Journal of proteome research

Article

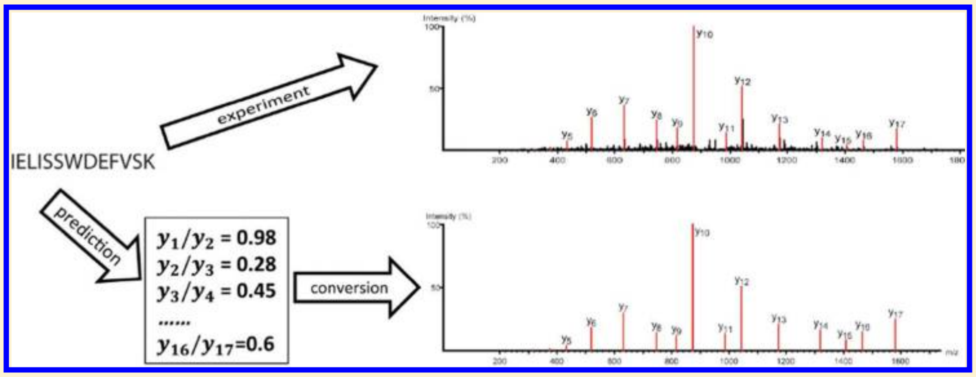
pubs.acs.org/jpr

MS-Simulator: Predicting *Y*-Ion Intensities for Peptides with Two Charges Based on the Intensity Ratio of Neighboring Ions

Shiwei Sun,[†,‖] Fuquan Yang,[‡] Qing Yang,[†,⊥] Hong Zhang,[#] Yaojun Wang,[†] Dongbo Bu,*[,†,‖] and Bin Ma*[,§]

[†]Advanced Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[‡]Proteomics Platform, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China
[§]School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
[‖]Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing, China
[⊥]School of Computer Science, University of Science and Technology, Beijing, China
[#]College of Food Science and Biological Engineering, Zhejiang Gongshang University, Hangzhou, China

S *Supporting Information*



ABSTRACT: For the identification of peptides with tandem mass spectrometry (MS/MS), many software tools rely on the comparison between an experimental spectrum and a theoretically predicted spectrum. Consequently, the accurate prediction of the theoretical spectrum from a peptide sequence can potentially improve the peptide identification performance and is an important problem for mass spectrometry based proteomics. In this study a new approach, called MS-Simulator, is presented for predicting the *y*-ion intensities in the spectrum of a given peptide. The new approach focuses on the accurate prediction of the relative intensity ratio between every two adjacent *y*-ions. The theoretical spectrum can then be derived from these ratios. The prediction of a ratio is a closed-form equation that involves up to five consecutive amino acids nearby the two *y*-ions and the two peptide termini. Compared with another existing spectrum prediction tool MassAnalyzer, the new approach not only simplifies the computation, but also improves the prediction accuracy.

KEYWORDS: *mass spectrometry, spectrum prediction, mobile proton hypothesis, peptide identification*

■ INTRODUCTION

Tandem mass spectrometry (MS/MS) has emerged as an indispensable method for the high-throughput protein identification and characterization in proteomics.[1,2] In a typical "bottom-up" protein identification experiment, the proteins are first enzymatically digested into shorter peptides, and then the peptides are measured with LC−MS/MS to produce a large amount of MS/MS spectra. Database search software is then employed to match each spectrum with the peptides in a protein sequence database and report the most probable peptide for the spectrum.

Many software tools are available for database search, including the earlier tools SEQUEST,[3] Mascot,[4] X!Tandem,[5] OMMSA,[6] SCOPE,[2] ProbID,[7] and the more recent pFind,[8] MS-GFDB[9] and PEAKS DB.[10] Each of these tools defines its own scoring function to measure the quality of the peptide-spectrum matches (PSMs) and uses the PSM score to separate the true and false identifications. The quality of a scoring function heavily affects the performance of a database search tool. Most of these tools, explicitly or inexplicitly, require the prediction of a theoretical spectrum from a given peptide sequence. For each PSM, the similarity between the input spectrum and the predicted spectrum from the peptide is one of the most important features for the scoring function.

Another approach for peptide identification with MS/MS is de novo sequencing. De novo sequencing can be used to both confirm the database search results and to derive sequences that

Received: March 9, 2012
Published: July 13, 2012

are not in a database.[10,11] Among the de novo sequencing tools developed, the more popular ones are PEAKS[12,13] and PepNovo.[14−16] In theory, de novo sequencing can be regarded as a special database search that searches in all the amino acid combinations (instead of any given protein database). However, because of this extensive search space, an efficient computer algorithm becomes an equally important factor as the scoring function.

The requirement for a high-quality PSM scoring function necessitates the accurate spectrum prediction. The reasonableness of spectrum prediction is essentially rooted in the reproducibility of tandem mass spectra.[16,17] In particular, Li et al.[17] reported that for the same peptide, the reproducibility of its tandem mass spectra is substantially high (Pearson correlation coefficient CC > 0.85) within an experiment. The Pearson CC drops to 0.70 across experiments, but is still sufficient to distinguish one peptide from another.

Although theoretically possible, the accurate spectrum prediction is a challenging problem due to the lack of quantitative understanding of the complex peptide fragmentation process in MS/MS. To circumvent the difficulty, most of available database searching tools employ certain oversimplified models. For example, the widely used SEQUEST software assigns artificially determined fixed intensities to different types of ions.[3] This oversimplification usually incurs a significant deviation between the experimental spectrum and the prediction[18] and may lead to compromised peptide identification performance.

