Published in final edited form as:

J Proteome Res. 2008 July ; 7(7): 2605–2615. doi:10.1021/pr800002u.

Monte Carlo Simulation-Based Algorithms for Analysis of Shotgun Proteomic Data

Hua Xu and Michael A. Freitas *

Department of Molecular Virology, Immunology, and Medical Genetics, The Ohio State University, Columbus, Ohio 43210

Abstract

Two new statistical models based on Monte Carlo Simulation (MCS) have been developed to score peptide matches in shotgun proteomic data and incorporated in a database search program, MassMatrix (www.massmatrix.net). The first model evaluates peptide matches based on the total abundance of matched peaks in the experimental spectra. The second model evaluates amino acid residue tags within MS/MS spectra. The two models provide complementary scores for peptide matches that result in higher confidence in peptide identification when significant scores are returned from both models. The MCS-based models use a variance reduction technique that improves estimation precision. Due to the high computational expense of MCS-based models, peptide matches were prefiltered by other statistical models before further evaluation by the MCS-based models. Receiver operating characteristic analysis of the data sets confirmed that MCS-based models improved the overall performance of the MassMatrix search software, especially for low-mass accuracy data sets.

Keywords

shotgun proteomics; tandem mass spectrometry; database search; Monte Carlo Simulation; amino acid residue tag

Introduction

Tandem mass spectrometry and shotgun proteomics have been widely used to identify and characterize proteins and peptides from complex protein mixtures.¹ In shotgun proteomics, peptides elute from a HPLC column, are ionized by an electrospray ionization (ESI) ion source, and are introduced into a tandem mass spectrometer. Ions are separated in the gas phase based on their mass to charge ratio and fragmented into product ions by various fragmentation techniques.¹ The signature mass spectra of the product ions produced by fragmentation can then be used to identify the sequences and post-translational modifications (PTMs) of the precursor peptide ions.^{2–4}

Database search programs for shotgun proteomic experiments are widely used to identify and characterize proteins, peptides, and their PTMs.^{1,5} Database search algorithms employ a process that involves creation of all proteolytic peptides contained within a protein sequence

* To whom correspondence should be addressed. Dr. Michael A. Freitas, The Ohio State University Medical Center, 460 West 12th Avenue, Columbus, OH 43210. Phone (614) 688–8432, fax (614) 688–8675, freitas.5@osu.edu.

Supporting Information Available: Supplementary Tables 1 and 2 provide the complete lists of peptide matches for the merged LCQ data set and the LTQ-Orbitrap data set from MassMatrix respectively. This material is available free of charge via the Internet at <http://pubs.acs.org>.

database. Theoretical MS/MS spectra are generated for these peptides based on the fragmentation technique used. Experimental LC–MS/MS spectra are then compared with the theoretical spectra created from the sequence database.¹

Several database search algorithms have been developed and automated to analyze high-throughput shotgun proteomic data sets. These programs rely on score models to evaluate potential peptide matches to the experimental spectra and differentiate true peptide identifications from false ones. There are four types of score models that are currently used in these database search programs: descriptive, interpretative, stochastic, and statistical/probabilistic models.¹ SEQUEST⁶ was the first automated approach and is still one of the most commonly used database search programs that use a descriptive score model. PeptideSearch⁷ and GutenTag⁸ are examples of database search programs that employ an interpretative score model. Database search programs that use stochastic models include SCOPE⁹ and OLAV.¹⁰ The statistical/probabilistic score models have gained increased interest, because probability-based scores are a direct measure of the statistical confidence of a peptide match and these scores from different models can be directly compared.¹¹ Examples of programs with statistical/probabilistic score models include Mascot, OMSSA, and X-Tandem as well as others.^{12–20}

At present, most of the automated database search algorithms do not fully utilize the peptide sequence tags inferred from the LC–MS/MS spectra^{7,8} or the information of abundances of peaks in the experimental data. Abundance and sequence tag-based score models used in database search are normally very complex. The abundance- and sequence tag-based score models that have been reported are either empirical models or models with empirically estimated parameters or approximations. Sadygov et al. and Xu et al. independently developed two ion abundance-based score models based on Central Limits Theorem approximations.^{13, 20} Mann et al. developed an empirical model that takes sequence tags into account,⁷ and Tabb et al. developed a database search program that used inferred sequence tags from experimental LC–MS/MS spectra. However, the score model in the algorithm did not account for those sequence tags in the score.⁸

Here, we report an approach based on Monte Carlo Simulations (MCS) to perform the abundance- and amino acid residue tag-based statistical scoring of peptide matches in database searches of shotgun proteomic data sets. Monte Carlo methods are especially useful in solving complex statistical models that cannot be solved without making assumptions or simplifications, such as score models in database search programs involving many variables or variables with unknown distributions. By this approach, complex score models may be implemented in database search software to take into account as much bioinformatics information as possible but yet remain solvable and provide statistical scores without introducing additional assumptions or simplifications. Monte Carlo methods normally suffer high computational expense, and the estimated scores have a variance that is influenced by sampling size. To make the methods practical for use in high-throughput database search programs, the computational algorithms for the methods need to be optimized for speed and must employ variance reduction techniques to obtain reasonable estimation precision.

