PepHMM: A Hidden Markov Model Based Scoring Function for Mass Spectrometry Database Search

Yunhu Wan,[†] Austin Yang,[‡] and Ting Chen*,[§]

Department of Mathematics, Department of Pharmaceutical Sciences, and Department of Biology, University of Southern California, Los Angeles, California 90089

An accurate scoring function for database search is crucial for peptide identification using tandem mass spectrometry. Although many mathematical models have been proposed to score peptides against tandem mass spectra, our method (called PepHMM, http://msms.cmb.usc.edu) is unique in that it combines information on machine accuracy, mass peak intensity, and correlation among ions into a hidden Markov model (HMM). In addition, we develop a method to calculate statistical significance of the HMM scores. We implement the method and test them on two sets of experimental data generated by two different types of mass spectrometers and compare the results with MASCOT and SEQUEST under the same condition. One experimental results show that PepHMM has a much higher accuracy (with 6.5% error rate) than MASCOT (with 17.4% error rate), and the other experimental results show that PepHMM identifies 43 and 31% more correct spectra than SEQUEST and MASCOT, respectively.

Mass spectrometry, especially tandem mass spectrometry, has become the most widely used method for high-throughput identification of peptides and proteins. Computational analysis of mass spectrometry data is essential for all applications that are based on this technique. Corresponding methods have also been developed for (1) identification of peptides[2,3,12,13,15−17,20,21,23,25,28,37,43,53−55] and proteins[19,33,36,44] via protein database searches, (2) de novo peptide sequencing,[4,9,14,18,30,32,50] protein sequencing,[6] identification of sequence tags,[18,34,48,49] and decomposition of b and y ions,[8,52] (3) identification of modified or mutated peptides,[22,31,38,39] (4) identification of cross-

* To whom the correspondence should be addressed. Phone: 1-213-7402415. Fax: 1-213-7402424. E-mail: tingchen@usc.edu.

† Department of Mathematics.

‡ Department of Pharmaceutical Sciences.

§ Department of Biology.