A pioneer research in the computational prediction of a peptide's spectrum was the MassAnalyzer program.[18,19] On the basis of the mobile proton hypothesis, a kinetic model was proposed to simulate the peptide fragmentation process. The model includes most fragmentation pathways listed in literatures and additional pathways observed by the author. A total of 236 parameters were trained from the spectra of known peptides and used in the computer simulation algorithm for predicting the spectra of new peptides. The model was first developed for charge 2 spectra[18] and then extended to spectra with more than 2 charges.[19] Besides MassAnalyzer, there have been other works for studying the relationship between peak intensities and peptide sequences, without explicitly predicting the spectrum.[20−25] Different from the kinetic model, another program, PeptideART,[17] uses a machine learning approach to predict the probability that a specific ion is observed. There are also works that attempt to only predict the intensity ranks for fragment ions.[4,24,26]

This study aims to predict the peak intensities of the spectrum from a peptide's sequence. This shares the same goal with the earlier MassAnalyzer program. However, a new computational model, MS-Simulator, is proposed. Instead of predicting all the peak intensities simultaneously, the new model focuses on the accurate prediction of the relative intensity *ratio* between every two adjacent *y*-ion peaks. The overall spectrum is then derived from the predicted ratios. The parameter training process is also designed to optimize the accuracy of the predicted ratio. It turns out that this seemingly simple shift of focus drastically simplifies the parameter training process and the prediction. Intuitively, because the two *y*-ions y_i and y_{i+1} are adjacent, many "global" or "remote" factors (such as a distant amino acid) may have approximately equal effects on the intensities of the two *y*-ions and get canceled out in the ratio y_i/y_{i+1}. In fact, one does not even need to know the exact mechanisms of these factors. The only requirement is that they affect y_i and y_{i+1} approximately equally. As a result, our model is greatly simplified. This fact is revealed more clearly during the mathematical derivation in the Prediction Model section.

In terms of prediction accuracy for *y*-ion intensities, MS-Simulator compares favorably to MassAnalyzer on a few large spectrum data sets studied in the paper. However, unlike MassAnalyzer, MS-Simulator does not make an effort to predict the other ion types. To showcase the application of spectrum prediction in peptide identification, the predicted spectrum with MS-Simulator was used in a straightforward way to rescore the PSMs reported by SEQUEST. The experiments showed that the rescoring can significantly improve SEQUEST's peptide identification results.

■ MATERIALS AND METHODS

Peptide fragmentation under collision-induced dissociation (CID) is a complex process that involves multiple competing pathways, some of which are still not fully understood.[27,28] The MS-Simulator method proposed in this study is largely based on the widely accepted mobile proton model for peptide fragmentation,[29] which is very briefly reviewed in the following. A better review of the mobile proton model can be found in refs 22 and 30. The mobile proton model states that if the number of charges of the precursor ion is more than the number of basic side-chain groups, then at least one proton is not localized and can migrate along the peptide backbone to an amide nitrogen. The protonation of the amide nitrogen profoundly weaken the amide bond and makes the carbon atom of the protonated amide group a likely target attacked by nearby electron-rich groups, which lead to bond cleavage and the possible generation of *b* and *y* fragment ions.[22,30] These fragmentation pathways that involve protons at the cleavage sites are called charge-directed. In contrast, when the number of protons is not more than the number of arginine residues, the peptide fragmentation is dominated by charge-remote pathways. In this case, all protons are localized on the side-chains of basic residues and seldom migrate along backbone under the low energy CID conditions. The charge-remote and charge-directed fragmentations are the two main peptide fragmentation pathways. The actual fragmentation may be a combination of the two, and include some other minor pathways to produce fragment ions other than *b* and *y* series.[18,31]

Prediction Model

While the real peptide fragmentation process is complex and not fully understood,[27,28] to enable the computational prediction of the spectrum, one can use the basis of a simplified model that captures the main fragmentation pathways. The MS-Simulator model is based on a few reasonable assumptions that are derived from the mobile proton hypothesis. Assumptions similar to our assumptions 1 and 2 were previously made in ref 18.

A peptide sequence with length *n* is denoted by $A_1A_2\cdots A_n$, where A_i denotes the *i*-th amino acid starting from the peptide's N-term. For the sake of presentation clarity, we relabel the *y*-ions so that the y_i ion is the result of the fragmentation at the left of A_i (Figure 1). Note that this relabeling makes the *y*-ion indices to be in the reverse order of the convention.

Assumption 1. The intensity of y_i is proportional to the probability that the amino acid A_i is protonated, denoted by P_i. That is,

$$y_i = \alpha \cdot P_i$$

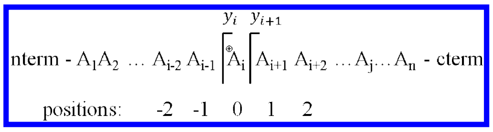

y_i y_{i+1}
nterm - A_1A_2 … A_{i-2} A_{i-1} $\oplus$ A_i A_{i+1} A_{i+2} … A_j…A_n - cterm
positions: -2 -1 0 1 2

Figure 1. The *y*-ions indices are labeled differently from convention so that y_i is caused by the fragmentation at left of amino acid A_i. The effects of A_j on both y_i and y_{i+1} become approximately equal when A_j is far away from A_i. Thus, the intensity ratio y_i/y_{i+1} is mostly determined by the five amino acids at positions −2 to 2.

Here $\alpha = 1/\sum_{j=1}^{n} y_j$ is a normalization factor that normalizes the intensities of all the *y*-ions. This assumption is a natural simplification of the mobile proton model.

Assumption 2. The distribution of protons across a peptide is calculated according to Boltzmann distribution. More specifically, a specific configuration regarding the locations of charging protons in a peptide ion is considered as a microstate. Denote by E_i the energy level of the microstate that residue *i* is protonated. Then the probability that residue *i* is protonated is calculated by

$$P_i = \exp(\beta \cdot E_i)$$

Here $\beta = -1/(R \cdot T_{\text{eff}})$ for R being the gas constant and T_{eff} being the effective temperature of the peptide. Note that T_{eff} correlates linearly to the normalized collision energy used by the mass spectrometer, and thus in our model β is regarded as a constant in a single experiment.

