Article

Deriving the Probabilities of Water Loss and Ammonia Loss for Amino Acids from Tandem Mass Spectra

Shiwei Sun, Chungong Yu, Yantao Qiao, Yu Lin, Gongjin Dong, Changning Liu, Jingfen Zhang, Zhuo Zhang, Jinjin Cai, Hong Zhang, and Dongbo Bu

J. Proteome Res., **2008**, 7 (01), 202-208 • DOI: 10.1021/pr070479v • Publication Date (Web): 20 December 2007

research articles Journal of proteome research

Deriving the Probabilities of Water Loss and Ammonia Loss for Amino Acids from Tandem Mass Spectra

Shiwei Sun,[§,†] Chungong Yu,[§,†] Yantao Qiao,[†] Yu Lin,[†] Gongjin Dong,[†] Changning Liu,[†] Jingfen Zhang,[‡] Zhuo Zhang,[△] Jinjin Cai,[‡] Hong Zhang,[⊥] and Dongbo Bu*[,†]

Bioinformatics Group, Center for Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, Joint Research Development Laboratory for Advanced Computer and Communication Technologies, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100035, China, and College of Food Science and Biological Engineering, Zhejiang Gongshang University

Received July 30, 2007

In protein identification through tandem mass spectrometry, it is critical to accurately predict the theoretical spectrum for a peptide sequence. The widely used prediction models, such as SEQUEST and MASCOT, ignore the intensity of the ions with important neutral losses, including water loss and ammonia loss. However, ignoring these neutral losses results in a significant deviation between the predicted theoretical spectrum and its experimental counterpart. Here, based on the "one peak, multiple explanations" observation, we proposed an expectation−maximization (EM) method to automatically learn the probabilities of water loss and ammonia loss for each amino acid. Then we employed these probabilities to design an improved statistical model for theoretical spectrum prediction. We implemented these methods and tested them on practical data. On a training set containing 1803 spectra, the experimental results show a good agreement with some known knowledge about neutral losses, such as the tendency of water loss from Asp, Glu, Ser, and Thr. Furthermore, on a testing set containing 941 spectra, the improved similarity between the experimental and predicted spectra demonstrates that this method can generate more reasonable predictions relative to the model that ignores neutral losses. As an application of the derived probabilities, we implemented a database searching method adopting the improved theoretical spectrum model with neutral loss ions estimated. Experimental results on Keller's data set demonstrate that this method can identify peptides more accurately than SEQUEST. In another application to validate SEQUEST's results, the reported peptide−spectrum pairs are reranked with respect to the similarity between experimental and predicted spectra. Experimental results on both LTQ and QSTAR data sets suggest that this reranking strategy can effectively distinguish the false negative predictions reported by SEQUEST.

Keywords: protein identification • tandem mass spectrum • expectation−maximization • neutral loss probability

1. Introduction

Tandem mass spectrometry (MS/MS) has become a powerful tool for the sensitive and high-throughput identification of proteins.[1,2] In a typical MS/MS experiment, proteins of interest are first selected and digested into peptides with an enzyme such as *trypsin*. Then, these generated peptides are separated in a mass analyzer according to their mass to charge ratio (m/z value). During the subsequent collision-induced dissociation (CID) step, these peptides are further fragmented and ionized into a set of ions. The m/z value and intensities of these generated ions are measured and recorded as an experimental MS/MS spectrum.[3]

To date, database searching is one of the widely used methods for peptide identification. A typical database searching method starts by constructing a theoretical spectrum for each peptide in a protein database, then adopts a scoring function to compare this theoretical spectrum with the experimental one, and finally reports the peptides with a score above a threshold as potential solutions.[2−10] Theoretical spectrum prediction is important to database searching since an inaccurate theoretical spectrum will prevent positive protein identification.

1.1. Related Work. There are mainly two types of theoretical spectrum prediction models: one is the chemical kinetic model to simulate the peptide fragmentation process,[11] and the other is the statistical model to learn the rules of the spectrum generation process. For example, Dancik et al. introduced the

* To whom correspondence should be addressed. E-mail: bdb@ict.ac.cn.
§ These two authors contributed equally to this paper.
† Bioinformatics Group.
‡ Joint Research Development Laboratory for Advanced Computer and Communication Technologies.
△ Institute of Biophysics.
⊥ Zhejiang Gongshang University.

10.1021/pr070479v CCC: $40.75
Published on Web 12/20/2007

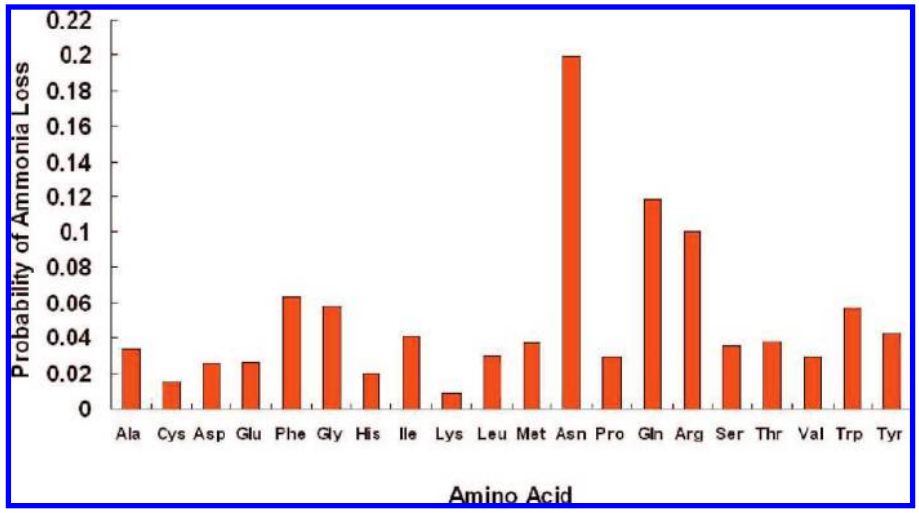

Figure 1. Probability of ammonia loss for each amino acid.