We have developed two score models based on Monte Carlo Simulations and incorporated them in the newly developed database search program, MassMatrix. The pure statistical models involve no empirical parameters. To lower the computational expense, potential peptide matches were prefiltered by the score models in MassMatrix before being scored by the MCS-based models. The simple variance reduction technique was implemented to improve the estimation precision and to make the approach practical for use in automated database searches.

Experimental

Sample Preparation and Mass Spectrometry

Bovine histones were isolated from bovine thymus tissue as described by Sures et al.^{21,22} The mixture of bovine histones was digested by trypsin in 100 mM ammonium bicarbonate buffer (pH = 8.0). Enzymes were used in 25:1 ratio (substrate/enzyme), and the mixture was incubated at 37 °C for two hours. The digested peptides were identified by use of data-dependent LC–MS/MS on a LCQ Deca XP ion trap and a LTQ-Orbitrap mass spectrometer (ThermoFisher, San Jose, CA). Two microliters of bovine histone peptides with a total concentration of 0.1 µg/µL were injected into the LC–MS/MS system and eluted off the capillary HPLC column into the mass spectrometer with a linear gradient. Solvent A was water with 0.1% acetic acid and solvent B was acetonitrile with 0.1% acetic acid. Ions were fragmented by use of collision induced dissociation (CID).

Database Search and Search Parameters

The .RAW data files obtained from the mass spectrometer were converted to mzXML files by use of ReAdW (<http://tools.proteomecenter.org/ReAdW.php>). The mzXML files were then converted to .MGF files by use of an in-house developed program mzXML-2MGF (www.massmatrix.net). The .MGF files from multiple experiments were merged into a single .MGF file by use of a Perl script. For low mass accuracy data collected on the LCQ Deca XP ion trap mass spectrometer, LC–MS/MS spectra that were not derived from singly charged precursor ions were searched as both doubly and triply charged precursors. The data sets in .MGF formats are available at www.massmatrix.net. The merged LCQ data set from 23 separate runs containing 62 117 MS/MS spectra, and the merged LTQ-Orbitrap data set from 5 separate runs containing 4606 MS/MS spectra were searched by use of the online version of MassMatrix (www.massmatrix.net) against a database containing both the bovine histone database (117 protein sequences) and a decoy reversed human database (96 997 decoy protein sequences) with the following options: (i) No variable or fixed modifications; (ii) Enzyme: trypsin; (iii) Missed Cleavages: 3; (iv) Peptide Length: 6–40 amino acid residues; (v) Mass tolerances of 2.0 Da and 20 ppm for the precursor ions on LCQ Deca XP ion trap and LTQ-Orbitrap mass spectrometers, respectively; and (vi) Mass tolerances of 0.8 Da for the product ions.

The two merged data sets were also evaluated by use of Mascot (www.matrixscience.com), OMSSA (<http://pubchem.ncbi.nlm.nih.gov/omssa/>), and X!Tandem (<http://www.thegpm.org/TANDEM/>). Each search program was loaded onto the same computer and ran in serial mode using a single CPU. The counterpart search parameters in Mascot, OMSSA, and X!Tandem were identical to those in MassMatrix except that the mass tolerance for the precursor ions on LTQ-Orbitrap was set as 0.02 Da for the OMSSA searches. For X!Tandem searches, refinement was enabled and performed for the peptide matches with expectation values greater than or equal to 1.0. Results from MassMatrix and Mascot were output as html files. Results from X!Tandem and OMSSA were output as pepXML format. The scan number, charge, calculated mass, observed mass, mass difference, missed cleavages, scores, and peptide sequences were extracted from the original output file into a tab delimited .TXT file by use of Perl scripts for receiver operating characteristic (ROC) analysis. The Perl scripts are available at www.massmatrix.net. For each spectrum, the highest scoring peptide match was assumed to be the best peptide hit. All outputted peptide matches without any filtering from the four search algorithms were considered when evaluating the score models. The score in Mascot, e-value in OMSSA, and expectation value in X!Tandem were considered during ROC analysis. The decoy reversed human database creates ~1000 times as many theoretical peptides as the bovine histone database. False positive peptide matches from the bovine histone database were assumed to be negligible. Therefore, the peptide matches

returned from the bovine histone database were considered as true positives (TPs) while those from the decoy database were considered as false positives (FPs).²³ ROC curves were created by plotting the number of TP against the number of FP as the score thresholds decreased in the search results.

Results and Discussion

Abundance-Based Score via Monte Carlo Simulation. Statistical Model Based on Ion Abundance

The ion abundance-based score model via MCS was derived from the probabilistic model in the MassMatrix software as described previously.¹³ For a pair of spectra, one experimental and one theoretical, the statistical model tests two hypotheses: the null hypothesis, H_0 , stating that a match is random, that is, the theoretical spectrum is independent of the experimental; and the alternative hypothesis, H_A , stating that the match is not random, that is, the theoretical spectrum is related to the experimental. To describe the test, we define W_e and W_t as the precursor masses of the experimental and theoretical spectra respectively. The scoring process is performed in two stages as described previously.¹³ In the first stage, the model matches W_e against W_t within the specified precursor ion mass accuracy, τ_{pep} . Potential peptide matches that pass stage 1 will be matched in stage 2 where the model then matches all product ions in the experimental spectra against the theoretical spectra within the specified mass accuracy, τ_{msms} . For those matches that do not pass stage 1, the second stage is skipped and their scores are assigned a value of 0.

