RECOMB 2005, Pages 342-356

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,
[§]Department of Biology, University of Southern California, Los Angeles, CA 90089

Abstract

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, PepHMM has a much higher accuracy (with 6.5% error rate) than MASCOT (with 17.4% error rate), and covers 43% and 31% more spectra than SEQUEST and MASCOT, respectively.

1 Introduction

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, 16, 17, 20, 23, 25, 21, 28, 37, 43, 53, 54, 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-linked peptides [5, 10], (5) verification of genes on the genome [9, 47], and (6) pre-processing mass spectra [7, 24, 35, 41, 51]. In addition, prediction of peptide fragmentation patterns have also been extensively studied in [26] and [27]. Reviews of these methods can be found in [45] and [44].

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 matching 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. (1993) [17] used a cross-correlation scoring function. Perkins et al. (1999) [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], OMASS [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

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

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.

2 Material and Methods

2.1 Datasets

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 1,649 curated $[M+2H]^{2+}$ spectra and 1,010 curated $[M+3H]^{3+}$ spectra and 125 curated $[M+H]^{+}$ spectra. Fixed on complete trypsin digestion, the datasets have 857 $[M+2H]^{2+}$ spectra and 646 $[M+3H]^{3+}$ spectra and 99 $[M+H]^{+}$ spectra. In this study, we first consider charge 2+ spectra, then apply the same method into charge 1+ and charge 3+ spectra.

We also obtained a spectra data set in Austin Yang's lab, which consists 2 runs of LTQ data, a total of 20,980 spectra from a mixture of human proteins containing a protein called microtuble-associated protein tau isoform. The data has been interpreted by SEQUEST and MASCOT. We will use this dataset to compare PepHMM with SEQUEST and MASCOT.

2.2 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 1(a). Figure 1(b) shows a uniform distribution for noise, obtained by excluding the matched mass peaks from the training data set.

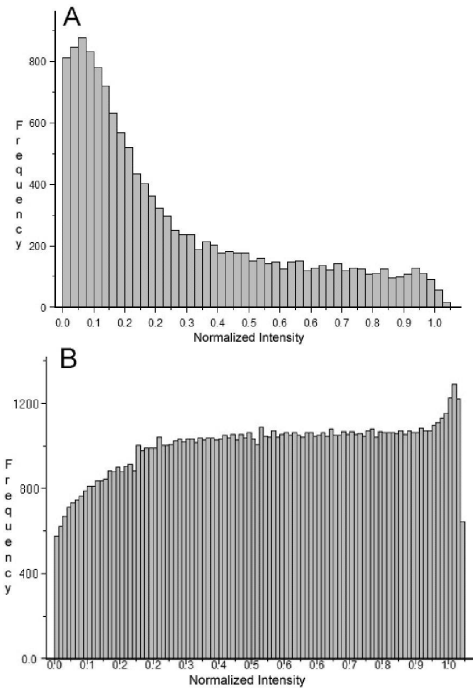

Figure 1: Distribution of peak intensity in the training set. (a) b ion intensity. (b) Noise intensity.

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

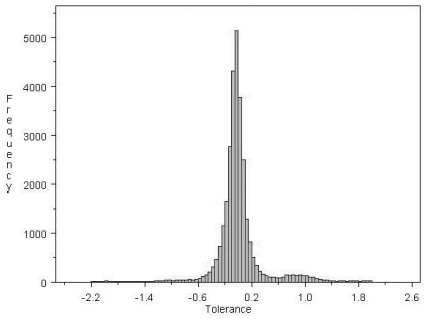

Figure 2: Distribution of match tolerance of b and y ions in the training set.

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.

2.3 Framework of Database Search

For the database search, we use a non-redundant protein sequence database called MSDB, which is maintained by the Imperial College, London. We downloaded the release (20042301) that has 1,454,651 protein sequences from multiple organisms. We pre-process 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.

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 ion, y ion, b-H_2O, y-H_2O, a-ion, 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 Section 2.4.

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.

2.4 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 labelled as missing, and (2) each mass peak is either matched to the closest ion within the machine accuracy or labelled 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 upon the $\Pr(s_j|h_i))$ and other ion matches, and will be determined through the HMM framework.

