Algorithms for Automatic Interpretation of High Resolution Mass Spectra

Parminder Kaur,[†*] and Peter B. O'Connor[†*]

Department of Biochemistry, Mass Spectrometry Resource, Boston University School of Medicine, Boston, Massachusetts, USA

Automated interpretation of high-resolution mass spectra in a reliable and efficient manner represents a highly challenging computational problem. This work aims at developing methods for reducing a high-resolution mass spectrum into its monoisotopic peak list, and automatically assigning observed masses to known fragment ion masses if the protein sequence is available. The methods are compiled into a suite of data reduction algorithms which is called MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction). MasSPIKE includes modules for modeling noise across the spectrum, isotopic cluster identification, charge state determination, separation of overlapping isotopic distributions, picking isotopic peaks, aligning experimental and theoretical isotopic distributions for estimating a monoisotopic peak's location, generating the monoisotopic mass list, and assigning the observed monoisotopic masses to possible protein fragments. The method is tested against a complex top-down spectrum of bovine carbonic anhydrase. Results of each of the individual modules are compared with previously published work. (J Am Soc Mass Spectrom 2006, 17, 459–468) © 2006 American Society for Mass Spectrometry

The wide employment of Fourier Transform Mass Spectrometry (FTMS) [1, 2] instruments for Matrix Assisted Laser Desorption/Ionization (MALDI) [3] and Electrospray Ionization (ESI) [4] experiments results in thousands of high-resolution mass spectra every day, creating an information overload. Due to the high mass accuracy and resolving power of an FTMS, MALDI-FTMS [5–7] and ESI-FTMS [8, 9] are becoming the instruments of choice for proteomics [10], and experiments on proteins and large fragments of proteins, so called "top-down" [11–14] mass spectrometry. These experiments tend to slow down due to the lack of sophisticated methods for automatic spectrum analysis. Currently, spectrum interpretation is one of the biggest bottlenecks in a proteomics experiment. Manual interpretation of such complex data is very tedious and time consuming. While some instrument manufacturers have developed reasonably effective programs for this problem, they rarely publish these algorithms, and thus the strengths and limitations of these methods are difficult or impossible to assess. Hence, there is need for the development of advanced data analysis algorithms [15–21]. In this work, several new algorithms are discussed that allow for the improved automated reduction of high-resolution mass spectra into a monoisotopic peak list. The proposed name for the unified suite of methods is Mass Spectrum Interpretation and Kernel Extraction (MasSPIKE).

The *m/z* ratio of most ESI product ions lies in the range of 500–5000 Daltons. Since the same mass can have multiple charge states and there can be multiple isotopic peaks at each nominal *m/z* value, a very dense, complicated spectrum can be generated. The first attempt for automated spectrum interpretation was (somewhat erroneously) called "deconvolution" [15–17, 20]. "Deconvolution" was based upon the principle that charge states can take only integer values. It combined peaks of the same mass but different charge states to determine the mass of the ion. Currently, most of the published "deconvolution" algorithms result in spurious peaks due to mis-assignment of charge states. Furthermore, these methods generally perform poorly with low S/N and complex spectra resulting in missed peaks (false negatives). Also, most "deconvolution" methods bias against peaks represented by only one charge state which will be poorly represented in these deconvolution approaches, though the Z score [20] algorithm does not suffer from this drawback.

To overcome these limitations, Horn et al. developed a computer algorithm called THRASH [21] (Thorough High Resolution Analysis of Spectra by Horn). THRASH was the first comprehensive "non-deconvolution" algorithm that addressed the problem of reducing a complex mass spectrum into a mass list with minimal human intervention. It combines various modules of signal to noise (S/N) calculation, charge state determination using the Fourier/

Published online February 7, 2006

Address reprint requests to Professor P. B. O'Connor, Department of Biochemistry, Mass Spectrometry Resource, Boston University School of Medicine, 670 Albany St., Boston, MA 02118, USA. E-mail: poconnor@bu.edu

[*] Also at the Cardiovascular Proteomics Center, 670 Albany St., Boston, MA 02118, USA.

[†] Also at the Department of Electrical and Computer Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, USA.

1044-0305/06/$32.00
doi:10.1016/j.jasms.2005.11.024

Received June 28, 2005
Revised October 13, 2005
Accepted November 15, 2005

Patterson [19] method, and least-squares fitting for determination of monoisotopic mass. It was a remarkable step towards automated spectrum interpretation and represents the current benchmark in the field. However, THRASH is based upon certain modules that can be approached differently to achieve significantly better results. The work presented here aims to develop better individual modules, and then combine them for improved data reduction. The comparative results are presented in each section.

Experimental

The methods are presented here, but their performance characteristics will be discussed below. All the methods are being integrated as part of the open source software package BUDA, [22], and will be available at www.bumc.bu.edu/ftms. MasSPIKE starts with modeling the mean of the noise across the selected *m/z* range of the spectrum. It then identifies isotopic distributions, marks their location, and determines the charge state for each of the identified isotopic distributions to map *m/z* values to corresponding mass values. Overlapping isotopic distributions are then separated, and the charge state is assigned to each of the resolved overlapping distributions. Then, each experimental isotopic distribution (EID) is aligned with its theoretical isotopic distribution (TID) to arrive at the best alignment index for the two distributions. Finally, the monoisotopic mass for each of the resolved isotopic clusters is calculated using results from the previous steps, and the final, minimal, monoisotopic peak list is generated. The mathematical basis of each of these methods is discussed, and the critical equations are "boxed" for the convenience of the reader.

Modeling Noise

Baseline noise in a Fourier transform mass spectrum typically has white noise characteristics. Other sources of noise include random electronic RF (Radio Frequency) interference peaks and chemical noise due to unevaporated solvent clusters, which may make the noise look non-white. To detect peaks, it is critical to know noise levels in a particular region of spectra in the *m/z* domain. This module aims at modeling the mean of the noise. To calculate the S/N ratio across the spectrum, noise mean is characterized as follows.

