	98.7% (n=1550)
Similarity index (I)	test	78.7% (n=75)	88.0% (n=1590)	87.6% (n=1665)

naturally lead to probabilistic assessments of correctness.

4.4. Comparisons with SEQUEST: the benefit of refining identifications.

In this section, we evaluate the benefit of using our likelihood-based score to refine identifications made by SEQUEST, by comparing the results of our method with (unrefined) SEQUEST results, and with the similarity index. Again, because the hand-curated data consists only of identifications that are correctly identified by SEQUEST, it cannot be used to make meaningful comparisons with SEQUEST. Instead, we form a different subset of the ISB data for comparison, based on the fact that these spectra were generated from a known mixture of proteins. Specifically, we take the subset of the (test-set) spectra whose top 10 candidates selected by SEQUEST include at least one subsequence of a constituent protein in the known protein mixture. The resulting subset contains 504 charge +1 spectra and 3669 charge +2 spectra. When assessing a peptide identification for each spectrum, the assumption we make, standard in this context, is that the identified peptide is correct if and only if its amino acid sequence is a subsequence of a constituent protein in the known mixture.

In this dataset, our method correctly identifies substantially more real peptides than SEQUEST and the similarity index (I) for both charge +1 and charge +2 spectra (Table 3). Among them, our method not only correctly identifies most spectra that SEQUEST correctly identifies, but also identifies many of the spectra that are misidentified by SEQUEST. Similar to the hand-curated dataset, here our method is also able to identify a subset of high-confidence high-accuracy identifications (e.g. 98.1% of spectra are correctly identified on this subset, see Table 3) and provides a substantially lower false discovery rate than both SEQUEST and the similarity index at any given undetermined rate (Figure 7).

The gain of using our method to refine SEQUEST results is noticeably higher for charge +1 spectra than for charge +2 spectra. Sun et al. (2007) also observed a higher gain in charge +1 spectra when using the more refined theoretical spectra for peptide identification using similarity index in

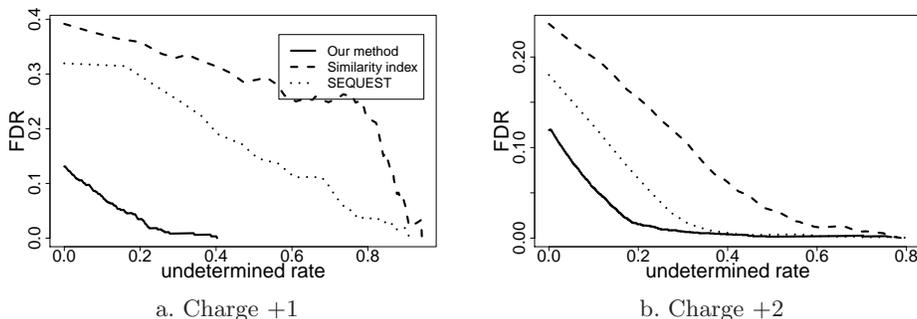


FIG 7. False discovery rate versus undetermined rate for the spectra whose top 10 candidates selected by SEQUEST include subsequence(s) of the constituent proteins in the ISB data. Shown is the comparison between posterior probability computed using our method (solid line), the difference in the similarity index between the best and second-best identifications (dashed line), and Delta Cn of SEQUEST (dotted line), which is the normalized difference between the top two Xcorr values. Xcorr performs worse than Delta Cn in this data set, thus it is not shown.

conjunction with a series of heuristic rules. They interpreted this as being due to the fact that charge +1 spectra tend to follow minor fragmentation pathways, which are excluded in the coarse theoretical spectra but included in the refined theoretical spectra, more frequently than charge +2 spectra (Wysocki et al., 2000), so refined theoretical spectra tend to fill in more information that coarse theoretical spectra miss on charge +1 spectra.