Putting the two assumptions together, one can derive that

$$y_i = \frac{1}{\sum_{j=1}^{n} y_j} \cdot \exp(\beta \cdot E_i)$$

This is a nonlinear function, and the calculation of y_i involves the calculation of all the *y*-ions because of the normalization factor $1/\sum_{j=1}^{n} y_j$. These make it difficult to design a model that can learn important parameters directly from known spectral data. However, the formula can be greatly simplified if one focuses on the log ratio between two adjacent *y*-ions:

$$\log\left(\frac{y_i}{y_{i+1}}\right) = \beta \cdot (E_i - E_{i+1}) \tag{1}$$

Notice that the normalization factor $1/\sum_{j=1}^{n} y_j$ is canceled out and β is a constant across the whole experiment. Thus, the remaining task for predicting the intensity ratio is the determination of the microstate energy difference $E_i - E_{i+1}$.

Assumption 3. The energy E_i is additively affected by other amino acids and the two termini of the peptide. More formally,

$$E_i = \sum_{j=1}^{n} E(A_j, j - i) + E(\text{N-term}, 1 - i) + E(\text{C-term}, n - i)$$

Here $E(x,d)$ is a function of x and d, indicating the additive effect of an amino acid or terminus x to E_i at distance d. We use $d < 0$ to indicate that x is to the left of A_i.

Since we only concern about $E_i - E_{i+1}$, let $\Delta(x,d) = E(x,d) - E(x,d - 1)$ be the difference between the effects at distances d and d −1, respectively. Then

$$E_i - E_{i+1} = \sum_{j=1}^{n} \Delta(A_j, j - i) + \Delta(\text{N-term}, 1 - i) + \Delta(\text{C-term}, n - i) \tag{2}$$

By combining eq 1 and eq 2, we get

$$\log\left(\frac{y_i}{y_{i+1}}\right) = \beta \cdot \left[\sum_{j=1}^{n} \Delta(A_j, j - i) + \Delta(\text{N-term}, 1 - i) + \Delta(\text{C-term}, n - i)\right] \tag{3}$$

If the values of $\Delta(x,d)$ can be determined, then the *y*-ion ratio can be computed easily from eq 3. If one fixes the amino acid (or terminus) x and varies the distance d, it is natural to reason that when the absolute value of d is large, the effect $E(x,d)$ becomes stabilized, and increasing the distance d by 1 will not fluctuate the value $E(x,d)$ by a large amount. Thus, the following assumption is reasonable.

Assumption 4. The difference between $E(x,d)$ and $E(x,d - 1)$ fades away when the distance d gets large. In other words, $\Delta(x,d) = E(x,d) - E(x,d - 1) \approx 0$ for sufficiently large d.

In our model, we only use the following $\Delta(x,d)$ and found them sufficient for the model's accuracy:

(1) $-4 \le d \le -1$ for x being the N-term.
(2) $1 \le d \le 4$ for x being the C-term and the C-term residue is not Lys or Arg.
(3) $1 \le d \le 7$ for x being a C-term Lys or Arg.
(4) $-2 \le d \le 2$ for all amino acids x that are not a C-term Lys or Arg.

In total, there are 122 values of $\Delta(x,d)$ that need to be determined, and most of them are from the 5 consecutive amino acids nearby the concerned y_i and y_{i+1} ions. These coefficients are optimized by the following parameter training process through spectral data of known peptides. Since $\beta = 1/(R \cdot T_{\text{eff}})$ is a constant in one experiment, eq 3 provides a linear equation for every two adjacent *y*-ions observed in the training data. By using a large number of training MS/MS spectra, thousands of linear equations are obtained on the 122 variables. The linear system is solved efficiently with the least-squares method to minimize the overall prediction error in the log intensity ratios. This provides 122 optimized coefficients. In our experiments, we actually trained $\beta \cdot \Delta(x,d)$ because β is an unknown constant factor for all the coefficients.

To predict the *y*-ion intensities of a new peptide sequence, these trained coefficients are used in eq 3 to compute the intensity ratios between every two adjacent *y*-ions. Then the *y*-ion intensities are derived from the ratios. If different collision energy is used in another experiment, the effective temperature T_{eff} will also change, resulting in different peptide fragmentations. Ideally, the coefficients should be retrained for each collision energy (and each instrument type) for best accuracy. However, for relatively small change of collision energy, we propose the following adjustment to reuse the coefficients. Note that the main difference caused by the change of T_{eff} in our model is the constant factor $\beta = -1/(R \cdot T_{\text{eff}})$ that is multiplied to each $\Delta(x,d)$. Therefore, to adjust for relatively small change of collision energy, all the trained coefficients $\Delta(x,d)$ are multiplied by a single factor ρ before being used in the prediction under a different experimental condition. The factor ρ is determined by minimizing the prediction error on a few high quality spectra of which the peptides are easily identifiable in the new experiment.

Data Sets

Four data sets were used in the study. The first data set SwedCAD[32] came from human cell lines, human milk and *E. coli*. The MS/MS experiment was performed on a 7T LTQ FT instrument. The 15897 highly confident unique tryptic peptides that were published together with the data set were used in this study. All these peptides are unmodified and doubly charged. There are 139 peptides of which the C-termini are not Lys or Arg, possibly because they are the protein C-terminal peptides. 3959 peptides have a missed cleavage, and 453 have two or more missed cleavages. These peptide-spectrum matches (PSMs) were split into two parts, for testing and training purposes, respectively. The testing part consists of 1000 PSMs, and the training part consists of 14899 PSMs. Because these are unique peptides, the division ensured that the two parts did not share any common peptide. These two parts are referred to by SwedCAD-testing and SwedCAD-training, respectively.