1. Find the mean of the signal every 1 Da, with 0.5 Da overlap between consecutive *m/z* windows to assure completeness
2. Every 10 Da (value can be changed by user), the window with the minimum mean is assumed to be the noise window
3. A histogram of the intensity values in the noise window is plotted. The histogram is then truncated to eliminate high intensity values caused by signal or RF interference noise peaks, i.e., the histogram is truncated once the intensity occurrence values reach less than 5% of the maximum occurrence value. Then the mean of the intensity values corresponding to the truncated histogram is calculated and this value is defined as the local noise value. This process is done iteratively until the mean converges. A typical example of this procedure is shown in Figure 1a and b. Note that this procedure does not take into account RF interference peaks, but is designed to find the baseline noise level. RF interference peaks will be filtered out in the subsequent modules to eliminate false positives.

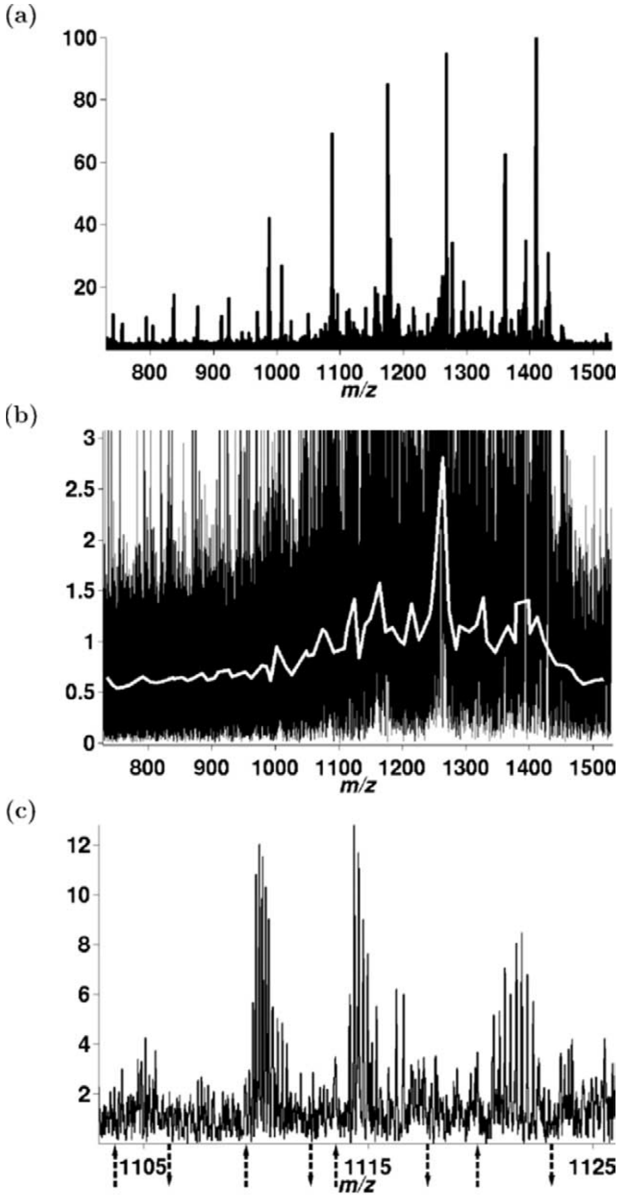


Figure 1. **(a)** Top down spectrum of bovine carbonic anhydrase; **(b)** zoomed in view of the baseline (black), modeled noise baseline (white); **(c)** zoomed-in view of the spectrum, "up" and "down" arrows denote the start and end of an ID respectively.

Isotopic Distribution Identification

The goal here is to identify the locations of isotopic distributions (IDs) based upon the S/N ratio in the spectrum. Here, an ID is identified based upon the fact that S/N ratio for an ID is higher than a particular user defined threshold. This principle is similar to that used in THRASH [21], but this module defines the isotopic distribution boundaries before assigning the charge state, instead of taking a ±0.5 Da window around the highest intensity peak. This step is very important for good performance of charge state determination. It is carried out by the following steps:

1. Scan the spectrum every 1 *m/z* unit from low *m/z* to high *m/z*
2. Check S/N in every 1 *m/z* window
3. S/N = max(signal)/mean(noise), with noise value defined as above
4. If S/N is greater than the user-defined threshold (default value = 3), mark the window as potentially containing an ID
5. Combine together consecutive potential ID windows and output: C1 (start ID *m/z* value), C2 (end ID *m/z* value)
6. If there is only one peak within {C1, C2} with S/N greater than the threshold, it typically indicates an RF interference noise peak and is discarded, because real mass spectral peaks almost always have an isotopic signature

Note that C1 and C2 are those values of *m/z* where the S/N hits the threshold value in the window first and last respectively. A table of {C1, C2} values is constructed and used as input to the charge state determination routine. A bovine carbonic anhydrase "top-down" spectrum (Figure 1a) was used to test MasSPIKE, and will be discussed below. {C1, C2} values for *m/z* region of 1103–1133 from this spectrum are plotted in Figure 1c as arrows below the spectrum with the "up" arrows indicating the start, C1, and "down" arrows indicating the end, C2, of individual IDs. If the threshold value is kept too low, some random noise spikes may be picked, while too high value will miss the low S/N isotopic distributions. So a value of threshold = 3 was found empirically to be an optimal balance between the two cases, but it can also be adjusted manually by the user.

Charge State Determination

Each entry in the {C1, C2} table constructed above is subjected to the process of charge state determination. Previously this problem has been approached by taking the Fourier Transform, Patterson, and combination [19] charge state maps (Z-maps) of the isotopic distribution. These methods generally work well with good signals, but all charge state determination methods fail under conditions of low S/N or overlapping IDs so that these methods should be compared against each other under those conditions. In addition, a new method using the Matched Filter [23, 24] approach has been developed and compared to the previous methods in the Discussion section.

The Matched Filter (MF) Method for Charge State Determination

In charge state determination, the goal is to design a detector for a specific known pattern (in this case, the theoretical isotopic distribution for a particular molecular weight while varying the charge state). In pattern recognition literature, a standard method for approaching this problem is the use of a matched filter [24]. Let E = vector representing experimental isotopic distribution (EID), T = matrix with N_z (number of possible charge states) rows, such that: Zth row of T, T(Z), is a vector representing the theoretical isotopic distribution (TID) for a given charge state Z. T(Z) is constructed as follows. An approximate average molecular weight (MW_{approx}) can be calculated from the location of EID and the charge state, Z, under consideration (MW_{approx} = *m/z* x Z, where *m/z* is the location of the center of the EID under investigation). For a given MW_{approx}, elemental composition is determined using the average composition of a model amino acid, averagine [18]. Based on the elemental composition, the Mercury [25] algorithm is used to generate the peak intensities of the TID. Peak width at half height for generating the TID is determined from EID, the value being the same as that of the highest peak of the EID. Knowing the peak heights and the width, each peak is generated assuming a Lorentzian [1] peak shape. The TID is finally generated as the sum of individual Lorentzian peaks.