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(2) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2* (2), 137−46.
(3) Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17* (Suppl. 1), S13−21.
(4) Bafna, V.; Edwards, N. *Proceedings of the seventh annual international conference on computational molecular biology*, 2003.
(5) Bailey-Kellogg, C.; Kelley, J. J., 3rd; Stein, C.; Donald, B. R. *J. Comput. Biol.* **2001**, *8* (1), 19−36.
(6) Bandeira, N.; Tang, H.; Bafna, V.; Pevzner, P. *Anal. Chem.* **2004**, *76* (24), 7221−33.
(7) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., 3rd. *Bioinformatics* **2004**, *20* (Suppl. 1), I49−54.
(8) Bern, M.; Goldberg, D.; Eigen, M. S. *RECOMB 2005*.
(9) Chen, T.; Kao, M. Y.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8*, 325−37.
(10) Chen, T.; Jaffe, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8* (6), 571−83.
(11) Clauser, K. R.; Baker, P. R.; Burlingame, A. L. *Anal. Chem.* **1999**, *71* (14), 2871−9.
(12) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. *Proteomics* **2003**, *3* (8), 1454−63.
(13) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2002**, *2* (10), 1426−34.
(14) Dancik, V.; et al. *J. Comput. Biol.* **1999**, *6* (3−4), 327−42.
(15) Demine, R.; Walden, P. *Rapid Commun Mass Spectrom.* **2004**, *18* (8), 907−13.
(16) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat Biotechnol.* **2004**, *22* (2), 214−9.
(17) Eng, J. K.; et al. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−89.
(18) Frank, A.; Tanner, S.; Pevzner, P. *RECOMB 2005*.
(19) Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75* (4), 768−74.
(20) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2* (1), 36−47.
(21) Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75* (3), 435−44.
(22) Gatlin, C.; Eng, J.; Cross, S.; Detter, J.; Yates, J. *Anal. Chem.* **2000**, *72*, 757−63.
(23) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3* (5), 958−64.
(24) Gentzel, M.; Kocher, T.; Ponnusamy, S.; Wilm, M. *Proteomics* **2003**, *3* (8), 1597−610.
(25) Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* **2001**, *73* (8), 1676−83.
(26) Huang, Y.; Triscari, J. M.; Pasa-Tolic, L.; Anderson, G. A.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *J. Am. Chem. Soc.* **2004**, *126* (10), 3034−5.
(27) Kapp, E. A.; Schutz, F.; Reid, G. E.; Eddes, J. S.; Moritz, R. L.; O'Hair, R. A.; Speed, T. P.; Simpson, R. J. *Anal. Chem.* **2003**, *75* (22), 6251−64.
(28) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74* (20), 5383−92.
(29) Keller, A.; et al. *Omics* **2002**, *6* (2), 207−12.
(30) Lu, B.; Chen, T. *J. Comput. Biol.* **2003**, *10* (1), 1−12.
(31) Lu, B.; Chen, T. *Bioinformatics* Suppl. 2 (ECCB), 113−121.
(32) Ma, B.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−42.
(33) MacCoss, M. J.; Wu, C. C.; Yates, J. R., 3rd. *Anal. Chem.* **2002**, *74* (21), 5593−9.
(34) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66* (24), 4390−9.
(35) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (5), 422−6.
(36) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75* (17), 4646−58.
(37) Perkins, D. N.; et al. *Electrophoresis* **1999**, *20* (18), 3551−67.
(38) Pevzner, P. A.; et al. *J. Comput. Biol.* **2002**, *7* (6), 777−87.
(39) Pevzner, P. A.; et al. *Genome Res.* **2001**, *11* (2), 290−9.
(40) Rabiner, L. R. *Proc. IEEE* **1989**, *77* (2), 257−286.
(41) Rejtar, T.; Chen, H. S.; Andreev, V.; Moskovets, E.; Karger, B. L. *Anal. Chem.* **2004**, *76* (20), 6017−28.
(42) Von Haller, P. D.; et al. *Mol. Cell. Proteomics* **2003**, *2* (7), 428−42.
(43) Sadygov, R. G.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75* (15), 3792−8.
(44) Sadygov, R. G.; Liu, H.; Yates, J. R. *Anal. Chem.* **2004**, *76* (6), 1664−71.
(45) Schutz, F.; Kapp, E. A.; Simpson, R. J.; Speed, T. P. *Biochem. Soc. Trans.* **2003**, *31* (Pt, 6), 1479−83.
(46) Searle, B. C.; Dasari, S.; Turner, M.; Reddy, A. P.; Choi, D.; Wilmarth, P. A.; McCormack, A. L.; David, L. L.; Nagalla, S. R. *Anal. Chem.* **2004**, *76* (8), 2220−30.

10.1021/ac051319a CCC: $33.50
Published on Web 00/00/0000

linked peptides,[5,10] (5) verification of genes on the genome,[9,47] and (6) preprocessing mass spectra.[7,24,35,41,51] In addition, prediction of peptide fragmentation patterns have also been extensively studied in refs 26 and 27. Reviews of these methods can be found in refs 44 and 45.

The more widely used method is the database search. In a database search framework, candidate peptides from a protein database are generated using specific enzyme digestion. A scoring scheme is used to rate the quality of match between an experimental mass spectrum and a hypothetical spectrum that is directly generated for a candidate peptide sequence from the protein database. If the database is a complete annotation of all coding sequences in a genome, ideally a good scoring function is able to identify the right peptide sequence with the best score. However, the actual MS/MS spectra are complicated because of unknown ion types, unknown charges, missing ions, noise, isotopic ions, and machine errors. As a result, the successful identification of peptide sequences using MS/MS remains a challenging task.