The second data set, referred to as ISB Mix1, was extracted from ISB Standard Protein Mix Database,[33] which was made from peptides in a mixture of tryptic digests of 18 proteins using 8 different mass spectrometers. Only the LTQ-FT analysis of mix1 was used in this study. The peptides were identified using SEQUEST against *Haemophilus influenzae* database with the 18 proteins of interest and some common contaminating proteins. A decoy database was searched together in order to control the false discovery rate (FDR). The PSMs (both target and decoy) were downloaded together with the spectral data. By requiring a strict 0.1% FDR, 17109 PSMs were selected for our experiment.

The third data set, referred to as IBP DROSOPHILA, was created from Drosophila proteome collected using nano LC coupled with LTQ (Thermo) instrument at the Institute of Biophysics, Chinese Academy of Sciences (isolation width: 2 Da, activation time: 30 ms, and normalized collision energy: 35%). Note that this data set is of low mass accuracy since it is a Trap instrument. The correct PSMs were identified using a local copy of SEQUEST to search against the Drosophila database, appended with the reverse decoy database. The same criteria as the ISB Mix1 data set was used to filter the PSMs, and 5385 were selected for our experiment.

The fourth data set, referred to as PAe000350, was downloaded from the public data repository PeptideAtlas (ftp://ftp.peptideatlas.org/pub/PeptideAtlas/Repository/PAe000350). The data was generated from *Leptospira interrogans* on an LTQ-FT instrument. The target and decoy PSMs were downloaded together with the spectral data. The same criterion as the ISB Mix1 data set was used to filter the PSMs, and 11257 PSMs were selected for our experiment.

Among these data sets, the SwedCAD-training was used for the training of the coefficients, and the others were used as testing data.

■ RESULTS

Conservation of Ion Intensity Ratio within 5-mers Segments

In addition to the theoretical justification outlined in the Prediction Model section, the correctness of our model can be illustrated by studying the conservation of y-ion intensity ratios within a 5-mer (five consecutive amino acids). More specifically, if the model is correct, then the ratio y_i/y_{i+1} should be roughly determined by the 5-mer $A_{i-2}A_{i-1}A_iA_{i+1}A_{i+2}$ nearby the two y-ions, as well as the position of the 5-mer relative to the N- and C-termini (Figure 1).

To validate this, an experiment was carried out to study the correlation of the intensity ratios for different occurrences of the same 5-mer in different peptides and different experiments. First, all 5-mers occurring more than once in different peptides were identified. The two occurrences must be at the same position relative to the N-term and C-term in both peptides. To remove the potential bias rooted in experiment setting, we further required the two occurrences come from different data sets, one from SwedCAD and the other from ISB Mix1. For every pair of occurrences of the same 5-mer, the two y_i/y_{i+1} ratios were calculated and plotted in Figure 2b. The figure suggests that the scatter points roughly form a straight-line (Pearson CC = 0.92). This highly suggests that the 5-mers, although coming from different experiments, have considerably stable peak intensity ratios.

To justify the choice of the length 5, the same experiment was repeated on k-mers for $k = 3$ and 7. The Pearson CC for 3-mers is only 0.74 (Figure 2a), suggesting that 3-mer is not sufficient for the accurate determination of the y-ion ratios. Whereas, for $k = 7$, the Pearson CC is 0.93 (Figure 2c), suggesting that making it longer than 5 is not much more useful for further improvement of the prediction accuracy.

Figure S1 in the Supporting Information shows the distribution of the y_i and y_{i+1} absolute intensities involved in the 3, 5, and 7-mer experiments shown above. The distributions are similar across the three experiments, suggesting that the correlation difference in the three experiments was not due to the possible bias in the ion selection of the three experiments.

Trained Parameters

The parameters trained from the SwedCAD-training data set (see the Data Sets section) are given in Tables 1 and 2.

When examining the $\beta{\cdot}\Delta(x,0)$ coefficients in Table 1, it is clear that proline and glycine have relatively large values, indicating a preferable cleavage at N-terminal side over the C-terminal side of these two amino acids; in contrast, histidine, lysine, and arginine have negative values, indicating a preferable cleavage at C-terminal side over the N-terminal side. This is consistent with what was previously reported in the literature.[18] Another trend is that $\Delta(x,0)$ has relatively larger values than the other columns, suggesting that the amino acid between y_i and y_{i+1} has the most contribution toward the ratio y_i/y_{i+1}. From Table 2 one can also observe that in general the coefficients $\beta{\cdot}\Delta(x,d)$ approach 0 when d grows. This is consistent with our assumption 4.

Because the training data set contains peptides with missed cleavages, there might be more than one Arg or Lys residues in one peptide. So, these coefficients should be regarded as the combined effect of both the charge-directed and the charge-remote pathways for producing the y-ions. In fact, some other unknown pathways might have also played a role in the coefficients. As such, the trained coefficients are meant to make the final prediction more accurate on an average peptide in a bottom-up proteomics experiment and should not be used as the sole evidence to draw conclusions about an amino acid's chemical property.

Prediction Accuracy on Relative Intensity Ratios

Since the method predicts the intensity ratio first before deriving the y-ion intensities, we first examined the accuracy of the predicted log ratios $\log(y_i/y_{i+1})$. For each observed pair of

