

**Evaluation of a Simultaneous
Mass-Calibration and Peak-Detection
Algorithm for FT-ICR Mass Spectrometry**

J.E. Eckel-Passow, T.M. Therneau, A.L. Oberg,
C.J. Mason, D.C. Muddiman

Technical Report #76
January 2006

Copyright 2006 Mayo Foundation

EVALUATION OF A SIMULTANEOUS MASS-CALIBRATION AND PEAK- DETECTION ALGORITHM FOR FT-ICR MASS SPECTROMETRY

Eckel-Passow, J.E.,^{1*} Therneau, T.M.,¹ Oberg, A.L.,¹ Mason, C.J.,² Muddiman, D.C.^{2-3,†}

¹Division of Biostatistics, Department of Health Sciences Research, ²W.M. Keck FT-ICR Mass Spectrometry Laboratory, Mayo Proteomics Research Center, ³Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine, Rochester, MN 55905

[†]Current Address: W.M. Keck FT-ICR MS Laboratory, Department of Chemistry, North Carolina State University, Raleigh, NC 27695

*To whom correspondence should be addressed:

Department of Health Sciences Research

Mayo Clinic

200 First Street SW

Rochester, MN 55905

Tel: (507) 538-6512

Fax: (507) 266-2478

e-mail: eckel@mayo.edu

Running Title: Mass-Calibration & Peak-Detection Algorithm

ABSTRACT

Electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (ESI-FT-ICR-MS) is a potentially superior biomarker discovery platform because it offers high mass-measurement accuracy and high mass-measurement precision as well as high resolving power over a broad mass-to-charge (m/z) range. The electrospray ionization of large molecules is characterized by multiple charge states for each molecular species, which is advantageous for mass calibration and peak detection. Herein, we describe and evaluate a simultaneous mass-calibration and peak-detection algorithm that exploits resolved isotopic peak-spacing information as well as space-charge frequency shifts across isotopic clusters that represent the same molecular species but differ in charge states by integer values. The algorithm is performed on un-windowed spectra in the frequency domain, where both peak shape and peak width are known, thus requiring fewer and more stable parameters in comparison to modified data. Furthermore, un-windowed frequency data are uncorrelated and equivariant and thus preferred for modeling fitting. Previous calibration work has focused on the importance of incorporating internal calibrants in every sample, using external calibrants, or a combination of both with the goal of improving mass-measurement accuracy and mass-measurement precision. Our work suggests that a single external calibration sample is sufficient for data with high mass resolving power and an ionization process that allows large molecules to be characterized by multiple charge states.

General Notation

c = number of ions ($j = 1, \dots, c$)

d = number of zero fills prior to FFT

f = cyclotron frequency

k = number of neutrons

m = mass

n = number of sample points ($i = 1, \dots, n$)

p = number of possible charge states ($h = 1, \dots, p$)

s = isotopic cluster

t = time

z = charge on the molecular ion

1. INTRODUCTION

Protein mass spectrometry (MS) is becoming a popular tool for biomarker discovery, and in particular, for the early detection of cancer and for understanding disease prognosis and progression. MS allows one to directly assess protein expression as opposed to inferring protein expression from mRNA expression profiles. MS consists of a diverse range of technologies and techniques and our efforts are directed towards accurately characterizing complex mixtures, e.g., the plasma proteome which requires high-end instrumentation. Characterizing the plasma proteome is a challenging problem and review papers by Diamandis [1] and Anderson and Anderson [2] delineate that not only is the plasma proteome complex, it spans a wide dynamic range ($>10^{10}$). For example, prostate specific antigen exists at approximately 1/4,000,000 the concentration of albumin and approximately 1/300 of other common protein biomarkers such as C-reactive protein. As a result, it is difficult to detect and identify low-abundance proteins in plasma even though these ultimately may be the most informative molecules for biomarker discovery. However, with continued improvements in mass-spectrometry technology, the potential is being more fully realized. A critical

issue that remains is the ability to analyze these complex datasets with sophisticated algorithms because it is not feasible to interpret the data manually.