offset frequency function to learn the ion type tendency and the intensity threshold from experimental spectra.[3,8] To comprehensively study the factors influencing the peptide fragmentation process, Schutz et al.[12] fitted training spectra into a linear model, which takes into consideration amino acid type and their position in the peptide. Yates et al.[1] attempted to identify a statistical trend in spectrum peak intensities, including the relationship between the peak intensity and ion type and a specific amino acid's preference for cleavage on its N-terminal or C-terminal bond, etc.[4] Applying a probability decision tree approach, Elias distinguished important factors influencing spectrum generation from a total of 63 attributes of peptide composition and fragmentation.[13] On the basis of expert observations, Huang proposed a rule-based program to enhance cleavage intensity for some specific amino acids and showed its success in peptide identification.[14] The above studies help both the understanding of the complicated fragmentation process and the accurate prediction of theoretical spectra.

There exist some difficulties which hinder an accurate prediction of the theoretical spectrum. First, some atoms have frequently occurring isotopes, causing the *isotopic shift*, i.e., about 1 Da heavier than the common monoisotope. Therefore, an ion may form a series of isotopic peaks because it contains a few heavier isotopic atoms. Second, the frequently observed neutral losses, i.e., loss of a water or an ammonia, lead to some new ions with a 17 or 18 Da deviation from the original ion, respectively.[15] These neutral losses are particularly important to tryptic peptides since the C-terminal Arg or Lys often leads to abundant *y* ions with ammonia loss.[16]

Even though the above-mentioned difficulties are identified, little attention has been given to them, especially to deriving the neutral loss probabilities for amino acids and predicting intensities for ions with neutral losses. Without this quantitative understanding of the spectrum generating process, the widely used database searching algorithms, such as SEQUEST[17] and MASCOT,[18] adopt a simple fragmentation model to predict the theoretical spectrum. For example, SEQUEST assumes that cleavage will occur at peptide bonds in a uniform manner and simply ignores the influence of neutral losses. Ignoring the influence of neutral losses, however, will result in a significant deviation between the predicted spectrum and the experimental one. This paper addresses the neutral losses probability learning problem and how to incorporate these probabilities into a statistical model to accurately predict the theoretical spectrum.

1.2. Our Contributions. Our contributions within this paper are as follows:

First, we proposed an EM method to derive the neutral loss possibilities for amino acids, including ammonia loss and water loss. This method is based on the "one peak, multiple explanations" observation; i.e., the ion with an offset of −17 Da from a *b* ion has two sources: one is an ammonia loss, and the other is a water loss along with an isotopic shift. Experimental results showed a good agreement with some known knowledge on mass spectra, such as that the tendency to lose water for Asp, Glu, Ser, and Thr is much higher than that for other amino acids.

Second, we used these probabilities to design an improved model for theoretical spectrum prediction. In this model, theoretical intensity is estimated for the ions with neutral losses. Experimental results on a testing data set demonstrate that this model can generate a more complete and more realistic theoretical spectrum relative to the model that simply ignores the neutral losses.

Third, as an application of the derived probabilities, we implemented a direct database searching package, called PI^{EM}, in which the intensities of the neutral loss ions are estimated by using the derived probabilities. On an ESI data set provided by Keller,[19] we performed comparison of PI^{EM} with SEQUEST and PI, the original version with neutral loss ions ignored. Experimental results suggest that PI^{EM} can identify peptides more accurately than SEQEUST does.

In addition, we applied this prediction model to distinguish the false positive peptide identification in SEQUEST's output. For each peptide sequence reported by SEQUEST, we used our model to predict the theoretical spectrum and reranked the peptide identification results according to the similarity between the theoretical spectrum and the experimental counterpart. On both LTQ and QSTAR spectra sets, this reranking technique shows its power to distinguish the false positive identification of SEQUEST.

We implemented these algorithms into an open source package PI (Peptide Identifier), which can be freely downloaded from http://www.bioinfo.org.cn/MSMS/.

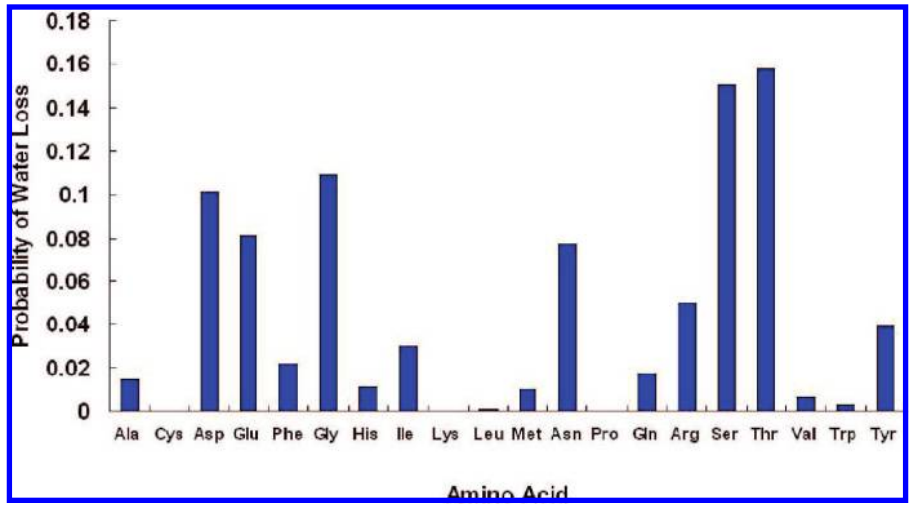

Figure 2. Probability of water loss for each amino acid.