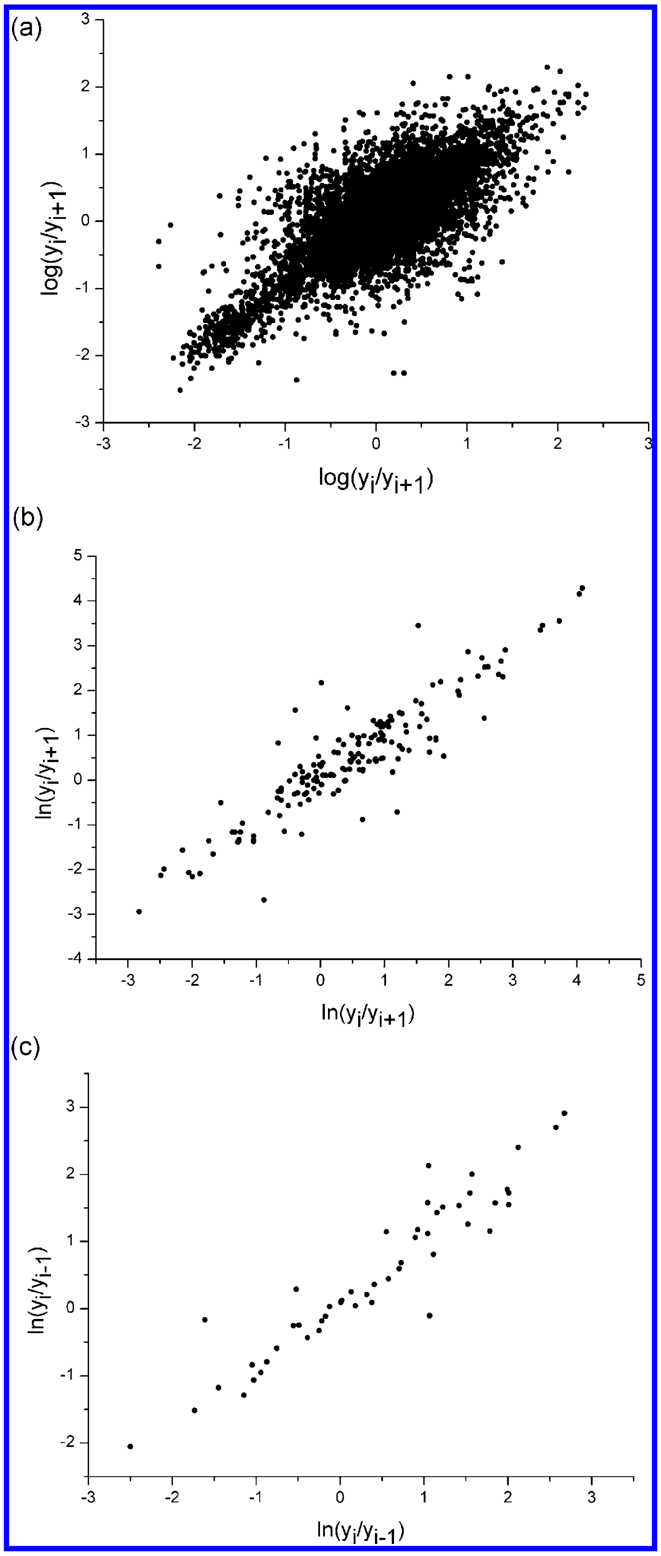


Figure 2. Conservation of logarithm of peak intensity ratio $\log(y_i/y_{i+1})$ within *k*-mer segments. Each data point represents two occurrences of the same *k*-mer in two different peptides and from two different data sets. The two $\log(y_i/y_{i+1})$ values are plotted as *x*- and *y*-axes, respectively. (a) 3-mer, CC = 0.74. (b) 5-mer, CC = 0.92. (c) 7-mer, CC = 0.93.

Table 1. Optimized Coefficients $\beta \cdot \Delta(x,d)$ for the Five Amino Acids Nearby the Concerned y_i and y_{i+1} Ions

residue	$\Delta(x,-2)$	$\Delta(x,-1)$	$\Delta(x,0)$	$\Delta(x,1)$	$\Delta(x,2)$
Ala	−0.37	−0.52	0.21	0.12	0.73
Cys	0.22	0.38	−0.59	0.00	−0.06
Asp	−0.76	−0.57	0.60	0.37	0.11
Glu	−0.63	0.26	−0.30	0.30	0.04
Phe	0.24	0.00	−0.19	−0.09	−0.09
Gly	−1.00	−1.69	1.90	0.00	−0.01
His	0.91	0.64	−1.70	−0.68	−0.35
Ile	0.42	0.58	−0.96	0.04	−0.02
Lys[a]	0.22	0.50	−1.76	−0.35	−0.54
Leu	0.24	0.11	0.06	−0.68	−0.01
Met	0.02	0.25	−0.39	0.13	−0.16
Asn	−0.21	−0.40	0.30	0.14	0.10
Pro	0.25	−0.44	1.95	−1.39	−0.01
Gln	0.49	0.68	−1.03	0.17	−0.11
Arg[a]	−0.39	0.19	−1.01	−0.43	−1.14
Ser	−0.35	−0.75	0.93	0.01	0.06
Thr	0.15	−0.30	0.20	0.07	−0.01
Val	0.27	0.53	−0.83	−0.01	0.00
Trp	0.55	0.45	−0.37	−0.38	−0.33
Tyr	0.32	0.09	−0.22	−0.05	−0.08

[a] These Lys and Arg coefficients are used only if they are not at the C-terminus of the peptide.

Table 2. Optimized Coefficients $\beta \cdot \Delta(x,d)$ for N-Term and C-Term with Lys, Arg, and Other Amino Acids, Respectively[a]

d	1	2	3	4	5	6	7
C-term (Arg)	2.95	2.15	1.48	0.97	0.56	0.24	−0.10
C-term (Lys)	2.03	1.31	0.89	0.52	−0.09	−0.29	−0.05
C-term (other)	0.27	0.37	0.16	0.03			
N-term	0.85	0.52	0.23	0.01			

[a] Note that *d* is negated for the N-term coefficients.