During the second stage, the variable b_i ($i = 1, 2, \dots, n$) is defined for the i^{th} product ion in the experimental spectrum as follows:

$$b_i = \begin{cases} 1 & \text{peak } i \text{ matches} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where n is the number of product ions in the experimental spectrum. Under H_0 , all matched product ions are independent random occurrences and the probability that a product ion randomly matches any of the product ions in the theoretical spectrum is:

$$p_2 = \frac{\Pi_{\text{theo}}}{\Pi} \quad (2)$$

where Π_{theo} is the total coverage of the mass detection range for all product ions in the theoretical spectrum and Π is the MS/MS mass detection range of the mass spectrometer. The mass window for each product ion in the theoretical spectrum is $\pm \tau_{\text{msms}}$ ($2 \times \tau_{\text{msms}}$). Accordingly, Π_{theo} can be calculated by use of the following equation

$$\Pi_{\text{theo}} = 2m \times \tau_{\text{msms}} \quad (3)$$

where τ_{msms} is the product ion mass accuracy and m is the number of product ions within the detection range in the theoretical spectrum. Π_{theo} is essentially the sum of the widths of all product ion mass windows in the theoretical spectrum given that the product ion mass windows

have the same width and there is no overlap in mass windows. Substitution of eq 3 into eq 2 yields

$$p_2 = \frac{2m \times \tau_{\text{msms}}}{\prod} \quad (4)$$

Because the theoretical spectrum is independent of the experimental spectrum under H_0 , all b_i ($i = 1, 2, \dots, n$) have identical and independent Bernoulli (p_2) distributions under H_0 .

The variable x is defined as the number of product ions in the experimental spectrum that match the theoretical spectrum (eq 5) in stage 2, with b_i values used as defined in eq 1.

$$x = \sum_{i=1}^n b_i \quad (5)$$

x has a binomial (n, p_2) distribution due to the fact that all b_i have identical and independent Bernoulli (p_2) distributions under H_0 . Consequently, its probability mass function is:

$$p(x) = C_n^x p_2^x (1 - p_2)^{n-x} \quad (6)$$

where p_2 is calculated from eq 4.

The variable Y is defined as the total abundance of the experimental product ions that match product ions within a given theoretical spectrum:

$$Y = \sum_{i=1}^n I_i b_i \quad (7)$$

where I_i is the abundance of the i^{th} product ion in the experimental spectrum and b_i is defined in eq 1.

The model next evaluates whether the total abundance of matched product ions in the experimental spectrum could be a random occurrence. The p value, α , is the probability that the total abundance of matched product ions in the experimental spectrum, Y , for a random match is greater than or equal to that of the actual match, I_{match} , under H_0

$$\alpha = P(Y \geq I_{\text{match}}) \quad (8)$$

The statistical score, β , defined in the model is the negative logarithm of the p value

$$\beta = -\log(\alpha) \quad (9)$$

According to the Law of Total Probability, the p value can be calculated by the following equation

$$\alpha = P(Y \geq I_{\text{match}}) = \sum_{x=0}^n [P(Y \geq I_{\text{match}} | x) p(x)] \quad (10)$$

where $p(x)$ is the probability mass function for x as shown in eq 6, and $P(Y \geq I_{\text{match}} | x)$ defines the probability that Y for a random match is greater than or equal to that of the actual match, I_{match} , given that the number of matched product ions in the experimental spectrum is equal to x under H_0 . Substitution of eq 6 into eq 10 yields

$$\alpha = \sum_{x=0}^n [P(Y \geq I_{\text{match}} | x) C_n^x p_2^x (1 - p_2)^{n-x}] \quad (11)$$

The statistical score then becomes:

$$\beta = -\log(\alpha) = -\log \left\{ \sum_{x=0}^n [P(Y \geq I_{\text{match}} | x) C_n^x p_2^x (1 - p_2)^{n-x}] \right\} \quad (12)$$

MCS Solution to the Abundance-Based Score Model

To accurately calculate the score in eq 12, the computational time would scale exponentially with the input size, n . To overcome this limitation, we have adopted an approach based on Monte Carlo Simulation to estimate the statistical score. The MCS approach is a linear time algorithm after optimization and computationally feasible for analysis of shotgun proteomics data sets. The algorithm is summarized as follows:

1. For a given number x , randomly select that number of product ions from the experimental spectrum without duplicates;
2. Calculate the total abundance of the selected x ions and compare that total abundance with I_{match} ;
3. Repeat procedures 1 and 2 for k times;
4. Calculate $P(Y \geq I_{\text{match}} | x)$ by use of

$$P(Y \geq I_{\text{match}} | x) = \frac{\text{\#of times that the total abundance of picked } x \text{ ions} \geq I_{\text{match}}}{k} \quad (13)$$