We also note that the similarity index identifies fewer real peptides than SEQUEST in our study. This is different from the results in Sun et al. (2007), who showed more favorable performance for the similarity score compared with SEQUEST and MASCOT, another widely-used peptide identification algorithm, using the similarity scores, in conjunction with other heuristic rules. This disparity could be attributed to a number of differences between their study and ours, such as selection criteria for test spectra (their selection criteria for test spectra enrich for spectra that are easy to discriminate and also involve filtering using the similarity index itself), choice of test datasets, preprocessing procedures (including data transformation) and their use of additional heuristic rules.

5. Discussion. We have developed a likelihood-based scoring approach to peptide identification using database search. This method is based on a rigorous statistical model, which provides a flexible framework to model the

TABLE 3

Correct identification rate for the spectra whose top 10 candidate sequences selected by SEQUEST include subsequence(s) of the constituent proteins in the ISB data. The confident subset consists of testing spectra whose top candidates are highly confident, i.e. $P(\mathbf{T}_{top} | \mathbf{O}) \geq 99\%$.

		Charge +1	Charge +2	All
Likelihood score (S)	test	86.9% (n=504)	87.7% (n=3669)	87.6% (n=4173)
	confident subset	99.1% (n=346)	98.0% (n=3008)	98.1% (n=3354)
SEQUEST	test	68.1% (n=504)	82.0% (n=3669)	80.2% (n=4173)
Similarity index (I)	test	60.7% (n=504)	76.3% (n=3669)	74.4% (n=4173)

fine details and noise structure in the spectra. By taking account of multiple sources of noise in the spectra using a model-based approach, the method makes use of the information of peak intensities on both observed spectra and theoretical spectra predicted by sophisticated chemical principles, in addition to peak locations, in scoring peptide-spectrum matches. Moreover, the use of a likelihood-based score leads naturally to an assessment of the probability that each individual identification is correct. In the ISB data we examined here, the probabilities produced by our method were well-calibrated and produced better identification accuracy than SEQUEST or similarity index.

Our results confirm that finer-scale spectra predicted from comprehensive fragmentation pathways can provide valuable information for peptide identification (Sun et al., 2007) and demonstrate the potential to improve the accuracy of spectra matching by modeling these structures. Similar improvement in peptide identification was also observed in Klammer et al. (2008), who developed a probabilistic model of peptide fragmentation chemistry using dynamic Bayesian network (DBN) and identified peptides using the features learned from DBN using support vector machine (SVM). Similar to Zhang (2004), Klammer et al. (2008) incorporated peptide fragmentation chemistry using widely accepted the mobile proton model (Dongre et al., 1996; Wysocki et al., 2000). It then trained DBNs using positive and negative spectra and generates a set of DBNs to capture the probabilistic relationships governing fragment intensities. Unlike Zhang (2004), it does not produce a theoretical spectrum for each peptide candidate; instead, it yields for each peptide-spectrum match a vector of features from each DBN to be discriminated by the SVM. One advantage of the likelihood-based scoring method over the SVM approach is that it can assess the uncertainty of identification for each identified peptide relative to other candidate peptides for the same observed spectra.

Although we have focussed here on using the likelihood-based score for

improving accuracy of peptide identification by peptide database search, it could also be usefully integrated into many other proteomics analyses that exploit such scores. For example, our score could be easily applied to spectral library search (Lam et al., 2007), where query spectra are identified by matching to a library of previously annotated observed spectra. Here the fact that our score models the fine structure on both query and library spectra should be expected to improve accuracy compared with simpler scoring algorithms. Similarly, peptide-spectrum match scores from our method could be used as input to software that use such scores in downstream analyses, such as identifying the proteins that are likely present in a mixture (Gerster et al., 2010; Keller et al., 2002; Li et al., 2010; Nesvizhskii and Aebersold, 2004; Shen et al., 2008). The improved performance we observed for the likelihood-based score in the peptide identification problem, compared with other scoring rules, should be expected to translate to improved accuracy of downstream protein identification.