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 modelled into the transition probabilities between states, and the information on the composite ions is modelled into hidden states. In order 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.

The input to the HMM are: sets of matches, missings, and noise. Each match is associated with an observation (T, I), where T is the match tolerance and I is the peak intensity. We model the information of (T, I) into the emission of each state. In our case, an observation state is the observed (T, I), and a hidden state is the true assignment of this observation. Since (T, I) can be an error observation, the fourth state emits an error observation. On the other hand, if there is no observation, the fourth state emits a missing observation. 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,b^{2+},y^{2+},a) 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.

HMM Algorithms

The dynamic programming algorithms and the EM algorithm can be found in the appendix.

2.5 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 are 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:

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

   where aa is one of the 20 amino acids and $mass(aa)$ returns the mass of aa. The size of A depends upon 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 3,000 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 $m' \in [m-\delta, m+\delta]$

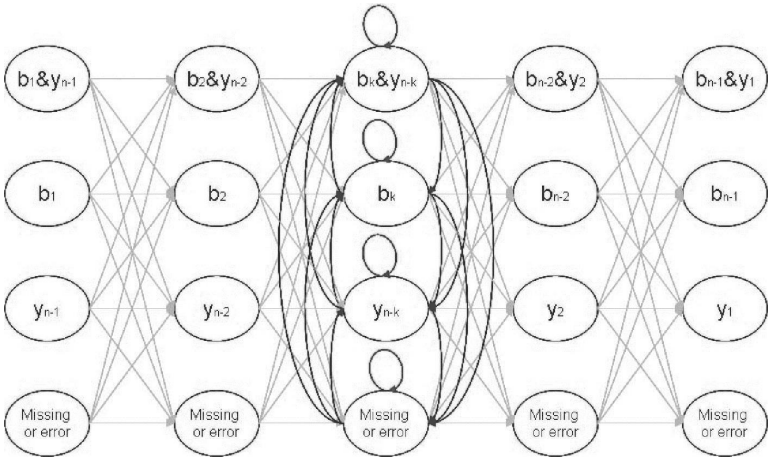

Figure 3: The hidden Markov model for the scoring function.

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.

3 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 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, Trypson-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 ten 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

Table 1: Comparison of MSDB search results by PepHMM and MASCOT on ISB's Charge +2 data

Groups	# 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	1,738	62 (3.6%)	148(8.5%)

Table 2: Comparison of MSDB search results by PepHMM and MASCOT for all of ISB's data

Charge	# 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%)

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 Dr. Austin Yang's Laboratory, and compare the database (MSDB) search results of PepHMM with those of SEQUEST and MASCOT. As being 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 correctly predicted spectra by PepHMM, SEQUEST and MASCOT respectively. PepHMM gave 43% more correct predictions (a total of 248 correct predictions) than SEQUEST (174 predictions), and 31% more than MASCOT (189 predictions). Clearly, PepHMM has a much better coverage than both SEQUEST and MASCOT.

Table 3: Correct predictions made by PepHMM, SEQUEST and MASCOT on Yang's data

Run	# Spectra	PepHMM	SEQUEST	MASCOT
1	11,246	153	110	112
2	9,734	95	64	77
Total	20,980	248	174	189

Table 4: False positive rates and true positive rates for different Z-score thresholds.

Z-Score	False positive rate	True positive rate
4	88.21%	100%
4.5	61.07%	100%
5	26.74%	99.88%
5.5	7.63%	99.26%
6	1.66%	97.06%
6.5	0.3%	93.50%

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 [**?**] 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.

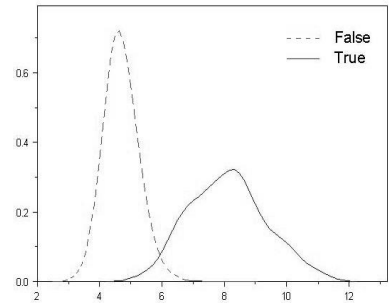

Figure 4: HMM Z-score distribution for the negative set and the positive set.