Database search programs that have been developed differ in their methods of computing the correlation score between a spectrum and a peptide sequence. The first program, SEQUEST, developed by Eng et al.[17] used a cross-correlation scoring function. Perkins et al.[37] later developed a program called MASCOT, which introduced a *p*-value based probabilistic scheme to access the significance of peptide identification. Similar programs that use probability-based scoring functions and other methods include Hypergeometric,[43] OMSSA,[23] OLAV,[12] ProbID,[54] Profound,[55] ProteinProspector,[11] SALSA,[25] SCOPE,[3] SHERENGA,[14] and SONAR.[20]

In this paper, we develop a probabilistic scoring function (called PepHMM) to calculate the probability that a spectrum s is generated by a peptide p, $\Pr(s|p)$. This scoring function combines information on correlation among ions and on peak intensity and match tolerance into a hidden Markov model (HMM). The model automatically detects whether there is a match between a mass peak and a hypothetical ion resulting from the fragmentation of the peptide. The detection is based on the local information on the intensity of the matched mass peak and the match tolerance and also on the global information on all matches between the spectrum and the peptide. Because $\Pr(s|p)$ varies in accordance with the density of s, the distribution of the peak intensities, and the mass of the precursor ion, we convert $\Pr(s|p)$ into a Z-score Z that measures the ranking of the score of this peptide among all possible peptides that have the same mass. For a given database, we can easily calculate the E-score E, the expected number of peptides that have a score better than Z.

(47) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, W.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440−5.
(48) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A.; Shevchenko, A. *Anal. Chem.* **2003**, *75* (6), 1307−15.
(49) Tabb, D. L.; Saraf, A.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75* (23), 6415−21.
(50) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11* (9), 1067−75.
(51) Venable, J. D.; Yates, J. R., 3rd. *Anal. Chem.* **2004**, *76* (10), 2928−37.
(52) Yan, B.; Pan, C.; Olman, V. N.; Hettich, R. L.; Xu, Y. *Bioinformatics* **2005**, *21* (5), 563−74.
(53) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67* (8), 1426−36.
(54) Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2* (10), 1406−12.
(55) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, 72 (11), 2482−9.

MATERIAL AND METHODS

Data Sets. We obtained a mass spectra data set from ISB.[29] Two mixtures, A and B, were obtained by mixing together 18 purified proteins of different physicochemical properties with different relative molar amounts and modifications. Twenty-two runs of LC/MS/MS were performed on the data sets, of which 14 runs were performed on mixture A and 8 on mixture B. The data sets were analyzed by SEQUEST and other in-house software tools, with 18 496 peptides assigned to spectra of $[M + 2H]^{2+}$, 18 044 to spectra of $[M + 3H]^{3+}$, and 504 to spectra of $[M + H]^{+}$. The peptide assignments were then manually scrutinized to determine whether they were correct. The final data set contains 1649 curated $[M + 2H]^{2+}$ spectra, 1010 curated $[M + 3H]^{3+}$ spectra, and 125 curated $[M + H]^{+}$ spectra. Fixed on complete trypsin digestion, the data sets have 857 $[M + 2H]^{2+}$ spectra, 646 $[M + 3H]^{3+}$ spectra, and 99 $[M + H]^{+}$ spectra. In this study, we first consider charge 2+ spectra and then apply the same method into charge 1+ and charge 3+ spectra.

We also obtained a spectra data set in A.Y.'s laboratory, which consists of two runs of LTQ data, a total of 20 980 spectra from a mixture of human proteins containing a protein called microtubule-associated protein tau isoform. The data have been interpreted by SEQUEST and MASCOT. We will use this data set to compare PepHMM with SEQUEST and MASCOT.

Analysis of Peak Intensity and Match Tolerance. The distributions of peak intensity and match tolerance play a crucial role in the scoring function. These distributions determine the quality of the match between a mass peak and a hypothetical ion of a peptide.

To obtain information on peak intensity, we use different formats to plot the peak intensity, relative intensity, absolute intensity, and relative ranking. The best characterization of the intensity information is the relative ranking. We compute the relative rankings of mass peaks as follows. We rank mass peaks according to their intensities in a descending order and then normalize them between 0 and 1, where 0 is for the highest intensity and 1 for the lowest. Figure 1 shows the distribution of b ion intensities using the relative ranking (y ions show the same trend). Clearly, the relative ranking of a matched mass peak conforms to an exponential distribution, as shown in Figure 1A. Figure 1B shows a uniform distribution for noise, obtained by excluding the matched mass peaks from the training data set.