Given an observation E, T(Z) vectors, for all the different possible Z values, are generated as discussed above, for each charge state Z, the matched filter output is then calculated as follows:

$$M(Z, n) = \sum_{k=-L}^{L} E(k)T(Z, k-n) \quad (1)$$

where L is the maximum of the lengths of E and T. This is equivalent to

$$M(Z, n) = E(n) * T(Z, -n) \quad (2)$$

where * denotes the convolution operator. Note that this is also equivalent to taking a cross-correlation of the EID and TID. Define

$$M_{\max}(Z) = \max_{n} M(Z, n) \quad (3)$$

$$N(Z) = \arg\max_{n} M(Z, n) \quad (4)$$

where arg $\max_n(M(Z,n))$ indicates the value of n that corresponds to the maximum value of $M(Z,n)$. Since the signal intensities and length of E may vary highly in a given experiment, it is important to normalize both E and T while calculating the "score" of closeness of E and T(Z). This "score" is given by the value of cross-correlation coefficient r(Z), which is given by the following expression:

$$r(Z) = \frac{\Sigma_i(E(i) - M_E)(T(Z, i - N(Z)) - M_{T(Z)})}{\sqrt{\Sigma_i(E(i) - M_E)^2}\sqrt{\Sigma_i(T(Z, i - N(Z)) - M_{T(Z)})^2}} \tag{5}$$

where M_E and $M_{T(Z)}$ are the means of E and T(Z) respectively. The theory of matched filters [24] tells us that the value of r(Z) will be maximum when E and T(Z) belong to the same class, which in this case means that they represent the same Z. So the charge state is estimated as follows:

$$Z_{est} = \arg \max_{Z} r(Z) \tag{6}$$

This means that the charge state that corresponds to the maximum value of r(Z) can be assigned as the estimated true charge state. This works satisfactorily provided the given input signal E is composed of only one charge state. In practice, a given input signal may represent multiple isotopic distributions of different charge states. Thus, the Z values corresponding to r(Z) greater than a certain user-defined threshold are assigned to be the true charge states. Isotopic cluster(s) corresponding to the above determined charge state(s) are then subtracted from the observed distribution, and the residual signal undergoes the same procedure to look for any more charge states represented by the residual data similar to the procedure used by THRASH. The process continues till the final residual cannot be assigned any charge state since r(Z) value is below the threshold for all values of Z. The procedure for determining useful threshold values is discussed below.

Alignment Between Theoretical and Experimental Isotopic Distribution

A mass spectrum does not generate a unique mass value for large molecules due to the presence of multiple isotopes of the constituent elements. So the question arises as to what mass value should be reported. One way is to report the chemical average mass using average of isotopic peaks, but this suffers from the problem that carbon isotope variability across different organisms limits the mass accuracy [26, 27] to about 10 ppm. The most significant and accurate mass that can be reported is the monoisotopic mass because its value is unaltered by isotopic variability. The monoisotopic mass (M) of a molecule is the sum of the masses of the lowest mass isotope for each of the elements present in the molecule. The relative abundance of the monoisotopic peak decreases with increase in the molecular weight because of the increased probability for the presence of heavier isotopes with increasing molecular mass. The monoisotopic peak is typically not visible experimentally when molecular weight is higher than 5 kDa because the tiny peak is buried in the noise. Thus, there is need for the development of a method that can estimate the monoisotopic mass based upon the experimentally observed isotopic profile. Previously, this problem was approached by Senko et al. [18] and Horn et al. [21] using a least-squares fit between the theoretical and experimental isotopic distribution. This method generally works well, but breaks down in the limit of low number of ions or low S/N ratio. This module targets at solving this problem rigorously by analyzing the isotopic distributions.

As discussed previously, [28] the EID can be interpreted as a result of a multinomial experiment (with the number of trials equal to the number of ions) having multiple outcomes, each with probability t_i, where t_i represents the area of each individual peak in the TID (i.e., t_0 fraction of the total ions in the cell contains no higher isotopes, t_1 fraction of the total ions contain exactly one +1 Dalton higher isotope (e.g., ^{13}C), etc.). Let vector E represent EID peaks areas, where e_i corresponds to t_i in the TID. E is a Gaussian random vector with mean T and covariance matrix Σ_N (eq 8), where T is composed of t_is, and is obtained using the polyaveragine [18] model and the Mercury [25] algorithm. The probability of observing E, given that the number of ions in the cell is N, is the given by [29]:

$$P(E|N) = \frac{e^{-0.5(E-T)'\Sigma_N^{-1}(E-T)}}{\sqrt{(2\pi)^n det(\Sigma_N)}} \tag{7}$$

where Σ_N is given by the following expression:

$$\Sigma_N = \frac{1}{N}\begin{pmatrix} t_1(1-t_1) & -t_1t_2 & \dots & -t_1t_n \\ -t_2t_1 & t_2(1-t_2) & \dots & -t_2t_n \\ . & . & \dots & . \\ . & . & \dots & . \\ -t_nt_1 & -t_nt_2 & \dots & t_n(1-t_n) \end{pmatrix} \tag{8}$$

where t_is are the components of T, defined by the theoretical isotopic abundances.