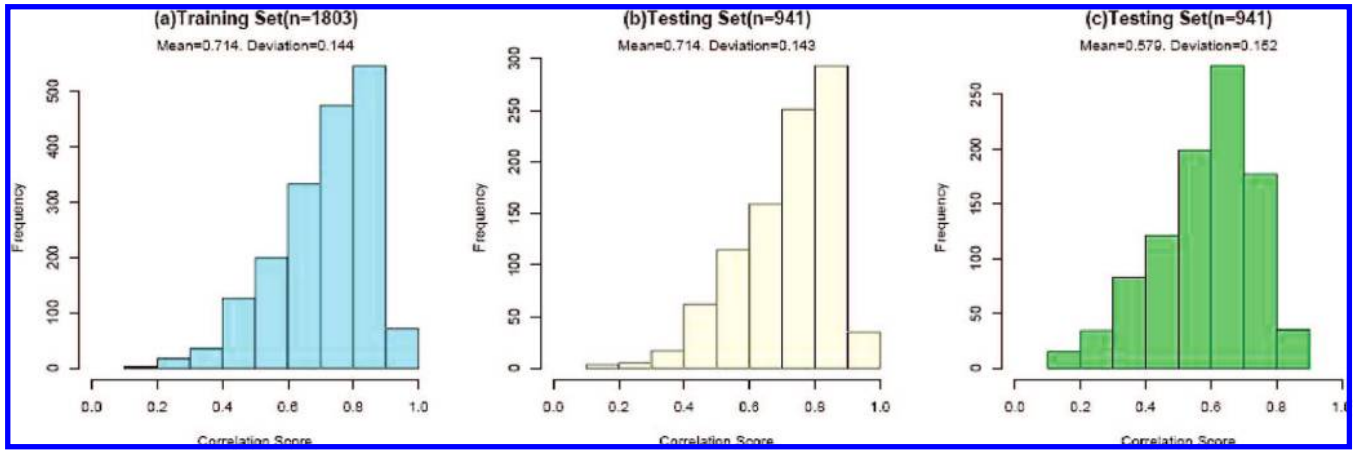

Figure 3. Distribution of correlation scores between experimental spectra and theoretical spectra. (a) The distribution acquired from the training set. (b) The distribution acquired from the testing set. (c) The distribution acquired from the testing set but with neutral loss ions ignored. The correlation scores were calculated by using Pearson correlation coefficient (see formula 2).

2. Methods

In this section, we formulated the neutral loss probability learning problem into an optimization model. On a training data set, this optimization problem aims to derive the neutral loss probabilities by maximizing the likelihood that a peptide generates its paired experimental spectrum. Before describing the optimization model, we give a brief introduction to the ion generating process and neutral losses first.

2.1. Ion Generating Process and Neutral Losses. According to the widely accepted *mobile proton* hypothesis, an ion is generated as described in the following two steps: the migration of the ionizing proton to an amide carbonyl oxygen along the peptide backbone and the cleavage of the N-terminal bond to this amide carbonyl oxygen. The peptide bond cleavage forms a *b* ion or a *y* ion, which depends on whether the N-terminus or C-terminus retains the ionizing proton. Occasionally, an *a* ion is generated from a *b* ion by loss of a carbon monoxide. Other possible backbone ions, such as *c*, *x*, or *z* ions, are not typically generated under the low-energy CID conditions.[4,20,21]

Both the *b* ion and the *y* ion usually have a few variants since some amino acids in the peptide may lose a water or an ammonia. For a charged peptide, the water loss may be generated by dehydrating the −COOH group of the C-terminal Asp or the side chain of Ser or Thr. It has also been observed that the N-terminal Glu may lose a water. Compared with the water losing process, the pathway leading to ammonia loss is much simpler. It has been reported that the ammonia loss occurs on the side chain of Asn, Gln, Lys, and Arg. Since none of the above pathways dominate the spectrum generating process for all the peptides, deriving a quantitative probability of neutral losses for each amino acid will deepen the understanding of neutral losses and improve prediction of the theoretical spectrum as well.[21]

2.2. Neutral Losses Probability Deriving Problem. Let us introduce some notations before describing the formal optimization model. Let $A = \{a_1, a_2, \ldots a_{20}\}$ be the amino acid set, with each amino acid $a \in A$ having a molecular mass $m(a)$. For a peptide $P = p_1 p_2 \ldots p_n$, $p_i \in A$, the cleavage at the ith bond between P_i and P_{i+1} generally forms two ions. One ion is b_i with mass $|b_i| = 1 + \sum_{1 \le j \le i} m(p_j)$, and the other ion is Y_{n-i} with mass $|y_{n-i}| = 19 + \sum_{i+1 \le j \le n} m(p_j)$. Additionally, a b ion with mass x generally has a series of variants, including an ion with mass $x - 18$ by losing a water, an ion with mass $x - 17$ by losing an ammonia, and an isotopic ion with mass $x + 1$. So do the y and a ions.