The histogram of the Z-score distribution of the positive set and the negative set is shown in Figure 4. The false positive rates and the true positive rates are shown in Table 4. Table 4 reveals that even at a high threshold PepHMM still has a high true positive rate, while the false positive rate becomes very small.

4 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 out-

performs both SEQUEST and MASCOT through large-scale test sets. The HMM structure is flexible in such a way that other ion types can be included. Currently, we do not separate charge +2 peptides into mobile, half-mobile and non-mobile 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 post-translational modifications.

5 Acknowledgement

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, and the Alfred P. Sloan Research Fellowship.

References

[1] Aebersold, R., Mann, M. Mass spectrometry-based proteomics, Nature, Vol 422,6928, 2003,198-207.

[2] Anderson DC, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res. 2003 Mar-Apr;2(2):137-46.

[3] Bafna, V., Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database,Bioinformatics,Vol 17, Suppl 1 (2001),S13-21.

[4] Bafna V. and Edwards N. On de novo interpretation of tandem mass spectra for peptide identification. *Proceedings of the seventh annual international conference on Computational molecular biology*, 2003.

[5] Bailey-Kellogg C, Kelley JJ 3rd, Stein C, Donald BR. Reducing mass degeneracy in SAR by MS by stable isotopic labeling. *J Comput Biol.* **8**(1):19-36, 2001.

[6] Bandeira N, Tang H, Bafna V, Pevzner P. Shotgun protein sequencing by tandem mass spectra assembly. Anal Chem. 2004 Dec 15;76(24):7221-33.

[7] Bern M, Goldberg D, McDonald WH, Yates JR 3rd. Automatic quality assessment of Peptide tandem mass spectra. Bioinformatics. 2004 Aug 4;20 Suppl 1:I49-I54.

[8] Bern M, Goldberg D. EigenMS: de novo analysis of peptide tandem mass spectra by spectral graph paritioning. RECOMB 2005.

[9] Chen T, Kao MY, Rush J, Church GM:A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry,J Comput Biol,Vol 8,3 (2001),325-37.

[10] Chen, T., Jaffe, J. and Church, G.M. 2001b. Algorithms for Identifying Protein Cross-links via Tandem Mass Spectrometry. *Journal of Computational Biology*, **8**(6):571-583.

[11] Clauser, K. R., Baker, P. R., and Burlingame, A. L. 1999. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry* **71**(14): 2871-9.

[12] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics. 2003 Aug;3(8):1454-63.

[13] Creasy, D. M.,Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data,Proteomics,Vol 2,10 (2002),1426-34.

[14] Dancik, V., et al. De novo peptide sequencing via tandem mass spectrometry,J Comput Biol,Vol 6:3-4, 1999 ,327-42.

[15] Demine R, Walden P. Sequit: software for de novo peptide sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. Rapid Commun Mass Spectrom. 2004;18(8):907-13.

[16] Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol. 2004 Feb;22(2):214-9. Epub 2004 Jan 18.

[17] Eng, J. K., et al. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,Journal of the American Society for Mass Spectrometry,Vol 5,11 (1994),976-989.

[18] Frank A, Tanner S, Pevzner P. Peptide Sequence Tags forFast Database Search in Mass Spectrometry. RECOMB 2005.

[19] Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem. 2003 Feb 15;75(4):768-74.

[20] Field HI, Fenyo D, Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics. 2002 Jan;2(1):36-47.

[21] Havilio, M., Haddad, Y.,Smilansky, Z. Intensity-based statistical scorer for tandem mass spectrometry,Anal Chem,Vol 75,3 (2003),435-44.

[22] Gatlin, C. and Eng, J. and Cross, S. and Detter, J. and Yates, J. 2000. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Analytical Chemistry*, 72:57-763.

[23] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004 Sep-Oct;3(5):958-64.

[24] Gentzel M, Kocher T, Ponnusamy S, Wilm M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics. 2003 Aug;3(8):1597-610.

[25] Hansen BT, Jones JA, Mason DE, Liebler DC. SALSA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. Anal Chem. 2001 Apr 15;73(8):1676-83.

[26] Huang Y, Triscari JM, Pasa-Tolic L, Anderson GA, Lipton MS, Smith RD, Wysocki VH. Dissociation behavior of doubly-charged tryptic peptides: correlation of gas-phase cleavage abundance with ramachandran plots. J Am Chem Soc. 2004 Mar 17;126(10):3034-5.