For big molecules, only a part of E is observed. The goal, therefore, is to align it with the appropriate indices of T to determine the monoisotopic mass. Thus:

$$P(E|N, i) = \frac{e^{-0.5(E-T_i)'\Sigma_{N_i}^{-1}(E-T_i)}}{\sqrt{(2\pi)^n det(\Sigma_{N_i})}} \tag{9}$$

which means that E is a normal (Gaussian) random vector with mean T_i and covariance matrix Σ_i (both mean and covariance matrix vary with the index). The index of T corresponding to first "visible" value of E is estimated using a Maximum Likelihood estimator as follows:

$$\text{index} = \arg\max_{i} P(E|T_i, \Sigma_{N_i}) \qquad (10)$$

where

$$T_i = T(i, \ldots, i + L_E - 1) \qquad (11)$$

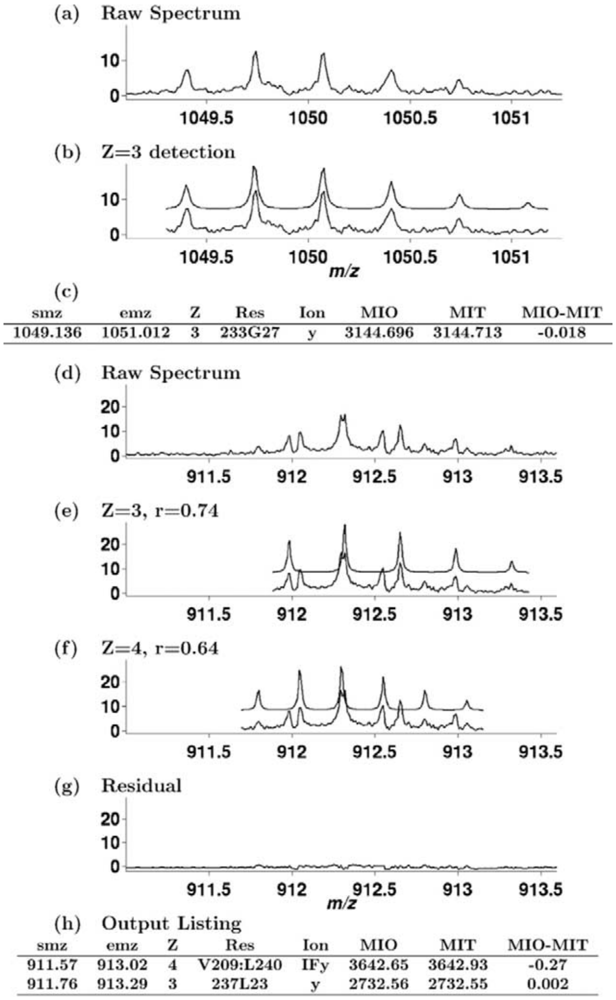
L_E = length of E.

This means that the first L_E values of T are aligned with E, calculate the probability of its occurrence from expression 9, and then T is shifted by 1, until all the possibilities have been considered. The shift of T that corresponds to the highest value of the probability is assigned to be the true index of the first observed value of E. This procedure is illustrated in detail in the Results and Discussion section.

Finally, the monoisotopic mass (M) is calculated as following:

$$M = \frac{m_1}{z} \times Z_{est} - (index - 1) \times 1.00235 - Z_{est} \times \mathrm{M}+ \qquad (12)$$

where m_1/z is the location of the first "visible" isotopic peak in the EID (i.e., peak location corresponding to E ((1), Z_{est} is the estimated charge state (eq 6), 1.00235 is the average mass difference between the centroid of each adjacent isotopic peak for poly-averagine [21], and M+ is the mass of the charge carrier (e.g., 1.0073 for a proton). Assuming the experiment was run in a positive ion mode, a charge state of Z usually means the ion carries Z protons, so the corresponding mass of Z protons (default) is subtracted to get M.



smz	emz	Z	Res	Ion	MIO	MIT	MIO-MIT
1049.136	1051.012	3	233G27	y	3144.696	3144.713	-0.018

smz	emz	Z	Res	Ion	MIO	MIT	MIO-MIT
911.57	913.02	4	V209:L240	IFy	3642.65	3642.93	-0.27
911.76	913.29	3	237L23	y	2732.56	2732.55	0.002

Figure 2. **(a)** Experimental data from Figure 1 showing an isotopic distribution of a fragment of bovine carbonic anhydrase **(b)** TID with Z=3 (top) and EID (bottom). TID shift corresponds to maximum value of cross-correlation coefficient (0.954) between the two, **(c)** snapshot of output listing corresponding to above fragment (y27). **(d)** Experimental distribution from Figure 1 showing two overlapping distributions **(e)** Z=3, r=0.74 **(f)** Z=4, r=0.64; **(g)** residual after subtracting TID's of **(e)** and **(f)** **(h)** final output listing.

Results and Discussion

Figure 1a shows a top down spectrum of carbonic anhydrase against which MasSPIKE has been tested. Figure 1b shows zoomed-in view of the baseline of Figure 1a and noise mean variation as a function of *m/z* (white line passing through the baseline). The plot is consistent with the variation of baseline noise in the spectrum. Noise modeling serves to provide a noise mean value to be used in S/N calculation for the identification of ID locations. Note that noise is not truly white (flat across *m/z* range), which is due to the "chemical noise" effect caused by unevaporated solvent clusters formed by the electrospray source.

Figure 1c shows the result of the ID identification module applied to one low S/N region of the spectrum. Figure 1c shows the ID boundaries for the *m/z* range of 1103–1132 with up and down arrows indicating the start and end of an ID respectively. Very closely spaced IDs (e.g., between *m/z* 1114 and 1117) are not separated as seen in the figure. Such cases and overlapping distributions are separated later in the charge state determination routine. ID determination allows MasSPIKE to identify the IDs representing both low and high charge states without bias. This method was found to correct a limitation of THRASH, which uses ±0.5 *m/z* window around the maximum intensity peak for the charge state determination, restricting the analysis to charge states greater than 2. MasSPIKE, therefore, can be used for both MALDI (typically representing 1+ or perhaps 2+ charge states) and electrospray (typically representing high charge states) spectra. Also, THRASH assumes that the isotopic distribution has a symmetrical Gaussian shape around the highest peak, which holds true for molecular weights greater than ~5 kDa, while MasSPIKE makes no such assumption, so is suited for any kind of ID shape.

One of the major challenges encountered in the interpretation of dense, complex spectra is that there is a high chance that the peak of interest is affected by interfering noise peaks or peaks from other signal components (e.g., other isotopic distributions). Figure 2 shows a simple case when input signal EID (Figure 2a) represents only Z = 3. Figure 2b shows the plot of T(3) (shifted up) and EID, with shift in T(3) corresponding to maximum value of cross-correlation coefficient (0.954) between the two. Note that r(Z) varies from 0 to 1, so r(3) = 0.954 indicates a very good match between EID and T(3). The output list from MasSPIKE, Figure 2c, shows the starting and ending values of the distribution (smz, emz), the charge state (Z), the assigned amino acid residue region (Res, given the sequence) and ion type (Ion), as well as the observed monoisotopic ion mass (MIO), theoretical ion mass (MIT), and the mass error in Daltons. However, this is an easy case with good S/N ratio, and no overlapping distributions. The real test of automated analysis methods comes at low S/N with distorted peak shapes. Figures 2 and 3 show a couple of such cases extracted from Figure 1a. It is important to note that Figures 2 and 3 are drawn on the same vertical scale as Figure 1a (which is normalized to 100). Thus, Figures 2 and 3, with the highest intensity values in the 3-15 range, represent parts of the spectrum where the S/N ratio is the lowest, and in particular, Figure 3 depicts a case where input signal came from one of the noisiest portions of the spectrum.