The distribution of the match tolerance of b and y ions is shown in Figure 2. Figure 2 shows that this distribution agrees with a normal distribution except that the right-hand side has a small bump at around +1 mass/charge, at which isotopic peaks appear. For simplicity, we use the normal distribution to model the match tolerance for all ions and a uniform distribution for noise.

Framework of Database Search. For the database search, we use a nonredundant protein sequence database called MSDB, which is maintained by the Imperial College, London. We downloaded the release (20051505) that has 2 011 572 protein sequences from multiple organisms. We preprocess the database as follows: for a specific enzyme such as trypsin, we digest in silico every protein sequence into small peptide sequences, and then we index these peptide sequences by their masses. The procedure of the database search follows a standard framework shown in the following paragraphs.

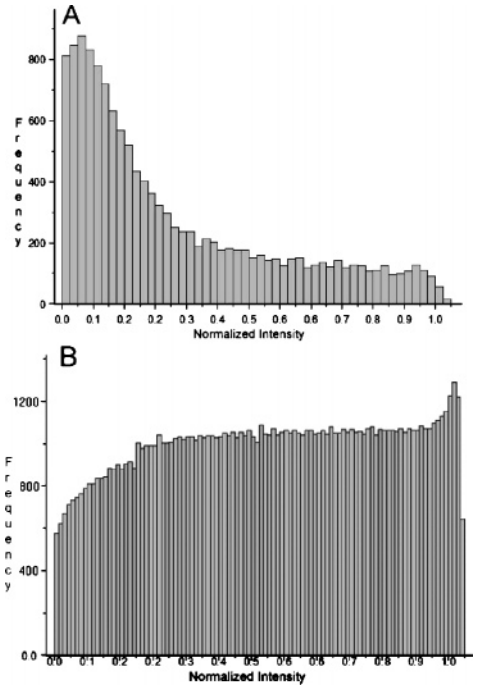

Figure 1. Distribution of peak intensity in the training set. (A) b ion intensity. (B) Noise intensity.

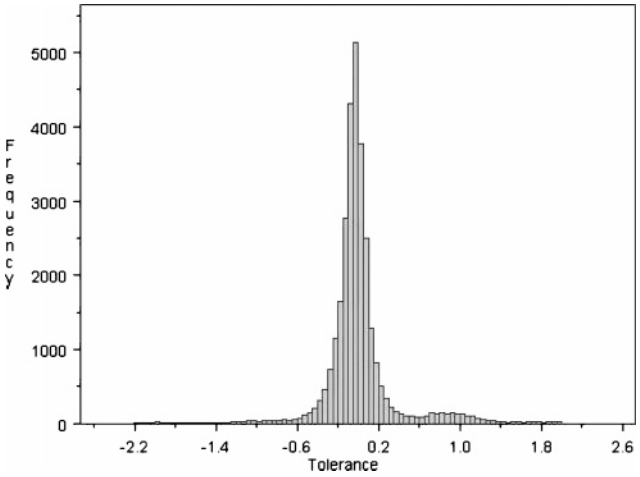

Figure 2. Distribution of match tolerance of b and y ions in the training set.

1. Extracting Peptides. For a given spectrum, we identify candidate peptide sequences the masses of which are within 2 Da of the precursor ion mass, m. The indexing of peptide masses can greatly speeded up this process.

2. Generating Hypothetical Spectra. For each candidate peptide, p, we generate a hypothetical spectrum h without weights (or intensities). In fact, the weights are embedded in the HMM framework. We consider the following seven ions: b, y, b $-$ H_2O, y $-$ H_2O, a, b^{2+}, and y^{2+}.