5. Repeat procedures 1, 2, 3, and 4 for all $x = 0, 1, \dots, n$;
6. Calculate the score using eq 12.

The statistical score calculated by MCS is henceforth referred to as pp2_{MCS} to be distinguished from the previously described pp2 value calculated with the Central Limit Theorem approximation.¹³

Amino Acid Residue Tag-Based Score Model via Monte Carlo Simulation. Statistical Model Based on Amino Acid Residue Tags

Peptide fragmentation by collision induced dissociation (and other thermal activations) predominantly results in the formation of y and b ions. Consecutive y and b ions can be used to infer the partial peptide sequences and are referred to as sequence tags.^{7,8} An amino acid residue tag is defined as a pair of observed peaks for two consecutive b or y ions whose mass difference is equal to the mass of the respective amino acid residue in the peptide as shown in Figure 1a. The presence of amino acid residue and sequence tags can be used as a measure of the confidence for a peptide match. In the statistical model described here, amino acid residue tags are evaluated to score the confidence in a peptide identification. In principle, a random peptide match will have randomly matched product ions and should have a smaller number of residue tags than a true peptide match. Figure 1a shows an example of a verified true peptide match with matched product ions clustered in non-neutral loss b/y ion series that resulted in 8 residue tags. In contrast, Figure 1b shows a random peptide match with randomly matched product ions in its theoretical ion table and without any residue tags. The residue tag-based score model tests two hypotheses: the null hypothesis H_0 , which states that the match has matched ions randomly distributed across the theoretical ion table, and the alternative hypothesis H_A , which states that the match has a nonrandom pattern of matched ions in the theoretical ion table.

For a peptide match, N is the total number of theoretical product ions, and M defines the number of theoretical b/y non-neutral loss ions. For the peptide match, n theoretical ions (non-neutral loss ions and neutral loss ions) match the experimental spectrum. The variable w is defined as the number of b/y non-neutral loss ions that randomly match the experimental spectrum given that n theoretical ions match those in the experimental spectrum. Under H_0 , all matched theoretical product ions are independent random occurrences. Therefore, w follows a Hypergeometric(n, M, N) distribution and its probability mass function is

$$p(w) = \frac{C_M^w C_{N-M}^{n-w}}{C_N^n} \quad (14)$$

The variable t for the peptide match denotes the number of residue tags, that is, the number of pairs of consecutive b/y product ions in the theoretical ion table that match those in the experimental spectrum. The model evaluates whether the number of residue tags inferred from matched product ions in the theoretical ion table for the peptide match could be a random occurrence. The p value, α , is the probability that the number of residue tags, t , for a random peptide match is greater than or equal to that of the actual match, n_T , under H_0

$$\alpha = P(t \geq n_T) \quad (15)$$

The statistical score defined in the model is the negative logarithm of the p value (eq 9).

According to the Law of Total Probability, the p value can be calculated by the following equation

$$\alpha = P(t \geq n_T) = \sum_{w=0}^n [P(t \geq n_T | w) p(w)] \quad (16)$$

where $p(w)$ is the probability mass function for w as shown in eq 14, and $P(t \geq n_T|z)$ defines the probability that t for a random match is greater than or equal to that of the actual match, n_T , given that the number of matched b/y non-neutral loss ions in the theoretical ion table is w , under H_0 . Substitution of eq 14 into eq 16 yields

$$\alpha = \sum_{w=0}^n \left[P(t \geq n_T|w) \frac{C_M^w C_{N-M}^{n-w}}{C_N^n} \right] \quad (17)$$

The statistical score is

$$\beta = -\log(\alpha) = -\log \left\{ \sum_{w=0}^n \left[P(t \geq n_T|w) \frac{C_M^w C_{N-M}^{n-w}}{C_N^n} \right] \right\} \quad (18)$$

MCS Solution to the Model Based on Amino Acid Residue Tags

Again, this statistical model cannot be practically solved. Therefore, MCS must be applied to produce a linear time algorithm to estimate the score. The MCS algorithm is summarized as follows:

1. For a given number w , randomly select that number of product ions from the M b/y non-neutral loss ions in the theoretical ion table without duplicates;
2. Calculate the number of b and y residue tags, t , based on the selected b and y ions and compare it with n_T ;
3. Repeat procedures 1 and 2 for k times;
4. Calculate $P(t \geq n_T|z)$ by use of

$$P(t \geq n_T|w) = \frac{\text{\#of times that } t \geq n_T}{k} \quad (19)$$

5. Repeat procedures 1, 2, 3, and 4 for all $w = 0, 1, n$
6. Calculate the score with eq 18.

The statistical score calculated by amino acid residue tag-based model is referred to as pp_{tag} .