We use a pair of numbers, (x,h), to denote an ion (also called a peak) in an MS/MS spectrum. In this pair, x is the ion mass, and h is the ion intensity. Thus, the spectrum can be repre-

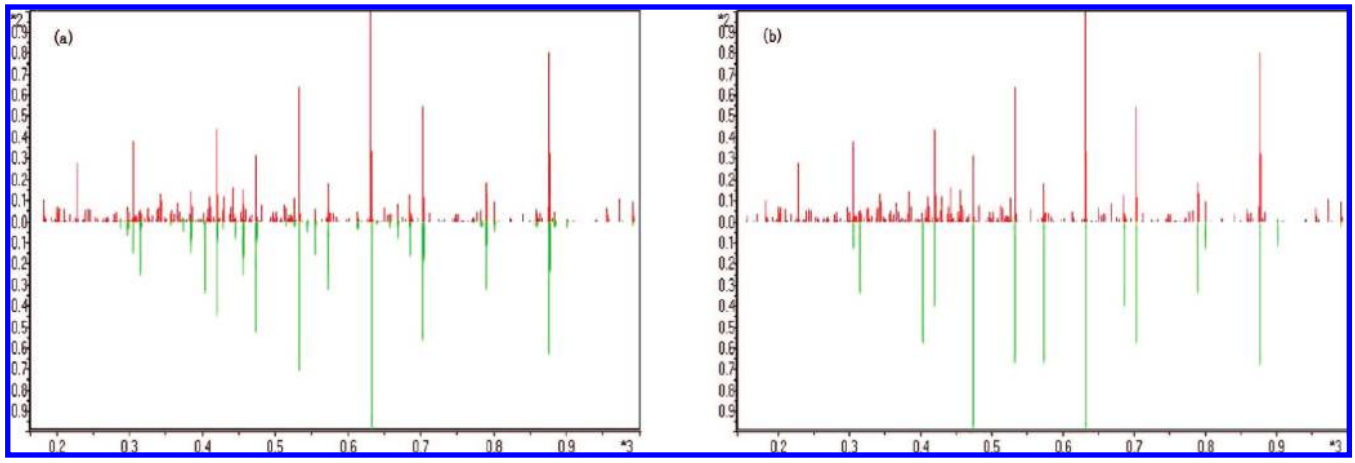

Figure 4. Experimental (above the axis) and theoretical spectra (below the axis) for peptide LDSSAVLDTGK. (a) Theoretical spectrum containing neutral loss ions with intensities estimated by our EM model. Correlation score = 0.702 (b). Theoretical spectrum with the neutral loss ions ignored. Correlation score = 0.567.

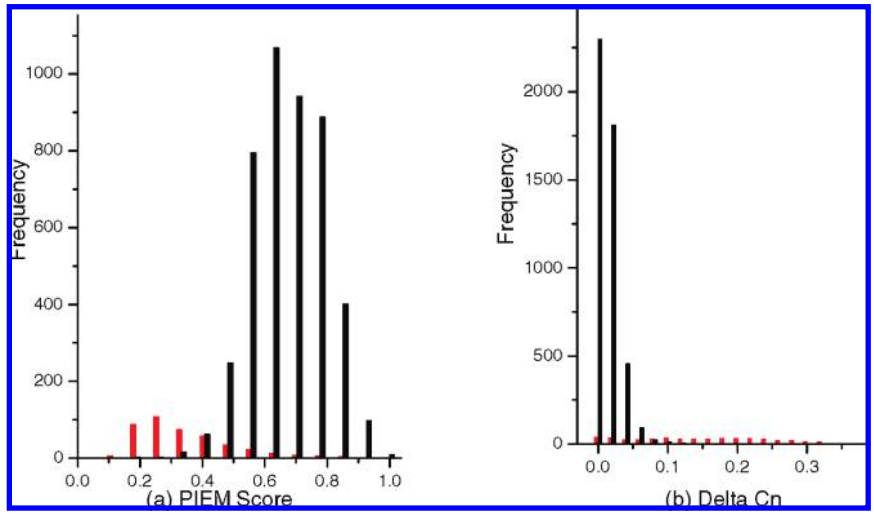

Figure 5. Distributions of the PI^{EM} score (a) and ΔCn (b) acquired from a validation data set. The validation data set contains 5000 spectra randomly selected from the 18 696 doubly charged spectra provided by Keller. The correct hits are shown in red, and the incorrect ones are shown in black.

sented as a peak list $S = \{(x_i, h_i)|1 \le i \le M\}$, where M is the mass of the precursor ion.