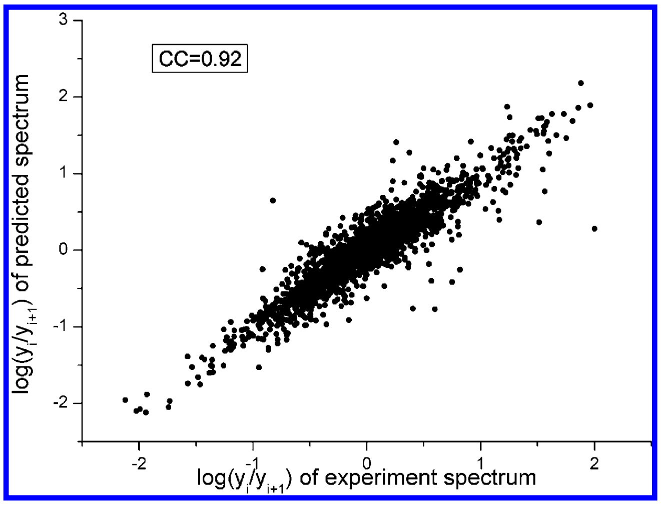


Figure 3. Prediction accuracy of $\log(y_i/y_{i+1})$ on data set SwedCAD-testing. The *x*-axis is the observed log intensity ratio and the *y*-axis is the predicted. The Pearson CC of the two values is 0.92.

ions y_i and y_{i+1} in the SwedCAD-testing data, Figure 3 plots the predicted log intensity ratio (*y*-axis) against the observed log intensity ratio (*x*-axis). The figure shows a strong correlation between the real and predicted ratios. The Pearson CC between the two is 0.92.

The Pearson CC between the real and predicted ratios for all four data sets are summarized in Table 3.

Table 3. Pearson CC between Predicted and Observed $\log(y_i/y_{i+1})$

data set	SwedCAD-testing	ISB Mix1	IBP DROSOPHILA	PAe000350
Pearson CC	0.92	0.85	0.73	0.84

Prediction Accuracy on *Y*-Ion Intensities

For each PSM in the SwedCAD-testing data set, the Pearson CC of the predicted and real *y*-ion intensities was calculated. Both MS-Simulator and MassAnalyzer were used in this experiment, and the Pearson CC distributions are shown in Figure 4. For MassAnalyzer, only the *y*-ions in the real and the predicted spectra were used in the calculation of the Pearson CC. Between the two tools, it is clear that MS-Simulator's prediction has an overall higher correlation with the real spectrum. MS-Simulator achieved an average Pearson CC of 0.92 on all the PSMs. In comparison, the MassAnalyzer tool was only able to get an average Pearson CC of 0.79.

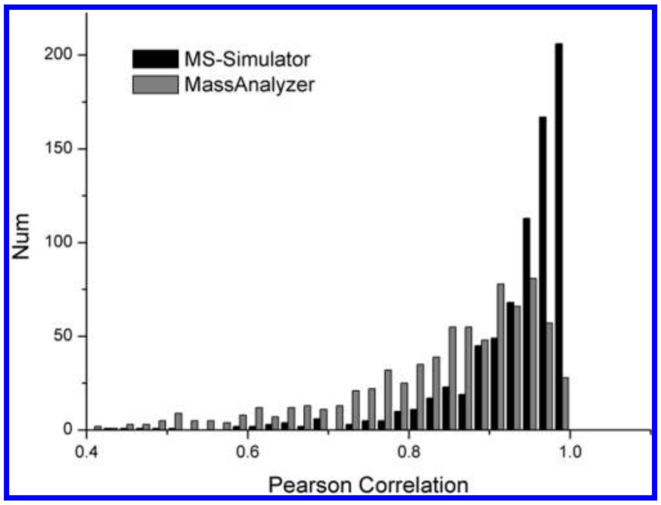


Figure 4. Prediction performance of MS-Simulator and MassAnalyzer on SwedCAD-testing. The *x*-axis is the Pearson CC between predicted and real *y*-ion intensities, and the *y*-axis is the number of PSMs.

Table 4 lists the average Pearson CC of the two tools on all of the four data sets studied. For MS-Simulator, two experiments were carried out. One used the same coefficients trained from the training data. The other adjusted the coefficients by multiplying a single factor ρ to all coefficients and optimized ρ according to several high quality spectra in the corresponding data set. This adjustment was to consider the possible different T_{eff} in different experiments.

Table 4. Average Pearson CC between the Real and Predicted Spectra of MassAnalyzer and MS-Simulator on the Four Testing Data Sets[a]

	MassAnalyzer	MS-Simulator	MS-Simulator (T_{eff} adjusted)
SwedCAD-testing	0.793	0.920	
ISB Mix1	0.745	0.800	0.818
IBP_DROSOPHILA	0.644	0.715	0.726
PAe000350	0.795	0.868	0.889

[a] The last column is MS-Simulator's performance after multiplying all the coefficients by a single factor ρ to reflect the possible variation of T_{eff} in experiments other than SwedCAD.

The table clearly shows that MS-Simulator achieved better prediction accuracy across all four data sets studied. The adjustment for the possible different T_{eff} also slightly improved the prediction accuracy. Note that the training of the parameters was conducted on the SwedCAD-training data set, which came from the same experiment as the SwedCAD-testing. Although the two training and testing data sets do not share any common peptides, the same instrument and experimental condition may explain why MS-Simulator's performance is better on SwedCAD-testing than on the other three testing data sets. Also because of this, the adjustment for T_{eff} was not necessary for SwedCAD-testing.

Improving SEQUEST Database Search with Predicted Spectrum

To investigate whether the predicted *y*-ion intensities help peptide identification, the predicted *y*-ion intensities were used in a straightforward way to rerank SEQUEST's database search results. For each PSM identified, SEQUEST provides two scores X_{corr} and ΔCn. According to the PEAKS DB paper,[10] the linear combination of the two scores $X_{\text{corr}} + 5\cdot\Delta Cn$ provides the optimal identification performance for SEQUEST. We further supplemented this score by adding 5 times the Pearson CC between the spectrum and the predicted spectrum from the peptide sequence. In total, five scoring strategies were used to rank the PSMs reported by SEQUEST: (1) X_{corr}, (2) $X_{\text{corr}} + 5\cdot\Delta Cn$, (3) $X_{\text{corr}} + 5\cdot\Delta Cn + 5\cdot\text{CC}_{\text{MassAnalyzer}}$, (4) $X_{\text{corr}} + 5\cdot\Delta Cn + 5\cdot\text{CC}_{\text{MassAnalyzerY}}$, and (5) $X_{\text{corr}} + 5\cdot\Delta Cn + 5\cdot\text{CC}_{\text{MS-Simulator}}$. Here, $\text{CC}_{\text{MassAnalyzer}}$ denotes the Pearson CC between the experimental spectrum and the predicted spectrum by MassAnalyzer,[19] and $\text{CC}_{\text{MassAnalyzerY}}$ is the Pearson CC restricted only on the *y*-ions of the experimental and predicted spectra. $\text{CC}_{\text{MS-Simulator}}$ is the Pearson CC between the *y*-ions of the experimental spectrum and the predicted spectrum by MS-Simulator.