Estimation Precision and Computational Expense of MCS Score Models. Estimation Precision

The estimation precision of the scores obtained via MCS depends on the sampling size of each simulation step, k . Increasing the sampling size in MCS will improve the precision of the estimation. However, it will also increase the computational expense. The optimal sampling sizes for the two models were determined empirically with shotgun proteomic data sets that contained 3068 MS/MS spectra collected from a LCQ Deca XP mass spectrometer. Four-hundred thirteen MS/MS spectra were scored by the two MCS-based score models. The scores for each spectrum were recalculated 100 times by the models to determine the relative standard deviation ($RSD = \text{standard deviation}/\text{mean} \times 100\%$)

Figure 2 shows the average RSD of the two scores for the 413 spectra. It can be seen that for the abundance-based score, the precision was poor ($RSD > 6\%$) even at a very large sampling

size. A simple variance reduction technique was then applied in which the simulation was performed R times and the score was estimated by the mean of the estimates from the R simulations. The computational expense depended on the total sampling size of the simulation, which was R multiplied by k . It can be seen from Figure 2 that the variance reduction technique improved the estimation precision for both models at small and median total sampling sizes. The RSD for different values of R should converge as the total sampling size approaches infinity. However, the values did not converge for the pp_{tag} estimates with R equal to 1, 9, and 25 due to the variance in RSD calculation (Figure 2b).

The actual implementation of the two models in MassMatrix employed a sampling size of $k = 200$ for each step and repeated each simulation for $R = 25$ times for each spectrum, that is, the total sampling size is equal to 5000. Under these conditions, the RSD improved to 3.1% for the abundance-based MCS model and 1.8% for the amino acid residue tag-based MCS model. Therefore, the two MCS models gave scores with reasonable precisions and were of practical use in shotgun proteomics database searches.

Computational Expense

For the MCS models implemented in MassMatrix with the total sampling size of 5000, the computational expense for processing 1000 peptide matches was ~ 10 s for abundance-based scoring and ~ 4 s for amino acid residue tag-based scoring on a single AMD 3200+ (2.2 GHz) PC after optimization for speed. Therefore, it is impractical to use MCS-based models to filter and score millions or even billions of potential peptide matches from actual database searches of real data. To overcome this limitation of the MCS-based approach, we use the MCS-based models for post hoc analysis of the data sets. In this manner, potential peptide matches were prefiltered by the non-MCS-based empirical and statistical models described previously in the MassMatrix search engine.¹³ The non MCS-based score models take constant time to score a peptide match and are normally hundreds of thousands of times faster than MCS-based models. For this reason, MCS-based models were applied to potential peptide matches after prefiltering the matches by other models (Figure 3).

By only searching the peptides that survive scoring by other faster models, we can then use the MCS approach to further validate these matches in a computationally affordable manner. The disadvantage of applying MCS in this way is that some unique peptide matches that may be found by MCS but missed by other models will not be returned in the final results. With current computational hardware this is a necessary tradeoff. The computational expense for processing the test data set containing 3068 spectra was ~ 4 s for the abundance-based MCS model and ~ 2 s for the amino acid residue tag-based MCS model. Because all statistical scores from the MCS models and other statistical models in MassMatrix are the negative logarithm of the probability that the peptide match is random, they all share the same standard to determine if a peptide match is significant.

Evaluation of Score Models Based on Monte Carlo Simulation

The MCS-based score models were evaluated by searching two merged data sets from 23 runs of bovine histone peptide mixture from a LCQ Deca XP mass spectrometer and 5 runs from a LTQ-Orbitrap mass spectrometer. The results were compared with those from the other statistical models in MassMatrix as described previously,¹³ and other database search algorithms including Mascot, OMSSA and X!Tandem. The data sets were searched against a database with a bovine histone database and the reversed human database appended as decoy sequences. The decoy database was much larger than the bovine histone database and created ~ 1000 times as many theoretical peptides as the bovine histone database does. From the combined LCQ data set, 13 397 out of 62 117 spectra were scored with potential peptide matches. From the combined LTQ-Orbitrap data set, 1243 out of 4606 spectra were scored

with potential peptide matches. A summary of the search results is listed in Table 1, and the complete lists of peptide matches for the merged LCQ data set and the LTQ-Orbitrap data set are provided in Supplementary Tables 1 and 2 (see Supporting Information).

The MCS-based models were evaluated by use of receiver operating characteristic (ROC) analysis.^{8,19,24} Because false positive peptide matches from the bovine histone database were considered negligible due to the large decoy database, peptides from bovine histone were considered as TPs and those from decoy database were considered as FPs.²³ The results of ROC analysis for the four programs without any allowed PTMs are shown in Figure 4 and discussed in details below. Although histones carry many PTMs, the majority of identified histone peptide matches were unmodified tryptic sequences. ROC analysis of searches that allowed for PTMs (acetylation of N-terminus, acetylation of Lysine, mono-, di- and trimethylation of Lysine, mono- and dimethylation of Arginine, and oxidation of Methionine) was also performed. The ROC curves showed similar profiles to searches without PTMs.

Performance of the Abundance-Based MCS Model

The $pp2_{MCS}$ from the abundance-based MCS score model was based on the same statistical model as the $pp2$ score described previously.¹³ The only difference between the two scores is that the $pp2_{MCS}$ score was calculated via MCS and the $pp2$ score was calculated by use of the Central Limit Theorem. It can be seen from the ROC analysis that the $pp2_{MCS}$ score outperforms the $pp2$ score. This is because the $pp2$ value estimated by the Central Limit Theorem cannot evaluate the quality of matches with a small number of product ions and the $pp2$ scores are normally overestimated when the score is >16 due to assumptions used in the Central Limit Theorem.¹³ The MCS-based models use random sampling methods to estimate the scores without assuming any distribution for any variables involved. Since the distribution of abundance of peaks in LC-MS/MS spectra varies from one spectrum to another, MCS is a better solution for abundance-based score models in MS/MS spectrometric data.