3. Computing the Probabilistic Score. We compare ions in the hypothetical spectrum with mass peaks in the experimental spectrum. The comparison results in three groups: *match*, where a peak in the experimental spectrum is within a range of mass tolerance of an ion; *missing*, where an ion does not match to any peak; and *noise*, where a mass peak does not match any ion in the hypothetical spectrum. Initially, we use a simple match tolerance threshold ($+/-$ 2 m/z) to classify the comparison into these three groups. Then we apply the initial classification as input for PepHMM. The PepHMM automatically determines whether they are actual matches, missings, or noise, and it returns a score $\Pr(s|p)$. The details of PepHMM are described in Probabilistic Scoring Function.

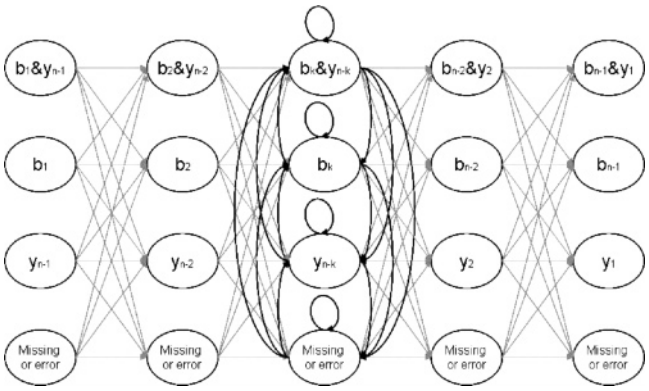

Figure 3. Hidden Markov model for the scoring function.

4. Computing the Z-Score. We simulate 500 random peptides the masses of which are within $[m - 2, m + 2]$, and we calculate HMM scores for these peptides using the above procedure. This simulation is done once for this spectrum. We adjust the HMM scores by the length of the peptides, and we calculate the mean μ and the standard deviation σ. Based on μ and σ, we compute a Z-score Z for peptide p.

5. Computing the E-Value. Given the size of the database, we calculate the expected number of peptides for which the Z-scores are better than Z.

Probabilistic Scoring Function. The notations are defined as follows. Let $s = \{s_1, s_2...\}$ be the given spectrum and p be a candidate peptide with N peptide bonds. For simplicity of description, we assume that only b and y ions are considered. We will describe how to incorporate other ions later. Therefore, the hypothetical spectrum h for p consists of $2N$ ions: $h_1, h_2,..., h_{2N}$. We match h with s into sets of matches, missings, and noise using the following two rules: (1) each ion is either matched to a mass peak within the machine accuracy or labeled as missing, and (2) each mass peak is either matched to the closest ion within the machine accuracy or labeled as noise. In training, we choose the closest mass peak s_j for ion h_i, while in testing, we choose the best mass peak according to the emission probability of $\Pr(s_j|h_i)$. In this paper, we use the probability for both the probability mass function and the probability. The probabilistic scoring function has two components: the matches and the missings as one component and the noise as the second. The probability of the first component can be calculated through an HMM framework. Note that whether s_j matches to h_i depends on the $\Pr(s_j|h_i)$ and other ion matches, and will be determined through the HMM framework.

1. HMM Structure. We model the information of consecutive and composite ions into an HMM framework as shown in Figure 3. For each fragmentation (or position), there are four possible assignments corresponding to four hidden states: (1) both the b ion and the y ion are observed, (2) the b ion is observed but the y ion is missing, (3) the y ion is observed but the b ion is missing, (4) neither of the two ions is observed. The information on

consecutive ions is modeled into the transition probabilities between states, and the information on the composite ions is modeled into hidden states. To deal with different lengths of peptides in the same fashion, we only include five positions of fragmentations here, the first two peptide bonds, the middle peptide bonds, and the last two peptide bonds. Analysis of the training data shows that the middle peptide bonds have similar properties: percentages of observed b ions and y ions, percentages of consecutive ions, and percentages of composite ions.