These five strategies were compared on each of the three data sets: ISB Mix1, PAe000350, and IBP DROSOPHILA. The performance of each strategy was measured by the FDR curve, as shown in Figure 5. The figure demonstrates the following: (1) Adding the Pearson CC could clearly improve SEQUEST's peptide identification, regardless of the choice of the prediction programs. At the same FDR, more target PSMs were reported. (2) For MassAnalyzer, including more ion types in the Pearson CC calculation did not result in further improvement than using *y*-ion alone. This may be because MassAnalyzer's prediction of the other ion types is less accurate than the *y*-ions. By using the SwedCAD-testing data set, Figure S3 (Supporting Information) shows the distribution of the Pearson CC of MassAnalyzer's prediction of *y*-ions, *b*-ions, and the whole spectrum, respectively. The figure clearly indicates that *b*-ions' prediction is inferior to the *y*-ions, and the whole spectrum's prediction has the lowest Pearson CC. (3) With a better prediction accuracy, MS-Simulator provided the greatest improvement over SEQUEST's results. (4) The improvement over SEQUEST is more significant when the fragment ion mass accuracy is low (IBP DROSOPHILA data set).

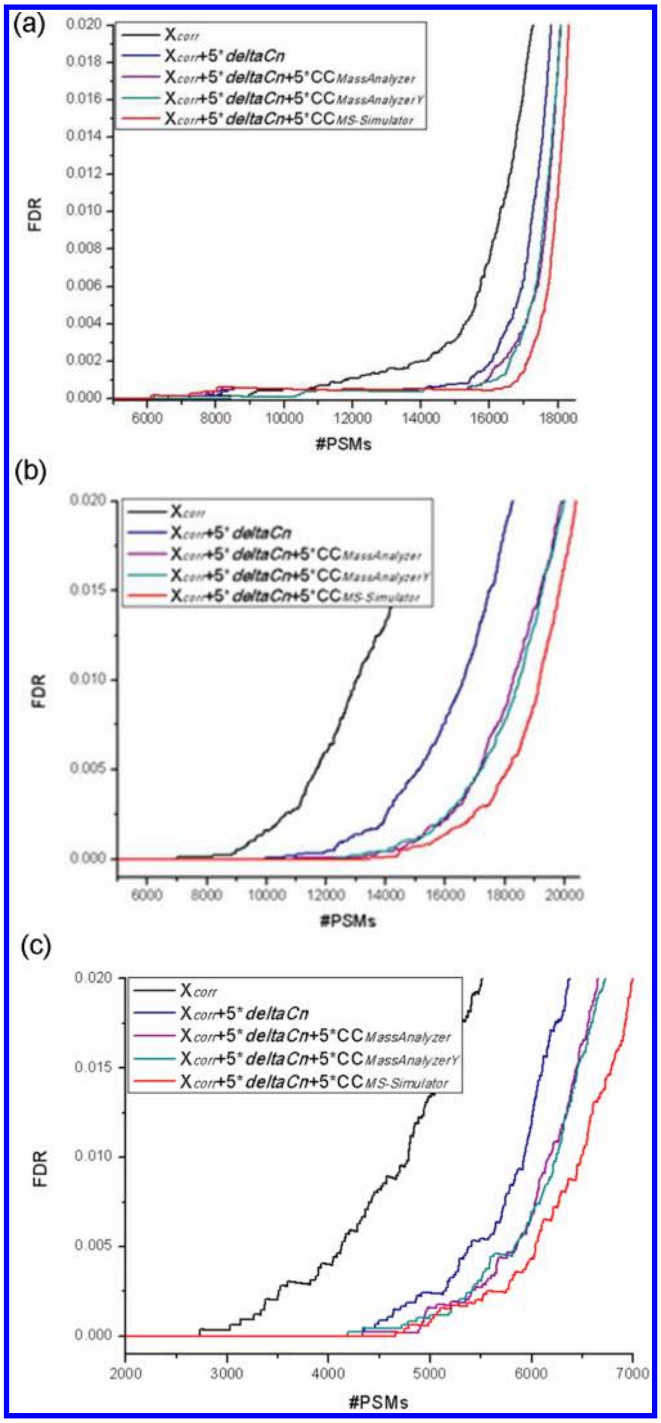


Figure 5. FDR curves for using the five scoring strategies on each of the three data sets: (a) ISB Mix1, (b) IBP_DROSOPHILA, (c) PAe000350.

■ DISCUSSION

We have presented a new computational model, MS-Simulator, for predicting the *y*-ion intensities of the CID spectrum of a peptide. The model focuses on the accurate prediction of the relative intensity ratio between every two adjacent *y*-ions and then derives the *y*-ion intensities from the ratios. The intuition is that many factors (such as a distant amino acid) remotely affecting the two *y*-ions may apply to the two *y*-ions approximately equally. By focusing on ratios, these remote factors may be canceled out. The ratio is therefore largely predictable by the 5-mer sequence around the two *y*-ions.

The outcome of the model is a much simplified training and predicting process. The optimal coefficients are learnable by solving a large linear system with the least-squares method. The prediction is also as simple as a closed form equation (eq 3) and does not require any intensive computing.