The distributions of $pp2_{MCS}$ for TPs and FPs are shown in Figure 5a and d. At a mass accuracy of 0.8 Da for product ions, there was an overlap between the distributions of TPs and FPs for $pp2_{MCS}$. However, it can be seen that TPs tend to have larger $pp2_{MCS}$ scores than FPs.

Performance of the Amino Acid Residue Tag-Based MCS Model

The amino acid residue tag-based model is also a MCS-based statistical model. Moreover, it takes into account the amino acid residue tags inferred from the LC-MS/MS spectra. Surprisingly, it outperformed all other statistical models and served as the best standard to discriminate TPs from FPs for both LCQ and LTQ-Orbitrap data sets. Figure 6 shows the distributions of pp_{tag} scores for TPs and FPs. For FPs, pp_{tag} had a much narrower distribution than $pp2_{MCS}$ scores. Furthermore, the overlap between FPs and TPs for pp_{tag} distributions was smaller than that for $pp2_{MCS}$ distributions. This observation was consistent with the ROC analysis and indicated that pp_{tag} performed better than $pp2_{MCS}$ at an accuracy of 0.8 Da for product ions. For TPs, the distributions of pp_{tag} were split into three groups labeled “1”, “2”, and “3” in Figure 6. Peptide matches in groups 1 and 2 had very high scores and narrow score distributions. Those peptides had a good ladder of b/y ions that matched the experimental spectra and were well separated from false positive matches. Peptide matches in group 3 normally had moderate to poor ladders of b/y ions and overlapped with false positive matches.

Effect of Peptide Charge on Monte Carlo Simulation Based Score Models

It is shown in Table 1 that 19.4 and 27.0% of doubly and triply charged spectra from LCQ and LTQ-Orbitrap mass spectrometers were identified as TPs respectively. However, only 5.4 and 9.2% of singly charged spectra from the two mass spectrometers were identified as TPs. ROC curves for spectra with different charges showed that all statistical score models in MassMatrix

performed better for doubly and triply charged spectra than for singly charged spectra (Figure 4). Therefore, singly charged spectra were less favorable due to their poorer fragmentation when compared with doubly and triply charged ions. Because of this under-performance for singly charged species, the MCS-based approaches may not be suitable for the data sets where MS/MS of singly charged precursors predominate.

Effect of Charge on the Abundance-Based MCS Model

The charge state of LC-MS/MS spectra also affected their scores from the MCS-based score models. For the abundance-based MCS model, the $pp2_{MCS}$ score distributions for doubly and triply charged matches were narrower than those for singly charged ones and the score distribution of TPs was better separated from that of FPs for doubly and triply charged matches than for singly charged ones. This was consistent with the ROC analysis and indicated that the model performed better for doubly and triply charged spectra.

Effect of Charge on the Amino Acid Residue Tag-Based MCS Model

For the amino acid residue tag-based MCS score model, the score distributions for singly charged matches (Figure 6b and e) and those for doubly and triply charged ones (Figure 6c and f) were quite different.

Singly charged peptides had one set of *b/y* ions with +1 charge. Therefore, high quality matches had almost a complete set of *b/y* ions that matched the experimental data. Therefore, the probability density of group 1 peptide matches was much higher than that of group 2 peptide matches. Due to the fact that singly charged FPs had a broad distribution, group 3 TPs almost completely overlapped with the distribution of FPs, and there was also overlap between FPs and TPs in groups 1 and 2 (Figure 6b and e).

Doubly and triply charged peptides had at least two sets of *b/y* ions with +1 and +2 charges. Doubly and triply charged matches of high quality normally had only one out of more than two sets of *b/y* ions in the theoretical product ion table that matched the experimental data. Thus these spectra had pp_{tag} scores that fell in group 2 instead of group 1 and the probability density of group 2 was greater than that for group 1 for doubly and triply charged TPs. Both groups 1 and 2 TPs were well separated from the FPs for doubly and triply charged matches due to the fact that the FPs had a narrow distribution.

In general, the score distributions for TPs and FPs were better separated for doubly and triply charged matches than for singly charged ones. Therefore, the model performed much better for doubly and triply charged spectra than singly charged, which confirmed the ROC analysis results.

Effect of High Mass Accuracy Precursor Ions on Monte Carlo Simulation Based Score Models

As shown in Table 1, 9.2% of singly charged spectra and 27.0% of doubly and triply charged spectra from the LTQ-Orbitrap mass spectrometer were identified as TPs. For comparison, only 5.4% of singly charged spectra and 19.4% of doubly and triply charged spectra from the LCQ mass spectrometer were identified as TPs. The ratio of FPs to TPs for the LCQ data set was much higher than that for the LTQ-Orbitrap data set. ROC analysis confirmed that the MCS-based statistical score models in MassMatrix performed better for the LTQ-Orbitrap data set than for the LCQ data set (Figure 4a and d). This result is directly related to the higher mass accuracy and higher resolving power of the precursor spectra obtained on the LTQ-Orbitrap. The high mass accuracy measurement of the precursor's *m/z* lowers the false positive rate during a database search by eliminating many putative peptide matches that fall outside the stricter mass tolerances. Furthermore, the higher resolution of the precursor ion spectra can be used to unambiguously determine the precursor's charge. Thus, the need to search multiply charged

precursors as both +2 and +3 charges is also eliminated, which lowers the likelihood for false positive peptide matches.