Given a spectrum and a peptide, we process them into sets of matches, missings, and noise. A mass peak is classified as noise if it is not close to any hypothetical ion. We consider noise peaks independently and identically distributed and having a uniform density. After filtering out noise peaks, the sets of the matches and the missings are the input data to the above HMM structure. Each match is associated with an observation (T, I), where T is the match tolerance and I is the peak intensity. We model (T, I) as the emission of each state. In HMM, the observation is (T, I), and the hidden state is the true assignment of this observation. In the case that (T, I) is not a real match (or a wrong observation), the fourth state emits a missing observation and (T, I) becomes an observation of a noise peak. On the other hand, if there is no observation, the fourth state emits a missing observation only. Note that noise peaks are considered in the model to calculate the score $\Pr(s|p)$. In general, the probability for observing a noise peak is much smaller than that for observing a match. Thus, a peptide with more matches usually has a higher score. For each pair of a spectrum and a peptide (s, p), a dynamic programming algorithm can calculate the probability that s is generated by p.

The HMM method has several advantages. First, the model emphasizes the global assignments of matches. True assignments of observations (the optimal path in HMM) are automatically selected through a dynamic programming algorithm along with the learned parameters. Second, we do not use a hard threshold for match tolerance or peak intensity. Instead, we model them into probability mass functions, of which parameters can be trained through an expectation−maximization (EM) algorithm. Third, we give weights (probability mass function) for matches and use all peaks for comparison (including low-intensity peaks).

Other types of ions ($b - H_2O$, $y - H_2O$, a, b^{2+}, y^{2+}) are also considered separately in our model. Due to the limited size of the training data set, we assume that the appearance of other types of ions is independent.

2. HMM Algorithms. The dynamic programming algorithms and the EM algorithm can be found in Supporting Information.

Significance of HMM Scores. The HMM scores vary in accordance with the length of peptides, the densities of spectra, the distributions of peak intensities, and so on. Here, we propose a general way to compute the significance of an HMM score. This method can be applied to any other scoring function. The central idea is to compute the ranking of a score among all possible scores. Given a spectrum s with precursor ion mass m and machine accuracy δ, we consider all peptides with masses within the range of $[m - \delta, m + \delta]$ as a complete set Q. If we can score every peptide in Q against s using our PepHMM, we can obtain a complete set of HMM scores and easily compute the ranking of the score. However, in general, there is an exponential number of peptides in Q, so just listing every peptide in Q is already unrealistic. For simplicity, we assume that the size of Q is infinite and that the HMM scores (logarithm) of peptides in Q follow a normal distribution. In the following, we describe how to compute the mean and standard deviation for the normal distribution, with which we can calculate the significance of a score.

1. Building a Mass Array. Without loss of generality, we assume that all masses are integers and that every amino acid is independently and identically distributed. Let A be a mass array, where $A[i]$ equals the number of peptides with mass exactly i. We compute A in linear time using the following recursion:

$$A[i] = \sum_{aa} A[i - mass(aa)], A[0] = 1$$

where aa is one of the 20 amino acids and $mass(aa)$ returns the mass of aa. The size of A depends on the accuracy and the measurement range of mass spectrometry machines. In our study, we build an array with an accuracy of 0.01 Da and a range of up to 3000 Da. The size of A is 300 000. We build this array once for all applications. We can easily adapt our method to the case that difference amino acids have different frequencies.

2. Sampling Random Peptides. We describe how to generate a random peptide. First, we randomly select a peptide mass $m' \in [m - \delta, m + \delta]$ using the following probability:

$$A[m'] / \sum_{i=m-\delta}^{m+\delta} A[i]$$

With m', we generate amino acids from the last one to the first one. The last amino acid aa is selected using the following probability:

$$A[m' - mass(aa)] / A[m']$$

We repeat this process to generate a random peptide with mass m'. We sample 500 random peptides, calculate the HMM scores for them, and compute the mean and standard deviation of the normal distribution. This step is done once for a spectrum.

3. Calculating the Z-Score. We use the above normal distribution to calculate the Z-score for each HMM score. The Z-score is a measure of the distance in standard deviations of a sample from the mean.