Data reduction methods for MS data are an active area of research and one that we pursue here. In particular, an algorithm to reduce an individual MS spectrum that consists of about one-million data points down to a set of biologically-meaningful peaks while simultaneously performing mass calibration is presented. This work is motivated by data that utilizes electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (ESI-FT-ICR-MS). Here, components are charged using ESI [3-4], a technique that attaches one or more charges to each molecular species such that the average number of charges is related to the size of the component with a general rule for peptides being one charge per 1000 Daltons. The mass analyzer is based on FT-ICR technology [5-6] that has the advantage of extremely high resolving power, high mass-measurement accuracy, high mass-measurement precision, and wide dynamic range. The proposed algorithm is able to utilize these qualities to approach mass calibration and peak detection simultaneously.

The proposed algorithm is based on work by Horn et al. [7], who took advantage of the high-resolving power of ESI-FT-ICR-MS data and developed an automated peak-detection algorithm called thorough high resolution analysis of spectra by Horn (THRASH). Our proposed algorithm differs from THRASH in two primary respects. First, it is implemented entirely in the frequency domain where both peak shape and peak width are known to follow a *sinc* function [9]. Second, mass calibration is

combined with the peak-detection process by taking advantage of the individual peaks that comprise an isotopic cluster as well as the presence of multiple charge states for each molecular species as a direct result of the ESI process. The benefits of multiply charged molecules for mass calibration have also been discussed by Bruce et al. [8]. Bruce and colleagues developed DeCAL, a calibration procedure that estimates space-charge effects by alternating between frequency space and m/z space to estimate a weighted average space-charge effect across molecular species with multiple charge states. In contrast, our algorithm estimates space-charge effects using un-windowed and zero-filled FFT data by means of a nonlinear model. By taking a modeling approach all necessary mass-calibration parameters, in addition to space-charge effects, are estimated concurrently and classical statistical techniques are applicable.

2. BACKGROUND

FT-ICR Background

The overall signal from an ICR cell is a sum of sinusoids,

$$y(t_i) = \sum_{j=1}^c \alpha_j \cos[2\pi \gamma_j (t_i - \delta_j)] + \varepsilon(t_i) \quad (1)$$

where $y(t_i)$ is the transient signal at time t_i , $i = 0, 1, \dots, n-1$, n is the number of sample points collected across the sampling interval, c is the number of ions in the sample, α_j is the abundance (amount) of ion j , γ_j is the cyclotron frequency (typically expressed in Hz = cycles/second) of rotation of ion j , δ_j is the phase offset of ion j , and $\varepsilon(t_i)$ is the random error. The values α_j and γ_j are the primary quantities of interest, whereas the phase offset δ_j merely reflects that ion j was not directly under the detection plate at

time zero. For computational ease, the number of sample points n is typically a power of two.

The analysis of FT-ICR data proceeds by taking the discrete Fourier transform (DFT) of the transient signal $y(t_i)$ in equation 1, most commonly using the fast-Fourier transform (FFT),

$$DFT[y(t_i)] = W(f_i) = \sum_{j=1}^c [\phi_{j1} \cos(2\pi f_i t_i) + \phi_{j2} \sin(2\pi f_i t_i)]. \quad (2)$$

The FFT returns the cosine and sine coefficients, ϕ_{i1} and ϕ_{i2} , respectively, for a set of frequency values $f_i = i$, $i = 0, 1, \dots, n-1$, where the f_i are in units of total cycles. The coefficients of the FFT in equation 2 represent the covariance of each probe frequency f_i with the true signal, i.e.,

$$\begin{aligned} \phi_{i1} &= Cov[\cos(2\pi f_i t_i), \cos(2\pi \gamma_j(t_i - \delta_j))] \quad \text{and} \\ \phi_{i2} &= Cov[\sin(2\pi f_i t_i), \cos(2\pi \gamma_j(t_i - \delta_j))] \end{aligned} \quad (3)$$

and are used to transform the abundance coefficients from the time domain to the FFT domain. Typically the data are zero filled before implementing the FFT to improve local resolution. Assuming n is a power of two, if the data are zero filled d times then $(2^d - 1)n$ zeros are placed at the end of the transient and the corresponding frequencies are $f_i = i/(2^d)$ where $i = 0, 1, \dots, (n2^d) - 1$. For example, if the data are zero filled $d = 2$ times, then $3n$ zeros are placed at the end of the transient and the frequencies are equal to $f = 0, 1/4, 1/2, 3/4, 1, \dots, n-1$.