Figure 2d shows the case when input signal represents two charge states (Z = 3 and Z = 4), which share a central peak at *m/z* = 912.3. The two charge states are successfully identified and subtracted from the input signal as shown, with TID shifted and plotted on the top of the EID. Note that here MasSPIKE is simultaneously detecting Z = 3 and 4 (Figure 2e and f) and the residual after subtraction is free of peaks (Figure 2g). By comparison, THRASH proceeds by identifying the charge state represented by the combo [19] routine, and then subtracts the TID from the experimental data. With such an approach, if any of the Z = 3 or Z = 4 is detected by the combo routine (which is likely), the peak at *m/z* = 912.3 (common peak to both Z = 3 and Z = 4) will be removed and the next charge state will not be assigned because the isotopic pattern is perturbed due to subtraction. MasSPIKE attempts to find all the charge states that give cross-correlation coefficient, r(Z), value greater than a certain threshold (default = 0.45) before carrying out the subtraction. This allows for assignment of a greater number of charge states. Note that assignment of this threshold represents a balance between missing peaks and generation of false positives. The default threshold value of 0.45 was empirically determined to be a moderate value, but this value can also be altered by the user.

Figure 3a shows an input signal from region *m/z* = 1380.7–1385.5 of the bovine carbonic anhydrase spectrum. MasSPIKE was used for determination of various charge states present in the signal. In this case, four isotopic distributions are identified with multiple distributions sharing isotopic peaks and the Z = 20 distribution at *m/z* is identified (Figure 3b) and removed (Figure 3c). For higher charge states, especially when the sampling rate of the spectrum is low (which is the case at higher *m/z* since sampling rate in *m/z* domain drops as *m/z* increases in FTMS instruments), it is sometimes difficult to distinguish between the consec-

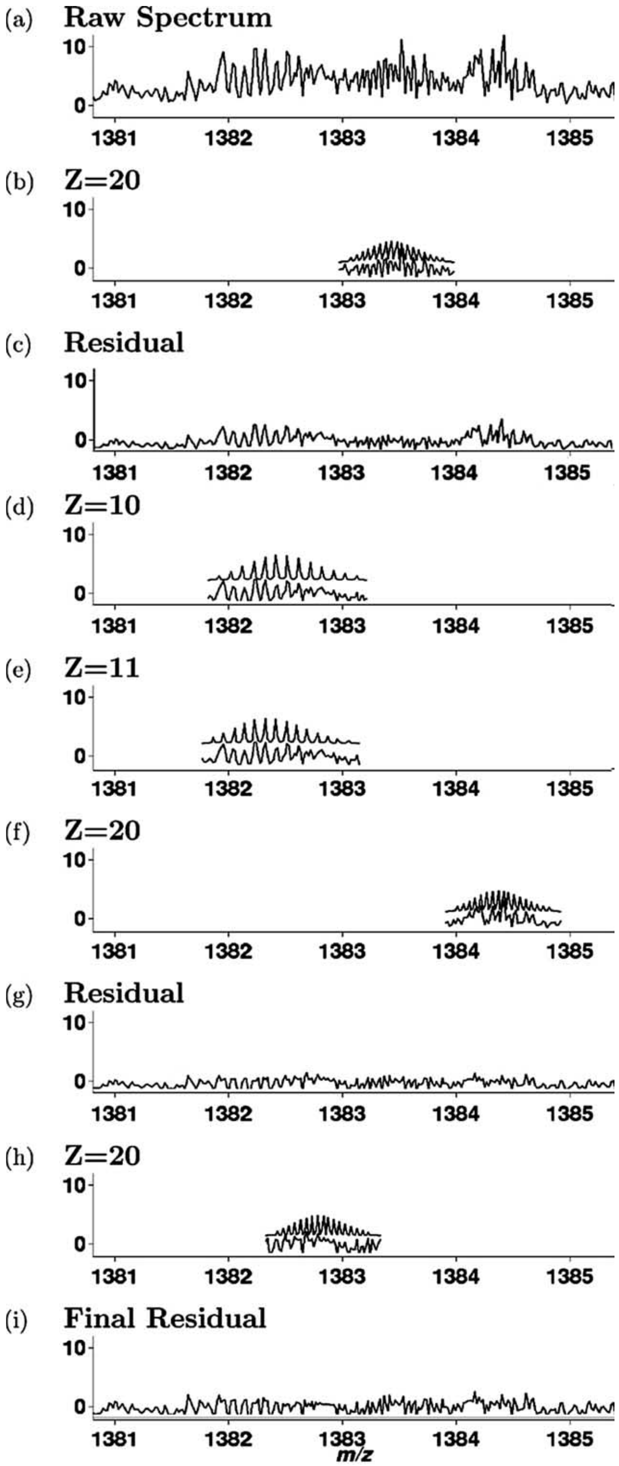


Figure 3. Experimental data from Figure 1 showing an extremely low S/N region of the spectrum that contains 4 overlapping distributions. **(a)** Raw data **(b)** Z=20, r=0.68; **(c)** residual after subtraction of (b); **(d)** Z=10 (r=0.576), **(e)** Z=11 (r=0.566) and **(f)** Z=20 (r=0.65) detected simultaneously; **(g)** residual after subtraction of **(d)**, **(e)** and **(f)**; **(h)** Z=20 (r=0.496); **(i)** residual of experimental signal after subtraction of **(h)**.