Search Speed and ROC Comparison of MCS Models with Other Algorithms

The search speed of MassMatrix using the MCS-based score models was compared with that of other database search algorithms (Figure 7). All searches were performed on the same computer equipped with an Intel quad core CPU (2.4 GHz) running a Linux operating system. All searches were executed in serial mode on one CPU core (with no other applications running) for comparison of the wall clock time for each application. It can be seen in Figure 7 that MassMatrix had a similar search speed as Mascot and X!Tandem but was significantly faster than OMSSA when searching the LCQ and LTQ-Orbitrap data. MassMatrix required 74 min to complete the search of the LCQ data set containing 62 117 MS/MS spectra against a database of 97 114 protein sequences and <1 min to complete the search of the LTQ-Orbitrap data set containing 4606 MS/MS spectra. The search time for MassMatrix included the calculation of 3 other scores in addition to the 2 MCS-based scores.

Next we compared the ROC curves for the MCS-based scores in MassMatrix with those for the score in Mascot, the e-value in OMSSA and the expectation value in X!Tandem. The ROC curves for the search results from low mass accuracy MS/MS data obtained on a LCQ and high mass accuracy precursor MS/MS data obtained on a LTQ-Orbitrap are plotted in Figure 8 (in ROC curve analysis, values toward the top of the graph indicate higher sensitivity and values toward the left indicate higher specificity). The ROC curves show that the abundance-based MCS score model in MassMatrix, had poorer discriminatory power than that of Mascot and OMSSA but better discriminatory power than X!Tandem for searches on the low mass accuracy LCQ data. The situation was similar when comparing searches of the LTQ-Orbitrap data. However, the abundance-based MCS model obtained similar discriminatory power to that of OMSSA.

We also used ROC curves to compare the discriminatory power of the amino acid residue tag-based MCS model with the three other search algorithms. In this case, the amino acid residue tag-based MCS model in MassMatrix, outperformed Mascot, OMSSA and X!Tandem when searching the low mass accuracy MS/MS data from the LCQ. This result highlights the benefit of an amino acid residue tag-based approach for database searching of low-mass accuracy MS/MS data. The performance gap narrowed upon analysis of the ROC curves for the high mass accuracy precursor MS/MS data obtained on the LTQ-Orbitrap. In this case, the amino acid residue tag-based approach still outperformed X!Tandem and OMSSA while achieving similar performance to that of Mascot.

Conclusions

Two new statistical score models based on MCS have been developed to evaluate peptide matches in shotgun proteomic data and incorporated in a database search program, MassMatrix. The first model is abundance-based and evaluates the peptide matches based on the total abundance of the matched peaks in the experimental spectra. It outperforms the abundance-based model that employs the Central Limit Theorem. This performance gain is due to the fact that the MCS-based model does not involve any simplification and makes no additional assumptions in the statistical model to calculate the statistical scores as other methods do. The second MCS-based model accounts for residue tags inferred from the LC-MS/MS spectra. This model outperforms all the other statistical models and serves as the best standard in MassMatrix for shotgun proteomic data from LCQ and LTQ-Orbitrap mass spectrometers. The score model outperformed the open-source applications OMSSA and X!Tandem when searching both LCQ and LTQ-Orbitrap data and outperformed Mascot when searching LCQ

data. As the two models are based on different search criteria, they are complementary to each other and combined to offer high confidence in peptide identification.

Although models based on MCS are normally of high computational expense, potential peptide matches can be prefiltered by other statistical or even simple empirical models prior to MCS-based scoring. Using such strategies the MCS-based score models in MassMatrix achieved a similar search speed as Mascot and X!Tandem and a faster speed than OMSSA. Application of simple variance reduction techniques resulted in excellent estimation precisions for the MCS score models with RSDs of 3.1% and 1.8% for the abundance-based model and the amino acid residue tag-based model respectively. By incorporating the two models in a new database search program, MassMatrix, it was further demonstrated that MCS can be practically used to solve complex statistical models in database searching for shotgun proteomic data. ROC analysis indicated that MCS-based score models improved the performance of MassMatrix for the data sets from LCQ and LTQ-Orbitrap mass spectrometers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Mitchell Meade, Lanhao Yang, and Liwen Wang for providing the data sets. The study was funded by The Ohio State University, the National Institutes of Health (CA107106, RR023647, CA101956), the V Foundation (AACR Translational Cancer Research Grant), and the Leukemia & Lymphoma Society (SCOR).