For each amino acid a_i, let $\Pr(NH_3|a_i)$ be the probability for this amino acid to lose an ammonia, $\Pr(H_2O)|a_i)$ be the probability to lose a water, and $\Pr(ISO|a_i)$ be the probability of an isotopic shift. For a peptide $P + p_1p_2 \ldots p_n$, $p_i \in A$, the cleavage occurring at the ith bond will form an ion with mass $|b_i| - 17$ with probability

$$\sum_{k=1}^{i} (\Pr(NH_3|P_i) \prod_{t=1, t\neq k}^{i} (1 - \Pr(NH_3|p_i)) \quad (1)$$

The probabilities of forming an ion with mass $|b_i|$, $|b_i| + 1$, $|b_i| - 18$ can be calculated similarly. Therefore, the spectrum generating process, in which many copies of the tested peptide are fragmented into ions, can be treated as a repeat trial under the reasonable assumption that fragmentations of different copies of the peptide are mutually independent. Since each trial produces an ion with a fixed probability, the number of ions observed at different mass conforms to a multinominal distribution. Here, for the simplicity of presentation, we simply but reasonably use the peak intensity h_i as the number of ions with mass x_i.[22] More complex estimations, such as $\log(h_i)$, can also be used without major changes to our algorithm.

The neutral losses probability deriving problem can be formally described as follows:

given a total of K pairs of peptides and the matched tandem mass spectra $M = \{(P_1, S_1), (P_2, S_2), \ldots, (P_k, S_k)\}$, derive the parameters $\theta = \{\Pr(NH_3|a_i), \Pr(H_2O|a_i), \Pr(H_2O|a_i)\}$ to maximize the likelihood $\Pr(M|\theta)$.

2.3. EM Method to Derive Probability of Neutral Losses. Our EM method to derive these probabilities is based on the one peak, multiple explanations observation; i.e., the peak with an offset of −17 Da from a b ion has two sources, an ammonia loss at an amino acid and a water loss along with a 1 Da increase of mass by an isotopic shift. The neutral loss probabilities can be reasonably estimated by determining the contribution of each source to these specific peaks.

For the simplicity of representation, only b ions are considered in the description of our algorithm; y ions are similar and thus omitted in the description. Let $b_{i,j}$ be the intensity of the jth b ion of spectrum S_i and $b_{i,j}^{(d)}$ be the intensity of the peak with an offset d to $b_{i,j}$. Let the hidden variables be denoted as

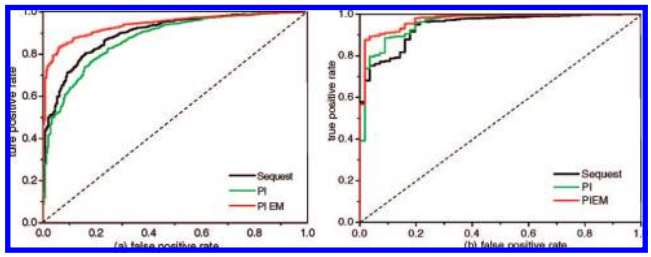

Figure 6. FDR Curves of the SEQUEST (ΔCn) threshold 0.29 (blue), SEQUEST (ΔCn) threshold 0.10 (black), PI (green), and PI^{EM} (red). The *x*-axis denotes the number of reported hits, and the *y*-axis denotes the false positive rate.

$\{W_{i,j}, A_{i,j}, N^{(1)}_{i,j,1}, N^{(1)}_{i,j,2}, \ldots N^{(1)}_{i,j,j}\}$ where $W_{i,j}$ is the contribution of water loss to $b^{(-17)}_{i,j}$; $A_{i,j}$ is the contribution of ammonia loss to $b^{(-17)}_{i,j}$; and $N_{i,j,k}(1 \le k \le j)$ is the contribution of the kth amino acid of peptide P_i to $bi,j^{(1)}$.

E-Step. We estimate the expectation of these hidden variables as follows:

First, $W_{i,j}$ and $A_{i,j}$ are estimated as follows

$$W_{i,j} = b^{(1)}_{i,j} \cdot \frac{w}{w+a}$$

where

$$W = [\sum_{k=1}^{j} \Pr(\mathrm{ISO}|a_{i,k}) \prod_{t=1,t\neq k}^{j} (1 - \Pr(\mathrm{ISO}|a_{i,k}))][\sum_{k=1}^{j} \Pr(\mathrm{H_2O}|a_{i,k}) \prod_{t=1,t\neq k}^{j} (1 - \Pr(\mathrm{H_2O}|a_{i,k}))]$$

$$a = [\prod_{k=1}^{j} (1 - \Pr(\mathrm{ISO}|a_{i,k}))][\sum_{k=1}^{j} \Pr(\mathrm{NH_3}|a_{i,k}) \prod_{t=1,t\neq k}^{j} (1 - \Pr(\mathrm{NH_3}|a_{i,t}))]$$

Hence, the number of the ions with an isotopic shift is $Ni,j^{(1)}$, and the number of the ions without an isotopic shift is $N_{i,j}$. Here

$$N^{(1)}_{i,j} = W_{i,j} + b^{(-16)}_{i,j} + b^{(1)}_{i,j}$$

$$N_{i,j} = b^{(-18)}_{i,j} + A_{i,j} + b_{i,j}$$

Then, $N^{(1)}_{i,j,k}(1 \le k \le j)$ is estimated as follows

$$N^{(1)}_{i,j,k} = N^{(1)}_{i,j} \cdot \frac{\gamma(a_{i,k})}{\sum_{t=1}^{j} \gamma(a_{i,k})}$$

where

$$\gamma(a_{i,t}) = \frac{\Pr(\mathrm{ISO}|a_{i,t})}{1 - \Pr(\mathrm{ISO}|a_{i,t})}$$

M-Step. On the basis of the estimation of these hidden variables, we computed θ to maximize the likelihood by solving the following formula

$$\frac{\Pr(\mathrm{ISO}|a_t)}{1 - \Pr(\mathrm{ISO}|a_t)} = \frac{\sum_{i=1}^{K} \sum_{k=1, a_i,k=a_t}^{|pi|} \sum_{j=k}^{|pi|} N^{(1)}_{i,j,k}}{\sum_{i=1}^{K} \sum_{k=1, a_i,k=a_t}^{|pi|} \sum_{j=k}^{|pi|} N_{i,j}}$$

$\Pr(\mathrm{NH_3}|a_t)$ and $\Pr(\mathrm{H_2O}|a_t)$ can be calculated similarly and are thus omitted here.