utive charge state values. For example, for the *m/z* region between 1381.6 and 1383.2 (Figure 3d and e), the method identifies the charge state values to be either 10 or 11 (though 10 is slightly more likely to be true, r = 0.576, than the case of Z = 11 where r = 0.566). In ambiguous cases like this, a flag is marked and it is left for the user to decide about the true charge state based on the knowledge from the protein sequence, or supplementary information from other portions of the spectrum. Furthermore, MasSPIKE identified two more EIDs with Z = 20 in this region of mass spectrum. These masses could not be assigned to a particular fragment ion from the given sequence. However, the approximate difference between the two higher Z = 20 ion masses corresponds to the loss of a water molecule, which commonly appears at high molecular weight. For example, the approximate molecular weight for the EID represented by the *m/z* region 1383.9–1384.9 is 1384.4 × 20 = 27,688, while that for the *m/z* region 1383–1384 is 1383.5 × 20 = 27,670. The difference of the two species (27,688–27,670 = 18) corresponds to the loss of a water molecule, which suggests the assignment of Z = 20 is correct. Also, the final residual from this region, Figure 3i seems to contain one or two remaining isotopic distributions. This is an artifact that arises due to the imperfect subtraction of TID from EID, and often happens because of the non-ideal peak shapes of the EID in low S/N conditions as seen in Figure 3d and e. Since the residual in Figure 3i contains an artifact and not real signal, no further charge state assignments are generated because MasSPIKE does not yield high enough quality assignment (cross-correlation coefficient, r >0.45) for any further charge states. Note that EID and TID take on negative values in some cases (Figure 3b–i) because both the EID and TID are normalized, which involves subtraction of the mean, while computing their cross-correlation coefficient as shown in eq 5. The Supplementary Material data (which can be found in the electronic version of this article) shows the case when the input signal represents 3 isotopic distributions (Z = 1, 3, and 4), sharing multiple peaks in the region of *m/z* 1221–1227.

It is important to test the matched filter method of charge state determination against established methods in an unbiased manner. To this end, 26 electrospray spectra of myoglobin, representing 775 isotopic distributions (resulting from charge states for the whole molecule, water losses and phosphate adducts for the whole protein, and one contaminant species with Z = 1) with S/N of 1–100, were acquired and each *m/z* region corresponding to Z = 1–22 in each spectrum (regardless of the presence/absence of signal) was analyzed by four different methods. The percentage correct answers for each method are plotted in Figure 4. BUDA (Boston University Data Analysis) [22] was used to determine the charge states using the Fourier, Patterson, and combo charge state determination methods [19]. In this analysis, the MF method gave correct answers 91% of the time. Of the missed 77 assignments, manual post analysis showed no apparent signal in 50 of them, and the remaining 27 misassigned the charge state by ±1.

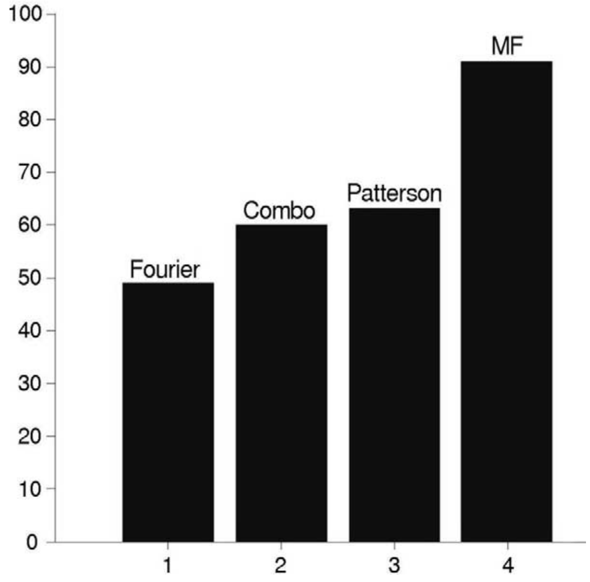


Figure 4. Comparison of different charge state determination methods on 775 isotopic distributions from 26 electrospray spectra of myoglobin.

There are certain points that need to be addressed while generating the TID. Generating good model distributions is the key to good results. Although the sampling rate is constant in the frequency domain, due to the inverse proportionality relation between frequency and *m/z*, the sampling rate of an FT mass spectrum is not the same over the whole *m/z* range, it decreases with an increase in *m/z*. Thus, parameters for generating good model distributions (peak width, sampling rate, maximum and minimum possible Z (MAXZ and MINZ)), must vary with the mass spectral region of interest, and in MasSPIKE they are based upon the observed data and shift through the *m/z* range. Furthermore, it is important to use unapodized spectra, zero filled once for the experimental data and true line shapes for the TIDs to generate the best matches. For the TID model, the peak width for generating the Lorentzian peaks in the TID is defined by the width of the highest peak in the EID. MINZ is defined by the observed isotopic distribution width. e.g., If the EID spans 1.1 Da, MINZ = 1 but if the EID is only 0.9 Da wide, MINZ = 2, so that it contains at least two peaks. This helps eliminate most of the RF interference noise peaks which usually consist of a single high spike. Also, special consideration is given to the number of TID peaks involved in the resulting cross-correlation coefficient. For example, if there is only one peak of the TID matching with the EID that results in the maximum cross-correlation value, it is discarded as a false positive, since it is highly unlikely for biomolecules to have an isotopic distribution with only one peak. MAXZ is defined by the peak width of highest peak, e.g., Peak Width at Half Height < (1/Maximum Z). Sometimes,