This approach to the significance of a score is unique in that it assumes a database of random sequences and computes the ranking of a score as its significance. Given a specific database, we can calculate an E-score, the expected number of peptides with scores better than the Z-score.

RESULTS

Training of Parameters. We randomly partition the ISB's 857 $[M + 2H]^{2+}$ data set into a training set with about 687 spectra and a testing set with about 170 spectra (5-fold validation). Using the training set, the EM algorithm converges after 40 iterations. The parameters for the normal distribution of match tolerance are $\mu = -0.0385$ and $\sigma = 0.119$. The parameters for the exponential distributions of peak intensities are $\lambda_b = 4.223$ for b ions and $\lambda_y = 6.421$ for y ions. We also trained the same

Table 1. Comparison of MSDB Search Results by PepHMM and MASCOT on ISB's Charge +2 Data

groups	no. of testing spectra	errors by PepHMM	errors by MASCOT
1	170	2	16
2	173	11	17
3	171	4	12
4	176	8	14
5	171	4	14
6	181	7	18
7	182	6	14
8	171	7	14
9	166	8	15
10	177	5	14
sum	1738	62 (3.6%)	148 (8.5%)

Table 2. Comparison of MSDB Search Results by PepHMM and MASCOT for All of ISB's Data

charge	no. of testing spectra	errors by PepHMM	errors by MASCOT
1	99	15	27
2	857	25	97
3	646	64	155
sum	1602	104 (6.5%)	279 (17.4%)

parameters using other data sets, and the parameters change very little compared to the above values.

Comparison with MASCOT. MASCOT[37] is generally considered to be the best available program for mass spectrometry database search. We compare the accuracy of our program against that of MASCOT using the same database of MSDB. We use 5-fold validation using ISB's charge +2, Trypsin-digested data set as mentioned before, and repeat it 10 times to obtain 10 groups of training and testing sets. For each group, we train the HMM and use the trained HMM for prediction. The parameters trained by the EM algorithm are very similar across all the training sets. We also run MASCOT from its website on these testing spectra. Both programs use MSDB for searches. A prediction is considered to be correct if and only if the correct peptide has the highest score. Table 1 lists the number of testing spectra and the number of errors by PepHMM and MASCOT for each of the 10 groups. Clearly, PepHMM outperforms MASCOT in every group. The average error rate for PepHMM (3.6%) is less than half of that of MASCOT (8.5%).

In addition, we run a thorough test for all of ISB's data using the parameters estimated from the charge +2 data and the same database, MSDB. PepHMM outperforms MASCOT in all three different charges of +1, +2, and + 3. Table 2 shows the number of incorrect predictions by PepHMM and MASCOT. In general, PepHMM's error rate (6.5%) is less than one-third of that of Mascot(17.4%).

Comparison with SEQUEST and MASCOT. Using the parameters trained from ISB's data, we test our program on 20 980 spectra (two runs) generated by the LTQ mass spectrometer at A.Y.'s laboratory and compare the database (MSDB) search results of PepHMM with those of SEQUEST and MASCOT. In the comparison, we used the default parameters set by SEQUEST and MASCOT except that we specified the trypsin digestion and no modification. We set similar parameters for PepHMM. As being

Table 3. Comparison of Spectra and Unique Peptides (Numbers in Parentheses) Correctly Predicted by PepHMM, SEQUEST, and MASCOT on A.Y.'s Data

run	no. spectra	PepHMM	SEQUEST	MASCOT
1	11246	153(16)	110(12)	112(11)
2	9734	95(11)	64(11)	77(13)
total	20980	248(16)	174(14)	189(14)

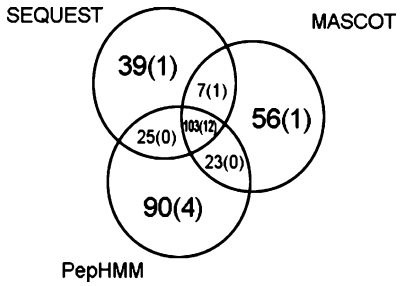

Figure 4. Overlaps of spectra and unique peptides (the numbers in parentheses) predicted by MASCOT, SEQUEST, and PepHMM.

defined before, a prediction is correct if and only if the predicted peptide (with the highest score) is within the target protein sequence of human microtuble-associated protein tau isoform. Table 3 shows the numbers of spectra and unique peptides correctly predicted by the three programs. Figure 4 shows the overlaps of the predictions made by the three programs. PepHMM gives 43% more correct predictions (a total of 248 correct predictions) than SEQUEST (174) and 31% more than MASCOT (189). Moreover, PepHMM predicts more peptide sequences (16) than both SEQUEST (14) and MASCOT (14).