References

1. Sadygov RG, Cociorva DC, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Methods* 2004;1(3):195–202. [PubMed: 15789030]
2. Hunt DF, Buko AM, Ballard JM, Shabanowitz J, Giordani AB. Sequence-analysis of polypeptides by collision activated dissociation on a triple quadrupole mass-spectrometer. *Biomed Mass Spectrom* 1981;53:397–408. [PubMed: 7306675]
3. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass-spectrometry. *Proc Natl Acad Sci USA* 1986;57:6233–6237. [PubMed: 3462691]
4. Biemann K. Contributions of mass-spectrometry to peptide and protein-structure. *Biomed Environ Mass Spectrom* 1988;16:99–111. [PubMed: 3072035]
5. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 2005;(13):3475–3490. [PubMed: 16047398]
6. Eng JK, McCormack A, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
7. Mann M, Wilm M. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;66:4390–4399. [PubMed: 7847635]
8. Tabb DL, Saraf A, Yates JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003;75:6415–6421. [PubMed: 14640709]
9. Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001;17:S13–S21. [PubMed: 11472988]
10. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* 2003;3:1454–1463. [PubMed: 12923771]
11. Keller A, Nesvizhskii A, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74:5383–5392. [PubMed: 12403597]

12. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
13. Xu H, Freitas AF. A high mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data. *BMC Bioinformatics* 2007;8:133. [PubMed: 17448237]
14. Zhang N, Aebersold R, Schwikowski B. ProbiD: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2002;2:1406–1412. [PubMed: 12422357]
15. MacCoss MJ, Wu CC, Yates JR. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem* 2002;74:5593–5599. [PubMed: 12433093]
16. Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem* 2003;75:435–444. [PubMed: 12585468]
17. Sadygov RG, Yates JR. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem* 2003;75:3792–3798. [PubMed: 14572045]
18. 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;76:1664–1671. [PubMed: 15018565]
19. 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;3:958–964. [PubMed: 15473683]
20. Sadygov R, Wohlschlegel J, Park SK, Xu T, Yates JR. Central limit theorem as an approximation for intensity-based scoring function. *Anal Chem* 2006;78:89–95. [PubMed: 16383314]
21. Sures I, Gallwitz D. Histone-specific acetyltransferases from calf thymus. isolation, properties, and substrate specificity of three different enzymes. *Biochemistry* 1980;19:943–951. [PubMed: 7356970]
22. Zhang LW, Freitas MA, Wickham J, Parthun MR, Klisovic MI, Marcucci G, Byrd JC. Differential expression of histone post-translational modifications in acute myeloid and chronic lymphocytic leukemia determined by high-pressure liquid chromatography and mass spectrometry. *J Am Soc Mass Spectrom* 2004;15:77–86. [PubMed: 14698558]
23. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4(3):207–214. [PubMed: 17327847]
24. Elias JE, Haas W, Faherty BK, P GS. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* 2005;2(9):647–648. [PubMed: 16118632]

Figure 2. Estimation precision of various total sampling sizes for (a) abundance-based MCS model and (b) amino acid residue tag-based MCS model.

Figure 3. Flow diagram of MassMatrix database search algorithm that employs MCS-based score models.

Figure 4. ROC curves of different score standards in MassMatrix for peptides with (a) all charges; (b) +1 charges; and (c) +2/+3 charges from LCQ mass spectrometer and peptides with (d) all charges; (e) +1 charges; and (f) +2/+3 charges from LTQ-Orbitrap mass spectrometer.

Figure 5. Distributions of $pp2_{MCS}$ score from the MCS-based score models for peptides with (a) all charges; (b) +1 charges; and (c) +2/+3 charges from LCQ mass spectrometer and peptides with (d) all charges; (e) +1 charges; and (f) +2/+3 charges from LTQ-Orbitrap mass spectrometer.

Figure 6. Distributions of pp_{tag} score from the MCS-based score models for peptides with (a) all charges; (b) +1 charges; and (c) +2/+3 charges from LCQ mass spectrometer and peptides with (d) all charges; (e) +1 charges; and (f) +2/+3 charges from LTQ-Orbitrap mass spectrometer.

Figure 7. Search speed in terms of MS/MS spectra per second for MassMatrix, Mascot, OMSSA, and X!Tandem on a PC with an Intel quad core CPU (2.4 GHz) and Linux operating system. The two merged data sets from LCQ and LTQ-Orbitrap mass spectrometers were searched against a database containing 97 114 protein sequences.

Figure 8. ROC curves of pp_{tag} and pp_{2MCS} from the MCS-based score models in MassMatrix along with other statistical scores from Mascot, OMSSA and X!Tandem for (a) the LCQ data set and (b) the LTQ-Orbitrap data set.

Table 1
Summary of the Search Results for the Merged Data Sets of 23 Runs of Bovine Histone Samples from a LCQ Deca XP Mass Spectrometer and 6 Runs from a LTQ-Orbitrap Mass Spectrometer^a

data	charge	spectra	spectra of TP's	spectra of FP's	percentage of spectra of TP's	percentage of spectra of FP's
LCQ data set	+1	36196	1969	1915	5.4%	5.3%
	+2, +3	25921	5023	4472	19.4%	17.3%
LTQ-Orbitrap data set	+1	1202	111	19	9.2%	1.6%
	+2, +3	3404	918	195	27.0%	5.7%

^aThe searches were performed against a database containing bovine histone database and a decoy reversed human database with 97 114 protein sequences.