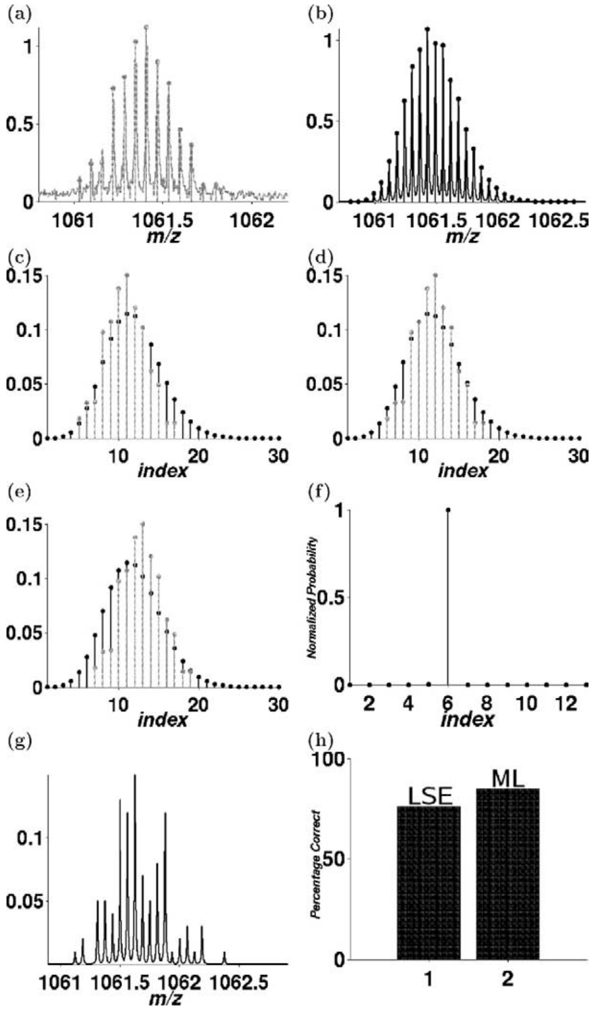


Figure 5. **(a)** EID of myoglobin when Z=16; **(b)** TID of myoglobin, alignment of the EID with **(c)** TID shifted by 5; **(d)** TID shifted by 6 **(e)** TID shifted by 7; **(f)** normalized probability of alignment as a function of varying TID indices; **(g)** alignment of myoglobin IDs using 3150 simulations (100 ions in each simulation); **(h)** a typical Monte-Carlo generated myoglobin isotopic distribution with only 100 ions.

when there are many noise peaks around the main peak, a false identification for a high charge state is generated. Filters have been added to remove these false positives by comparing the peaks of Fourier charge state maps of the theoretical and experimental data. Also, it is required that for a particular molecular weight, the observed isotopic distribution should be wide enough to represent the peaks that are of intensity at least 60% or greater than the maximum intensity. For example, for an observed distribution of molecular weight 10,000, the distribution should be wide enough to encompass at least 5 peaks (>60% of maximum intensity). Otherwise, it most likely arises as a false positive. These considerations lead to reduced number of false positives and overall better performance.

Figure 5 demonstrates the alignment of a typical experimental isotopic distribution (Figure 5a) with the theoretical isotopic distribution (Figure 5b). The EID and TID are represented by grey and black stick plots respectively in Figure 5c–e. Figure 5c–e shows the alignment of the EID with the TID, with the TID being shifted by 5, 6, and 7 in Figure 5c, d, and e, respectively. Figure 5f shows the probability of alignment of the EID against the TID with varying shift of TID. A shift of 6 in the TID gives the best alignment as depicted in Figure 5d and 5h. The normalized probability plot (Figure 5h) shows that the probability the EID and the TID are aligned properly when the shift is 6 is much higher than its nearest-neighbor (index = 5). These results are typical with such high S/N (~20) clean isotopic distributions. However, all alignment methods will work well under these conditions. It is important to test these methods under low S/N and low ion count conditions where large statistical variance occurs in isotopic abundance [28].

When only 100 ions are present in the isotopic distribution, large statistical variation in the isotopic abundance occurs; a typical 100 ion isotopic distribution for myoglobin (16.7 kDa, 16^+) is shown in Figure 5g. To test the ML method versus the least-squares method, 3150 Monte Carlo simulated distributions were generated with only 100 ions per simulation, and the two alignment methods were tested against these distributions. The tests revealed that ML method works correctly 85% of the time, compared to the least-squares error method which gave 76% correct results (Figure 5h). Note that it is more difficult to estimate the true index when the distribution is generated by a fewer number of ions since the EID deviates from the TID due to high variance among the isotopic peaks as discussed in our previous work [28].

After the determination of monoisotopic masses (as discussed above), it is desirable to automatically assign the protein fragments that generated those masses. This requires the knowledge of how a protein or peptide fragments in an experiment [30]. MasSPIKE was used to generate theoretical masses of the b and y ions. Internal fragment masses and masses with common losses (e.g., water loss from a molecule) were calculated knowing the sequence of the protein. The observed masses that match with the theoretical masses of the whole protein and its fragments are then evaluated. A complete analysis of the bovine carbonic anhydrase spectrum revealed the presence of 165 isotopic clusters after eliminating all false positives, which were matched to the closest masses of b or y ions, the corresponding internal fragment ions, and some common losses like water loss, or ammonia loss from a y-ion. The complete deconvolved spectrum representing monoisotopic masses is shown in Figure 6. Only abundant peaks are labeled, but the complete monoisotopic mass list is included in the supplemental data (which can be found in the electronic version of this article). Due to the high-energy used for fragmentation, the precursor ion is not observed.

One important limitation of MasSPIKE at this time is the assumption implicit in the poly-averagine model,

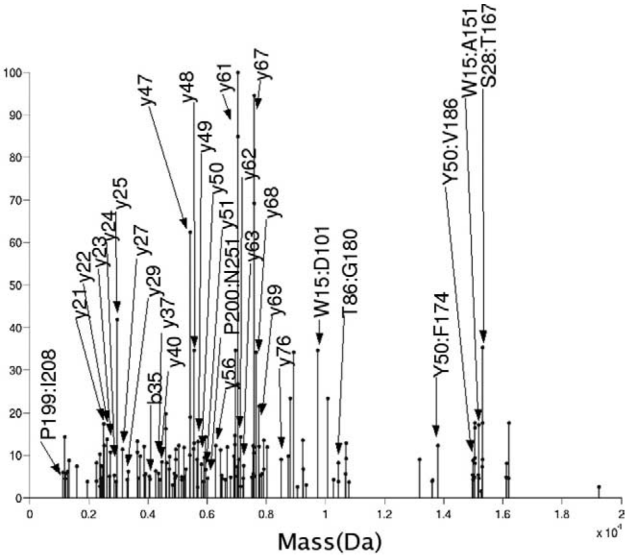


Figure 6. Final monoisotopic mass plot of bovine carbonic anhydrase (The full table of peaks is included in the Supplementary Material section).

specifically that the molecule of interest is an "average" protein. Clearly, this assumption fails routinely. A future modification to MasSPIKE will include a DNA and Glycan model as well as the ability to adjust the model manually.