While the work described here provides a solid foundation for a rigorous statistical approach to the problem of matching spectra to their generating peptides, there remain many opportunities for further development and refinement. For example, one issue that would need to be tackled in practical applications is how to estimate parameters of the model in the absence of a training set. While simply using parameters estimated from the ISB data used here might perform adequately in some cases, one could almost certainly do better using data generated in the specific context of the experiment to be analyzed. One simple possibility worth investigating would be to use the most confident matches identified by a simpler approach, such as SEQUEST, as a training set; more statistically rigorous approaches, based for example on using an EM algorithm (Dempster et al., 1977) to learn from unlabelled data, could also be developed. While the need to estimate parameters may initially seem like a drawback, it is also responsible for an important advantage of likelihood-based scores: specifically, it makes the likelihood-based score easily adaptable to the varying characteristics of spectra under, for example, different charge states and different machine instrumentation and settings.

Our method requires more computing than some other methods, such as the similarity score and Xcorr, because it involves optimizing the likelihood. The computational cost for identifying each peptide spectrum match is linear in the number of peaks on a cleaned spectrum. With our prototyping implementation in R, it takes on average 0.6s and 1.4s to evaluate each peptide spectrum match for charge 1+ and charge 2+ spectra, respectively. An implementation in C is expected to significantly improve the speed and

makes it more suitable for practical use. A software implementation of the scoring algorithm described here is available from the first author on request.

Appendix A: Estimation and scoring procedure. For any subset I of theoretical peaks, let $\hat{e}^o(I)$ denote the set of observed peaks that could have been generated by the theoretical peaks in I . That is,

$$(5.1) \quad \hat{e}^o(I) = \cup_{i \in I} \{j \in \{1, \dots, m\} : |T_i - O_j| \leq w\},$$

where I is an index set for theoretical peaks. Also define

$$(5.2) \quad L((\theta_0, \mu_1, \dots, \mu_N) \mid \mathbf{O}, \mathbf{T}, \mathbf{e}) = p_\theta(\mathbf{O} \mid \mathbf{T}, \mathbf{e})p_\theta(\mathbf{e} \mid \mathbf{T})$$

Table 4 describes the procedure for estimating parameters and scoring. In the training stage, the entire procedure is carried out. When scoring spectra in the test set, step 2(a) is omitted and $\hat{\theta}_0$ estimated from the training stage is used.

SUPPLEMENTARY MATERIAL

Supplement A: Preprocessing procedure

(<http://lib.stat.cmu.edu/aoas/???/???>). We describe the preprocessing steps in this supplement.

Acknowledgements. We would like to thank two anonymous reviewers and the editor for their constructive comments on improving the quality of the paper.

References.

- Coon, J. J., J. E. Syka, J. Shabanowitz, and D. Hunt (2005). Tandem mass spectrometry for peptide and proteins sequence analysis. *BioTechniques* 38, 519–521.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, B* 39(1), 1–38.
- Dongre, A. R., J. L. Johns, A. Somogyi, and V. Wysocki (1996). Influence of peptide composition, gass-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. *J. Am. Chem. Soc.* 118, 8365–8374.
- Elias, J. and S. Gygi (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4, 207–214.
- Eng, J., A. McCormack, and J. I. Yates (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom* 5, 976–989.
- Fenyo, D. and R. Beavis (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chemistry* 75(4), 768–774.

TABLE 4
Procedure for estimating parameters and scoring.

Let (T, O) denote a pair of theoretical and observed spectra in the training set. For each such pair, do 1(a) and 1(b).