3. Results

3.1. Estimating the Probabilities of Water Loss and Ammonia Loss. In this experiment, we applied the EM method to estimate the probabilities of water loss and ammonia loss for amino acids. A spectra data set downloaded from PeptideAtlas, known as A8IP,[23] was used to derive the probabilities. The spectra in A8IP were obtained from the Human Erythroleukemia K562 cell line through an LCQ Classic ion trap mass spectrometer and were converted into DTA format by using TurboSequest. In addition, each spectrum in A8IP has been annotated with a matched peptide by SEQUEST and Peptide-Prophet.[19]

In this proof-of-concept experiment, we restricted our analysis to the doubly charged spectra with a peptide-prophet score above 0.8. As results, we obtained a benchmark data set consisting of 2744 high-confidence peptide−spectrum pairs. These peptide−spectrum pairs were further randomly divided into two disjoint subsets: one is a training set with 1803 peptide−spectrum pairs, and the other is a testing set with 941 peptide−spectrum pairs (see http://www.bioinfo.org.cn/MSMS/ for the Supporting Information.)

On the training set, we applied the EM method described in Section 2 to derive neutral loss probabilities for each amino acid (see Figures 1 and 2). From Figure 1, it can be observed that some amino acids have a high ammonia loss probability, i.e., Asn(0.2), Gln(0.119), and Arg(0.1), while the others have a relatively lower probability to lose an ammonia. This observation is consistent with the reaction pathway analysis reported by Paizs and Suhai.[16,21]

Figure 2 shows the water loss probability for each amino acid, which supports the theoretical and practical observation that Asp(0.1) and Glu(0.081) tend to lose a water from the −COOH group in its backbone and that Ser(0.15) and Thr(0.157) often lose a water from its side chain. In addition, we also note that other amino acids, such as Gly(0.11) and Asn(0.08), may also lose a water.

3.2. Theoretical Spectrum Predicting. For a given peptide P, its theoretical spectrum is predicted through simulating the fragmentation process. More specifically, the number of events at each peptide bond is estimated by PI,[24] a statistical fragmentation model, and the intensity of the ions with neutral losses are estimated using Formula 1. Since the effective temperature of the peptide fragmentation process is unknown under general circumstances,[16] we adopted a rough assumption that a cleavage event generates an N-terminal ion and a C-terminal ion with equal probability.

In this paper, we adopted the following Pearson correlation coefficient function[25] to measure the similarity between a theoretical spectrum t and its experimental counterpart e

$$\text{correlation score} = \frac{\sum_i (s^e_i - \bar{s^e_i})(s^t_i - \bar{s^t_i})}{\sqrt{\sum_i (s^e_i - \bar{s^e_i})^2 \sum_i (s^t_i - \bar{s^t_i})^2}}, \text{ whereas } \bar{s^e_i} = \frac{\sum_i s^e_i}{\sum_i 1}, \bar{s^t_i} = \frac{\sum_i s^t_i}{\sum_i 1} \quad (2)$$

Here s^e_i represents the intensity of the ion with an m/z value of i in the experimental spectrum e, and s^t_i represents the intensity of the same ion in theoretical spectrum t. Since an experimental spectrum always has more peaks than its theoretical counterpart, only the common peaks shared by these two spectra are considered in calculating the above correlation

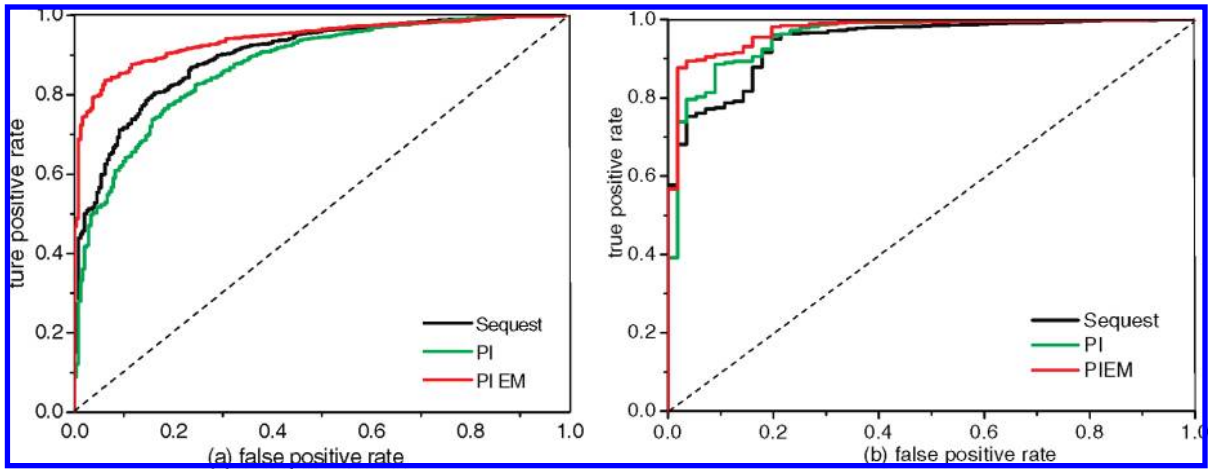

Figure 7. Comparison of ROC plots for SEQUEST (black line), PI (green line), and PI^{EM} (red line) (a). ROC curves on the Q_STAR data set. (b). ROC curves on the LTQ data set. The horizontal axis of a ROC curve is false positive rate = FP / (FP+TN) and the vertical axis is true positive rate= TP / (TP+FN) , where FP is false positive number; TN is true negative number; TP is true positive number; and FN is false negative number.