Conclusions

MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction), a suite of data analysis algorithms, has been developed. The goal is to reduce a high-resolution mass spectrum into a monoisotopic peak list. MasSPIKE identifies isotopic peak cluster locations, determines the charge state for each of the isotopic clusters, resolves overlapping isotopic distributions, aligns the experimental and theoretical distributions, and generates a monoisotopic mass list. If the protein sequence is available, the calculated masses are matched for possible assignments. The method has been applied and tested against complex top-down spectra of bovine carbonic anhydrase. The isotopic distribution identification method is able to identify and mark locations corresponding to both low and high charge states. The Matched Filter charge state determination routine worked correctly 91% of the time for unbiased test data as compared to the standard routines [19], which vary from 48–64% accuracy. MasSPIKE is capable of identifying multiple charge states in the input signal sharing multiple peaks. Alignment of the theoretical and experimental isotopic distributions with only 100 ions (and hence, high statistical variance) in the distribution gave 85% correct results as compared to 76% given by the least-squares fitting method.

Acknowledgments

The authors gratefully acknowledge Professor W. Clem Karl and Dr. Amit Juneja for their valuable suggestions for the design of MasSPIKE, Dr. Bogdan Budnik and Dr. Judith Jebanathirajah for scientific discussions, Dr. Cheng Zhao, Jason Cournoyer, Raman Mathur, Dr. Cheng Lin, Vera Ivleva, and Konstantin Aizikov for experimental assistance, and Dr. Alan Rockwood for sharing the Mercury algorithm source code. This work was supported in part by Federal funds from the National Center for Research Resources under grant no. P41-RR10888 and the National Heart, Lung, and Blood Institute under contract no. HHSN268200248178C, and the American Chemical Society.

References

1. Marshall, A. G.; Verdun, F. R. Fourier Transforms in NMR, Optical, and Mass Spectrometry; Elsevier: Amsterdam, 1990.
2. Amster, I. J. Fourier Transform Mass Spectrometry. *J. Mass Spectrom.* **1996,** *31,* 1325–1337.
3. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. Matrix-assisted Ultraviolet Laser Desorption of Non-Volatile Compounds. *Int. J. Mass Spectrom. Ion Phys.* **1987,** *78,* 53–68.
4. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989,** *246,* 64–71.
5. Budnik, B. A.; Moyer, S. C.; Pittman, J. L.; Ivleva, V. B.; Sommer, U.; Costello, C. E.; O'Connor, P. B. High pressure MALDI-FTMS: Implications for Proteomics. *Int. J. Mass Spectrom. Ion Processes* **2004,** *234,* 203–212.
6. O'Connor, P. B.; Budnik, B. A.; Ivleva, V. B.; Kaur, P.; Moyer, S. C.; Pittman, J. L.; Costello, C. E. A High Pressure Matrix-Assisted Laser Desorption Fourier Transform Mass Spectrometry Ion Source Designed to Accommodate Large Targets with Diverse Surfaces. *J. Am. Soc. Mass Spectrom.* **2003,** *15,* 128–132.
7. Moyer, S. C.; Budnik, B. A.; Pittman, J. L.; Costello, C. E.; O'Connor, P. B. Attomole Peptide Analysis by High Pressure Matrix-Assisted Laser Desorption/Ionization Fourier Transform Mass Spectrometry. *Anal. Chem.* **2003,** *75,* 6449–6454.
8. Henry, K. D.; Quinn, J. P.; McLafferty, F. W. High-Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Chem. Soc.* **1991,** *113,* 5447–5449.
9. Shen, Y. F.; Tolic, N.; Zhao, R.; Pasa-Tolic, L.; Li, L. J.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. High-Throughput Proteomics Using High Efficiency Multiple-Capillary Liquid Chromatography with On-Line High-Performance ESI FTICR Mass Spectrometry. *Anal. Chem.* **2001,** *73,* 3011–3021.
10. Aebersold, R. A Mass Spectrometric Journey into Protein and Proteome Research. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 685–695.
11. McLafferty, F. W. High Resolution Tandem FT Mass Spectrometry Above 10 kDa. *Acc. Chem. Res.* **1994,** *27,* 379–386.
12. Kelleher, N. L. Top-down Proteomics. *Anal. Chem.* **2004,** *76,* 196A–203A.
13. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *J. Am. Chem. Soc.* **1999,** *121,* 806–812.
14. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.* **1998,** *13,* 3265–3266.
15. Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass Spectra of Multiply Charged Ions. *Anal. Chem.* **1989,** *61,* 1702–1708.
16. Reinhold, B. B.; Reinhold, V. N. Electrospray Ionization Mass Spectrometry: Deconvolution by an Entropy Based Algorithm. *J. Am. Soc. Mass Spectrom.* **1992,** *3,* 207–215.
17. Henry, K. D.; McLafferty, F. W. Electrospray Ionization with Fourier-Transform Mass Spectrometery. Charge State Assignment from Resolved Isotopic Peaks. *Org. Mass Spectrom.* **1990,** *25,* 490–492.
18. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 229–233.
19. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply Charged Ions. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 52–56.

20. Zhang, Z.; Marshall, A. G. A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 225–233.
21. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 320–332.
22. O'Connor,, P. B. Boston University Data Analysis www.bumc.bu.edu/ftms.
23. Haykin, S. Communications Systems; John Wiley and Sons, Inc.: Singapore, 1994, pp 413–418.
24. Duda, R. O.; Hart, P. E.; Stork, D. H. Pattern Classification; Wiley Interscience: New York, 2001, pp 325–326.
25. Rockwood, A. L. Ultrahigh-Speed Calculation of Isotope Distributions *Anal. Chem.* **1996,** *68,* 2027–2030.
26. Beavis, R. C. Chemical mass of carbon in proteins. *Anal. Chem.* **1993,** *65,* 496–497.
27. Zubarev, R. A.; Hakansson, P.; Sundqvist, B. Accuracy Requirements for Peptide Characterization by Monoisotopic Molecular Mass Measurements. *Anal. Chem.* **1996,** *68,* 4060–4063.
28. Kaur, P.; O'Connor, P. B. Use of Statistical Methods for Estimation of Total Number of Charges in a Mass Spectrometry Experiment. *Anal. Chem.* **2004,** *76,* 2756–2762.
29. Poor, H. V. *An Introduction to Signal Detection and Estimation (Springer Texts in Electrical Engineering), 2nd ed.*; Springer Verlag: New York, 1994, p 53.
30. Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrom.* **1984,** *11,* 601.