[27] Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal Chem. 2003 Nov 15;75(22):6251-64.

[28] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search,Anal Chem,Vol 74,20 (2002),5383-92.

[29] Keller,A., et al. Experimental protein mixture for validating tandem mass spectral analysis,Omics,Vol 6,2 (2002),207-12.

[30] Lu B, Chen T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry, *J Comput Biol*, Vol 10,1 (2003),1-12.

[31] Lu B and Chen T. A suffix tree approach to protein identification via mass spectrometry: applications to peptides of non-specific digestions and amino acid modifications. *Bioinformatics* Suppl. 2 (ECCB), Page 113-121.

[32] Ma, B.,Doherty-Kirby,A., Lajoie, G.,:PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry,Rapid Commun Mass Spectrom,Vol 17,20 (2003),2337-42.

[33] MacCoss MJ, Wu CC, Yates JR 3rd. Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal Chem. 2002 Nov 1;74(21):5593-9.

[34] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994 Dec 15;66(24):4390-9.

[35] Moore RE, Young MK, Lee TD. Method for screening peptide fragment ion mass spectra prior to database searching. J Am Soc Mass Spectrom. 2000 May;11(5):422-6.

[36] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003 Sep 1;75(17):4646-58.

[37] Perkins, D. N., et al. Probability-based protein identification by searching sequence databases using mass spectrometry data,Electrophoresis,Vol 20,18 (1999),3551-67.

[38] Pevzner, P. A., et al. Mutation-tolerant protein identification by mass spectrometry,J Comput Biol,Vol 7,6 (2002),777-87.

[39] Pevzner, P. A., et al. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry,Genome Res,Vol 11,2 (2001),290-9.

[40] Rabiner, L. R. A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition,Proceedings of the IEEE,Vol 77,2 (1989),257-286.

[41] Rejtar T, Chen HS, Andreev V, Moskovets E, Karger BL. Increased identification of peptides by enhanced data processing of high-resolution MALDI TOF/TOF mass spectra prior to database searching. Anal Chem. 2004 Oct 15;76(20):6017-28.

[42] Von Haller, P. D., et al. The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry: II. Evaluation of Tandem Mass Spectrometry Methodologies for Large-Scale Protein Analysis, and the Application of Statistical Tools for Data Analysis and Interpretation,Mol Cell Proteomics,Vol 2,7 (2003),428-42.

[43] Sadygov RG, Yates JR 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem. 2003 Aug 1;75(15):3792-8.

[44] Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem. 2004 Mar 15;76(6):1664-71.

[45] Schutz F, Kapp EA, Simpson RJ, Speed TP. Deriving statistical models for predicting peptide tandem MS product ion intensities. Biochem Soc Trans. 2003 Dec;31(Pt 6):1479-83. Review.

[46] Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, McCormack AL, David LL, Nagalla SR. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Anal Chem. 2004 Apr 15;76(8):2220-30.

[47] Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm W, Vorm O, Mortensen P, Boucherie H and Mann, M. 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proceedings of National Academy of Sciences*, **93**:14440-5.

[48] Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal Chem. 2003 Mar 15;75(6):1307-15.

[49] Tabb, D.L., Saraf, A., Yates, J. R. 3rd. :GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem. Vol 75, 23 (2003), 6415-21.

[50] Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 1997;11(9):1067-75.

[51] Venable JD, Yates JR 3rd. Impact of ion trap tandem mass spectra variability on the identification of peptides. Anal Chem. 2004 May 15;76(10):2928-37.

[52] Yan B, Pan C, Olman VN, Hettich RL, Xu Y. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. Bioinformatics. 2005 Mar 1;21(5):563-74. Epub 2004 Sep 28.

[53] Yates JR 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database,Anal Chem,1995,67,8,1426-36.

[54] Zhang N, Aebersold R, Schwikowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data,Proteomics,Vol 2,10 (2002),1406-12.

[55] Zhang, W.,Chait, B. T. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information,Anal Chem,Vol 72,11 (2000),2482-9.