score. Though the shared ion counts vary according to spectra and peptides, the Pearson correlation coefficients are comparable since they have been normalized.

To validate the parameter estimation of our EM method, we compared two correlation score distributions: one is from the training data set, and the other is from the testing data set (see Figure 3a and 3b). In both cases, the theoretical spectra contain neutral loss ions with intensities estimated by the EM method in this paper. From Figures 3a and 3b, we can see that the correlation score achieves a mean of 0.714 and a standard deviation of 0.144 on the training set and a mean of 0.714 and a standard deviation of 0.143 on the testing set. The similarity between these two distributions demonstrates the validity of this prediction method.[26]

To evaluate the effect of estimating intensities for the neutral loss ions, we first predicted theoretical spectra with neutral loss ions ignored, and then compared these theoretical spectra with the experimental spectra (see Figure 3c). In this case, the correlation score has a mean of 0.58, which is lower compared with the case considering neutral loss ions (see Figure 3b). As a concrete example, the theoretical spectrum was predicted for peptide "LDSSAVLDTGK" and shown in Figure 4a. In Figure 4b, a control case is shown where the ions with neutral losses are ignored. In the first case, the correlation score between the theoretical and experimental spectra is 0.702, while the correlation score is only 0.567 in the second case.

We obtained similar observations when using another spectrum similarity measure, called the Jensen−Shannon divergence,[27] instead of the correlation score. These results suggest that considering neutral loss ions will improve the quality of the theoretical spectrum prediction.

3.3. Database Searching with Neutral Loss Ions Estimated. As an application of the neutral loss probabilities, we implemented a direct database searching package, called PI^{EM}, in which the intensities of the neutral loss ions are estimated by using the derived probabilities. On an ESI data set provided by Keller,[19] we performed a comparison of PI^{EM} with SEQUEST and PI, the original version with neutral loss ions ignored.

The data set contains spectra generated from 22 different LC/MS/MS runs on a sample of 18 known nonhuman proteins mixed in varying concentrations. Each spectrum was searched by SEQUEST against a human protein database with the known protein sequences appended. The top scoring peptide hits against the known 18 proteins, suffering a further manual verification, and were labeled as correct, and the hits to human proteins were labeled as incorrect. In this experiment, we restricted our analysis to the 18 496 doubly charged spectra.

The thresholds of PI and PI^{EM} were set based on the distributions of the Jensen−Shannon divergence score and ΔCn, the score difference between the first hit and the second hit. Specifically, on a validation set with 5000 spectra randomly selected from the data set, the distribution of the Jensen−Shannon divergence score and ΔCn are calculated and shown in Figure 5. From this figure, we can see that there is an obvious gap between the score distributions for correct and incorrect hits. The incorrect hits have a biased ΔCn distribution, while the correct hits have a uniform ΔCn distribution. On the basis of these observations, we set the Jensen−Shannon divergence score threshold to be 0.43, the cross-point of the two score distributions, and the ΔCn threshold to be 0.05, which can filter out most incorrect hits. For SEQUEST, we adopted two widely used threshold configurations, i.e., $\Delta Cn > 0.1$ and $\Delta Cn > 0.29$.

We compared the direct database searching performance of PI^{EM} with SEQUEST and PI. The performances are measured using false discovery rate (FDR), i.e., the ratio of incorrect hits reported. As illustrated by Figure 6, when we control FDR to be 0.1, PI^{EM} returns 10% more hits than SEQUEST with a ΔCn threshold of 0.1 and 26% more than SEQUEST with a ΔCn threshold of 0.29. Similar observations were obtained when FDR was set to other levels, such as 0.05 or 0.15. Take spectrum sergei_digest_B_full_5.2054.2054.2.dta as a concrete example. This spectrum was incorrectly matched to peptide WDNLIYYALGGHK by SEQUEST (Xcorr: 1.6753; SEQUEST ΔCn, 0.19). In contrast, PI^{EM} correctly matched this spectrum to peptide TAGWNIPMGLLYSK (PI^{EM} score, 0.34; PI ΔCn, 0.13). In other words, if considering the same number of top hits, PI^{EM} shows a higher accuracy than SEQUEST and PI.

3.4. Improving SEQUEST by Identifying False Positive Matching. As another application, this improved prediction model can also be used to validate the peptide identification results. During protein identification, SEQUEST compares each given spectrum against peptides in a database and reports a set of peptide−spectrum pairs ordered by their confidence

scores. However, since SEQUEST employs a simple theoretical spectrum prediction model, there are always false positive pairs in the identification results. Here, we attempt to improve peptide predictions by identifying these false positive pairs. More specifically, for each peptide−spectrum pair reported by SEQUEST, we first predict its theoretical spectrum by using the fragmentation model PI[24] and by using the EM model in this paper. Then we calculate the Jensen−Shannon divergence scores[12] between this theoretical spectrum and its experimental spectrum. We rerank the peptide identification results according to this score and report the pairs that have low Jensen−Shannon divergence scores. Ideally, the false positive pairs will be given relatively high scores.