1. Initialization:

- (a) Recall, from the main paper, that e denotes the unobserved “emission configuration” that maps each peak in T to the peak it gave rise to in O (or to no peak if it gave rise to no observed peak). Here we generate a set E containing all possible values for e .
 - i. For each T_i , $i = 1, \dots, n$, find $\hat{e}^\circ(\{i\})$.
 - ii. If $\hat{e}^\circ(\{i_1\}) \cap \hat{e}^\circ(\{i_2\}) \neq \emptyset$ for some i_1 and $i_2 \in \{1, \dots, n\}$, merge both the index sets and the mapped sets, and obtain $\hat{e}^\circ(\{i_1, i_2\}) = \hat{e}^\circ(\{i_1\}) \cup \hat{e}^\circ(\{i_2\})$.
 - iii. Repeat merging, until all mapped sets are mutually exclusive. Suppose G mutually exclusive sets are obtained with corresponding sets of theoretical peaks indexed by I_1, \dots, I_G . The emission configurations of peaks within each putative emission set $(I_g, \hat{e}^\circ(I_g))$, $g = 1, \dots, G$, are determined only by peaks within the set and are independent of peaks in other sets.
 - iv. Let e_g denote the emission configuration e restricted to the set of theoretical peaks in I_g . Thus e_g maps each theoretical peak in I_g to $\hat{e}^\circ(I_g)$. Enumerate all possible values for e_g , and call this set E_g .
- (b) Generate an initial configuration to start Step 2:
 - i. For each putative emission set g , assign $e_g^0 = \operatorname{argmin}_{i \in I_g, j \in \hat{e}^\circ(I_g)} |T_i - O_j|$. Denote the initial configuration as $\mathbf{e}^0 = (e_1^0, \dots, e_G^0)$.

2. Maximization:

In this section, we use the subscript s to label the different spectra in the training set. So the training set consists of pairs $\{(T_s, O_s) : s = 1, \dots, S\}$ and e_s denotes an emission configuration mapping peaks in T_s to peaks in O_s .

Alternate 2(a) and 2(b) until the log-likelihood $\sum_{s=1}^N \log L(\theta_0, \mu_1, \dots, \mu_N)$ converges:

- (a) Estimate spectrum-nonspecific parameters θ_0 :
 - i. $\hat{\theta}_0^{(t)} = \operatorname{argmax}_{\theta_0} \sum_s \log L((\theta_0, \mu_1, \dots, \mu_N) \mid \mathbf{O}_s, \mathbf{T}_s, \mathbf{e}_s^{(t)})$ for current $\mathbf{e}_s^{(t)} = (e_{s,1}^{(t)}, \dots, e_{s,G}^{(t)})$.
- (b) Update configuration and estimate μ_s :

For each pair of spectra \mathbf{T}_s and \mathbf{O}_s , repeat 2b(i-ii) until $\log L((\hat{\theta}_0, \mu_s) \mid \mathbf{O}_s, \mathbf{T}_s, \mathbf{e}_s^{(t)})$ converges.

 - i. Generate a random permutation $\phi = (\phi_1, \dots, \phi_G)$ of $(1, \dots, G)$.
 - ii. Repeat for $g = 1$ to G :
 - A. Fix current configurations $e_{s,\phi_1}^{(t)}, \dots, e_{s,\phi_{g-1}}^{(t)}, e_{s,\phi_{g+1}}^{(t)}, \dots, e_{s,\phi_G}^{(t)}$. For each $e_{s,\phi_g} \in E_{\phi_g}$, define $\mathbf{e}_{s,\phi_g}^{(t)} = e_{s,\phi_1}^{(t)}, \dots, e_{s,\phi_{g-1}}^{(t)}, e_{s,\phi_g}, e_{s,\phi_{g+1}}^{(t)}, \dots, e_{s,\phi_G}^{(t)}$, compute $\hat{\mu}_s = \operatorname{argmax}_{\mu_s} L((\theta_0, \mu_s) \mid \mathbf{O}_s, \mathbf{T}_s, \mathbf{e}_{s,g}^{(t)})$.
 - B. Update $e_{s,\phi_g}^{(t+1)} = \operatorname{argmax}_{e_{s,\phi_g} \in E_g} L((\theta_0, \hat{\mu}_s) \mid \mathbf{O}_s, \mathbf{T}_s, \mathbf{e}_{s,\phi_g}^{(t)})$