Despite its simplicity, the prediction accuracy is improved over an existing spectrum prediction program, MassAnalyzer. Since MassAnalyzer uses many more fragmentation pathways than MS-Simulator, the improvement on the prediction accuracy might have been due to better choices of parameters and more accurate parameter training. As mentioned above, our parameter training is simply solving a linear system. This guarantees the finding of the optimal set of parameters, whereas in MassAnalyzer's model, the system is highly nonlinear and no efficient algorithm can guarantee the finding of the optimal parameters.

The simplicity also makes it possible to embed the spectrum prediction in another peptide identification software tool, whether for database search or de novo sequencing. The hypothetical spectrum of each peptide candidate can be predicted *on-the-fly* and compared with the experimental spectrum. The ability to accurately predict the peak intensities may provide additional information for the identification software to utilize and lead to more accurate and sensitive identification. The simple experiment for reranking SEQUEST results in this paper already demonstrated such possibility. We suspect that a more sophisticated use of the predicted spectrum and the embedding of spectrum prediction in the software's searching process can further improve the peptide identification performance.

While the paper demonstrates the power of our prediction model, the MS-Simulator software still has a few limitations that will be improved in our future work. Currently the model has only been examined on the doubly charged tryptic peptides with no PTM. Moreover, it only predicts the *y*-ions. Future works include the extension of the model to study peptides with three or more charges, to predict other ion types such as the *b*-ions, and to predict for peptides with PTMs.

MS-Simulator can be accessed from http://bioinfo.ict.ac.cn/MS-Simulator/.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1: The intensity distribution of all ions from the SwedCAD-testing data set that are involved in the 3, 5, and 7-mer reproducibility experiment. Figure S2: Comparison between *y*-ion intensities predicted by MS-Simulator and that of experimental spectrum. The experiment spectrum (top) is shown head-to-tail with the predicted *y*-ions (bottom) for peptides. Figure S3: The Pearson CC distribution of MassAnalyzer's prediction of *y*-ions, *b*-ions and whole spectrum on the SwedCAD-testing dataset. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*(D.B.) Phone: 86-10-62601019. Fax: 86-10-62601356. E-mail: dbu@ict.ac.cn. (B.M.) Phone: 1-519-8884567 ext. 32747. Fax: 1-519-8851208. E-mail: binma@uwaterloo.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was funded by the National Basic Research Program of China (973 Program 2012CB316502, 2010CB833703), the National Natural Science Foundation of China (30800189, 30800168, 90914047), the Beijing Municipal Natural Science Foundation (5102029), and the "Strategic Priority Research Program" of the Chinese Academy of Sciences, Grant No. XDA06030200. Bin Ma was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery program. We thank Dr. Zhongqi Zhang for providing the MassAnalyzer program.

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Bafna, V.; Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **2001**, *17* (Suppl 1), S13−21.

(3) Yates, J. R., 3rd; et al. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **1995**, *67* (8), 1426−36.

(4) Perkins, D. N.; et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−67.

(5) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−7.

(6) Geer, L. Y.; et al. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958−64.

(7) 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* (10), 1406−12.

(8) Sun, R. X.; et al. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. *J. Proteome Res.* **2010**, *9* (12), 6354−67.

(9) Kim, S.; et al. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **2010**, *9* (12), 2840−52.

(10) Zhang, J.; et al. PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **2011**, DOI: 11: M111.010587.

(11) Ma, B.; Johnson, R. De novo sequencing and homology searching. *Mol. Cell. Proteomics* **2012**, *11* (2), O111014902.

(12) Ma, B.; et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−42.

(13) Ma, B.; Zhang, K. Z.; Liang, C. Z. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *J. Comput. Syst. Sci.* **2005**, *70* (3), 418−30.

(14) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−73.

(15) Frank, A. M.; et al. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.* **2007**, *6* (1), 114−23.

(16) Elias, J. E.; et al. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *22* (2), 214−9.

(17) Li, S.; et al. On the accuracy and limits of peptide fragmentation spectrum prediction. *Anal. Chem.* **2011**, *83* (3), 790−6.

(18) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908−22.

(19) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77* (19), 6364−73.

(20) Barton, S. J.; et al. Using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. *Anal. Chem.* **2007**, *79* (15), 5601−7.

(21) Lin, Y.; et al. A fragmentation event model for peptide identification by mass spectrometry. *Res. Comput. Mol. Biol. Proc.* **2008**, *4955*, 154−66.

(22) Paizs, B.; Suhai, S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun. Mass Spectrom.* **2002**, *16* (17), 1699−702.

(23) Schutz, F.; et al. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans.* **2003**, *31* (Pt 6), 1479−83.

(24) Frank, A. M. Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **2009**, *8* (5), 2226−40.

(25) Sun, S.; et al. Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *J. Proteome Res.* **2008**, *7* (1), 202−8.

(26) Bern, M.; Cai, Y.; Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **2007**, *79* (4), 1393−400.

(27) Mouls, L.; et al. Low energy peptide fragmentations in an ESI-Q-Tof type mass spectrometer. *J. Proteome Res.* **2007**, *6* (4), 1378−91.

(28) Neta, P.; et al. Dehydration versus deamination of N-terminal glutamine in collision-induced dissociation of protonated peptides. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (1), 27−36.

(29) Wysocki, V. H.; et al. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399−406.

(30) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508−48.

(31) Bythell, B. J.; Suhai, S.; Somogyi, A.; Paizs, B. Proton-driven amide bond-cleavage pathways of gas-phase peptide ions lacking mobile protons. *J. Am. Chem. Soc.* **2009**, *131* (39), 14057−65.

(32) Falth, M.; et al. SwedCAD, a database of annotated high-mass accuracy MS/MS spectra of tryptic peptides. *J. Proteome Res.* **2007**, *6* (10), 4063−7.

(33) Klimek, J.; et al. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7* (1), 96−103.