We tested this reranking strategy on two spectrum data sets downloaded from Gygi laboratory:[13] one is an LTQ spectrum data set, and the other a QSTAR spectrum data set. Among the peptide−spectrum pairs reported by SEQUEST, the false positive pairs have been identified through the reverse-database technique;[28] that is, a peptide−spectrum pair is thought to be false positive if the peptide is from the reverse database. We used this reverse-database technique to benchmark our method. Specifically, for the LTQ spectrum set, SEQUEST reports 8639 peptide−spectrum pairs. We used the first 2000 pairs to train our EM model and used the rest as a testing set (6639 pairs, 56 of them have been labeled false positive by the reverse-database technique). Similarly, in the 5865 peptide−spectrum pairs reported by SEQUEST for the QSTAR data set, we chose the first 2000 pairs as a training set and the rest as a testing set (3865 pairs, 242 of them have been labeled as false positive).

We compared SEQUEST, PI, and PI^{EM}. The relationship between the false positive rate and the true positive rate is graphically shown in Figure 7 as receiver-operating characteristic (ROC) plots. From Figure 7a, we can see that when we control the false positive rate to be 0.05 PI^{EM} has a significantly higher true positive rate (0.83) than SEQUEST (0.6) and PI (0.55). Figure 7b suggests similar results. In summary, these ROC curves demonstrate again that the EM method can increase the accuracy of peptide identification by taking neutral losses into consideration, and this reranking technique can discriminate between correct and random matches when validating the results from SEQUEST.

4. Conclusion and Discussion

An accurate prediction of the theoretical spectrum is important to improve the accuracy of identification using database searching methods. However, this process requires the full understanding of the fragmentation process and neutral losses. The EM model in this paper shows that the prediction of ions with neutral losses is feasible and can improve the peptide identification.

Currently, we have not taken the charge-remote fragmentation pathway into consideration and have restricted our efforts on the peptides with two charges. How to incorporate those factors in PI remains an open problem.

Acknowledgment. This work was supported by the National Sciences Foundation of China under grants 60496320, 30500104, and 30570393, the National Key Basic Research and Development Program under grants 2002CB713805 and 2003CB715900, and an opening task of Shanghai Key Laboratory of Intelligent Information Processing Fudan University with No. IIPL-04-001. The authors thank Fuquan Yang and Zhensheng Xie for their valuable help. (All results are available from http://www.bioinfo.org.cn/MSMS/.)

References

(1) Yates, J. R., III *J. Mass Spectrom.* **1998**, *33*, 1–19.
(2) Zhu, H.; Bilgin, M.; Snyder, M. *Annu. Rev. Biochem.* **2003**, *72*, 783–812.
(3) Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17*, S13–21.
(4) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. *Anal. Chem.* **2003**, *75*, 1155–1163.
(5) Sonar http://65.219.84.5/ProteinId.html.
(6) MOWSE http://www.hgmp.mrc.ac.uk/Bioinformatics/Webapp/mowse/mowsedoc.html.
(7) Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406–1412.
(8) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327–342.
(9) Zhang, Z.; Sun, S.; Zhu, X.; Chang, S.; Liu, X.; Yu, C.; Bu, D.; Chen, R. *BMC Bioinf.* **2006**, *7*, 222.
(10) Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8*, 325–337.
(11) Zhang, Z. *Proc. 50th Am. Soc. Mass Spectrom.*; Orlando, Fl,2002. Paper TPE-126.
(12) Schutz, F.; Kapp, E. A.; Simpson, R. J.; Speed, T. P. *Biochem. Soc. Trans.* **2003**, *31*, 1479–1483.
(13) Elias, J. E.; Hass, W.; Faherty, B. K.; Gygi, S. P.; *Nat. Biotechnol.* 20042.
(14) Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **2002**, *219*, 233–244.
(15) Ma, B.; Zhang, K.; Liang, C. *J. Comput. System Sci.* **2005**, *70*, 418–430.
(16) Paizs, B.; Suhai, S. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 103–113.
(17) Yates, J. R., III; Eng, J. K.; McCormack, A. L. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(18) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.
(19) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *Omics* **2002**, *6*, 207–212.
(20) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399–1406.
(21) Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2004**, *5*, 103–113.
(22) Wan, Y.; Chen, T. *RECOMB* **2005**, 342–356.
(23) Cheung, H. T., et al. *Anal. Chem.* **2004**, *76* (13), 3556–68.
(24) Yu, C.; Lin, Y.; Sun, S.; Zhang, Z.; Cai, J.; Zhang, J.; Chen, R.; Bu, D. *Computational Systems Bioinformatics Conference (CSB),* 2006, 353–360.
(25) Feller, W. An introduction to probability theory and its applications, 2nd ed.; Wiley: New York, 1971.
(26) Zhang, Z. *Anal. Chem.* **2004**, *76*, 3908–3922.
(27) Lin, J. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
(28) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. j.; Gygi, S. P. *J. Proteome Res.* **2003**, *2* (1), 43–50.

PR070479V