- Gerster, S., E. Qeli, C. H. Ahrens, and P. Buehlmann (2010). Protein and gene model inference based on statistical modeling in k-partite graphs. *PNAS* 107(27), 12101–12106.
- Hernandez, P., M. Muller, and R. D. Appel (2006). Automated protein identification by tandem mass spectrometry: issues and strategies. *Mass Spectrometry Reviews* 25, 235–254.
- Keller, A. (2002). Experimental protein mixture for validating tandem mass spectral analysis. *Omic*s 6, 207–12.
- Keller, A., A. Nesvizhskii, E. Kolker, and R. Aebersold (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.* 74, 5383–5392.
- Kinter, M. and N. E. Sherman (2000). *Protein sequencing and identification using tandem mass spectrometry*. Wiley.
- Klammer, A. A., C. Y. Park, and W. S. Noble (2009). Statistical Calibration of the SEQUEST XCorr Function. *Journal of Proteome Research* 8(4, Sp. Iss. SI), 2106–2113.
- Klammer, A. A., S. Reynolds, M. J. MacCoss, J. Bilmes, and W. Noble (2008). Modelling peptide fragmentation with dynamic bayesian networks for peptide identification. *Bioinformatics* 24, i348–i356.
- Lam, H., E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7(5), 655–667.
- Li, Q., M. J. MacCoss, and M. Stephens (2010). A nested mixture model for protein identification using mass spectrometry. *Annals of Applied Statistics* 4(2), 962–987.
- Nesvizhskii, A., A. Keller, E. Kolker, and R. Aebersold (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75, 4646–4653.
- Nesvizhskii, A. I. and R. Aebersold (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discovery Today* 9, 173–181.
- Sadygov, R., H. Liu, and J. Yates (2004). Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal Chem* 76, 1664–1671.
- Shen, C., Z. Wang, G. Shankar, X. Zhang, and L. Li (2008). A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics* 24, 202–208.
- Sun, S., K. Meyer-Arendt, B. Eichelberger, R. Brown, C. Yen, W. Old, K. Pierce, K. Cios, N. G. Ahn, and K. A. Resing (2007). Improved validation of peptide ms/ms assignments using spectral intensity prediction. *Molecular and Cellular Proteomics* 6, 1–17.
- Wan, Y., A. Yang, and T. Chen (2006). PepHMM: A hidden Markov model based scoring function for mass spectrometry database search. *Anal. Chem.* 78, 432–437.
- Wysocki, V. H., G. Tsaprasilis, L. Smith, and L. A. Breci (2000). Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom* 35, 1399–1406.
- Yen, C., S. Houel, N. G. Ahn, and W. Old (2011). Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol. Cell. Proteomics* 10, M111.007666.
- Yu, W., J. A. Taylor, M. T. Davis, L. E. Bonilla, K. A. Lee, P. L. Auger, C. C. Farnsworth, A. A. Welcher, and S. D. Patterson (2010). Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. *Proteomics* 10(6), 1172–1189.
- Zhang, Z. (2004). Prediction of low-energy collision-induced dissociation spectra of pep-

tides. *Analytical Chemistry* 76, 3908–3922.

Zhang, Z. (2005). Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Analytical Chemistry* 77, 6364–6373.

DEPARTMENT OF STATISTICS
PENN STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PA 16802, USA
E-MAIL: qunhua.li@psu.edu

DEPARTMENT OF GENOME SCIENCES
UNIVERSITY OF WASHINGTON
BOX 355065
SEATTLE, WA 98195, USA
E-MAIL: engj@uw.edu

DEPARTMENT OF STATISTICS AND HUMAN GENETICS
UNIVERSITY OF CHICAGO
ECKHART HALL ROOM 126
5734 S. UNIVERSITY AVENUE
CHICAGO, IL 60637, USA
E-MAIL: mstephens@uchicago.edu