Accessing False Positives. It is also important to estimate the false positive rate of PepHMM for unknown mass spectra. To calculate the false positive rate, we need to construct a *positive* set of annotated mass spectra and a database, as well as a *negative* set in which spectra and the database do not match. We choose the above ISB data and the human protein database plus the 18 purified proteins as a positive set. At the same time, we choose a set of published mass spectra of human proteins from ISB using ICAT experiments[42] and the reversed human protein database as the negative set. This human ICAT data set contains 21 592 charge +2 spectra from 41 runs. The reverse human database contains the reversed protein sequences of human protein sequences. Any match found in the negative set is incorrect.

The histogram of the Z-score distribution of the positive set and the negative set is shown in Figure 5. It reveals that even at a high threshold PepHMM still has a high true positive rate, while the false positive rate becomes very small.

DISCUSSION

We have developed an HMM-based scoring function, PepHMM, for mass spectra database search. We show that this scoring function is very accurate, with a low false positive rate, and that it outperforms both SEQUEST and MASCOT in two large-scale test sets. The HMM structure is flexible in such a way that other ion types can be included. We ran PepHMM on a PC with an AMD Opteron 2.0-GHz CPU and 4-GB RAM. It took less than 1 s to search one spectrum in the human protein database and about 5−10 s to do it in MSDB.

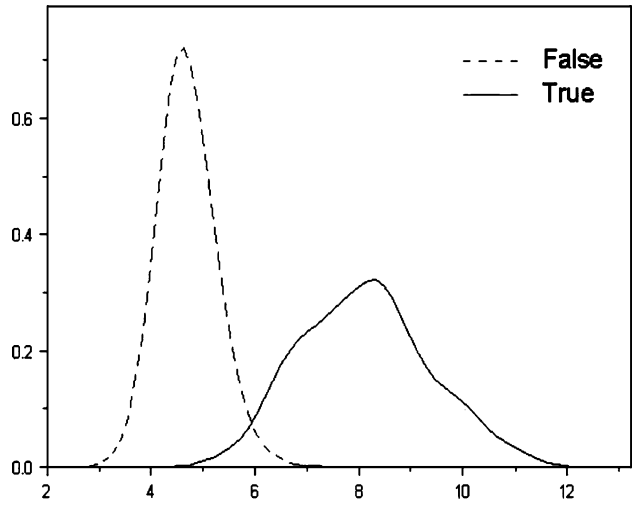

Figure 5. HMM Z-score distribution for the negative set and the positive set.

Currently, we do not separate charge +2 peptides into mobile, half-mobile, and nonmobile due to the limited size of the training data. We do not use the sequence information that is useful for predicting mass peak intensities as having being explored in 27 and 26 because we do not have the training data for this purpose. How we can incorporate these data into our model remains an open question. Another challenge is to score a mass spectrum with posttranslational modifications. Finally, the ultimate goal is to identify proteins. Our next step is to build a protein identification program based on PepHMM.

ACKNOWLEDGMENT

We thank Andrew Keller and Alexey Nesvizhskii from the Institute of Systems Biology for providing us data sets and test results. We thank Debojyoti Dutta for providing web support. This research is partially supported by NIH NIGMS 1-R01-RR16522-01, NSF ITR EIA-0112934, NIH MH59768, NIH AG25323, and the Alfred P. Sloan Research Fellowship.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review July 25, 2005. Accepted November 7, 2005.

AC051319A