Figure 1a displays the FFT of a FT-ICR mass spectrum. Since the cyclotron frequency of a species in the ICR cell is inversely proportional to the mass-to-charge (m/z) ratio of the ion in question, each peak (vertical line) in the figure corresponds to a distinct m/z ratio. Figure 1b expands the 219.5 to 220.5 kHz region, which displays one isotopic cluster of a doubly charged species. In Figure 1b, A denotes the monoisotopic peak and adjacent peaks in the figure differ in mass by one neutron, corresponding to the replacement of one ^{12}C by one ^{13}C isotope, or one $^1\text{H} \rightarrow ^2\text{H}$, $^{14}\text{N} \rightarrow ^{15}\text{N}$, etc. Note that because of the inverse relationship between frequency and mass, the peak with the largest mass, $A+4$, corresponds to the smallest frequency.

The FFT transform of each of the c ions ($j=1, \dots, c$) in equation 1, ignoring the phase information, is the *sinc* function displayed in Figure 2 [9]. The overall FFT is the sum of *sinc* functions,

$$Y(f_i) = \sum_{j=1}^c \alpha_{j(s)} \left| \frac{\sin[\pi(f_i - \tilde{\gamma}_{j(s)})]}{\pi(f_i - \tilde{\gamma}_{j(s)})} \right| + \varepsilon(f_i) \quad (4)$$

where $Y(f_i) = \sqrt{\phi_{i1}^2 + \phi_{i2}^2}$ denotes the FFT abundance at frequency f_i , $\alpha_{j(s)}$ denotes the abundance of ion j nested within isotope cluster s and $\tilde{\gamma}_{j(s)}$ denotes the frequency of ion j nested within cluster s . Note that the f_i returned from the FFT in equation 2 are in terms of total cycles and the γ_j in equation 1 are in cycles-per-second, so there is minor conversion ($\tilde{\gamma}_{j(s)} = \frac{n\gamma_j}{r}$) in transferring from one to the other, where r denotes the number of sample points collected per second. The nested notation $j(s)$, used from here on, is incorporated due to the fact that the FFT signal contains isotopic clusters

and each cluster is made up of multiple peaks (ions). For a point of reference, Figure 2 displays the fit of the *sinc* function against un-windowed observed FFT data with $d=2$ zero fills.

If the true frequency $\tilde{\gamma}_{j(s)}$ for a particular ion were an integer value and the FFT were not zero filled, then the FFT data would consist of a single non-zero peak at $\tilde{\gamma}_{j(s)}$ since the *sinc* function is equal to zero at integer spacings from the true frequency of the ion. Savitski et al. [10] used this fact to fit the *sinc* function to un-windowed and non zero-filled FFT data in a particular way. By pre-multiplying the transient data by $\exp(-i\nu)$ before applying the FFT, one obtains the discrete Fourier transform for the set of cyclotron frequencies $h+\nu$, $h = 0, 1, \dots, n-1$. Savitski and colleagues proposed doing this for a large number of offsets ν , and then ultimately chose the $f_i = h+\nu$ from that image that is closest to being a single isolated peak. Because windowing broadens peaks and compromises mass resolution [10], we also propose performing peak detection on un-windowed FFT data. We use a more computationally efficient method of directly estimating $\tilde{\gamma}_{j(s)}$ through the use of a *sinc* function.

Mass Calibration as a Function of Mass

The rational frequencies $\tilde{\gamma}_{j(s)}$ for each ion are of course not of specific interest, but instead serve as a marker of the mass. Zhang et al. [11] showed that frequency is inversely proportional to m/z ,

$$f = \beta_0 + \beta_1 \left(\frac{z}{m + zm_{charge}} \right) + \epsilon \quad (5)$$

where f refers to the cyclotron frequency of an ion as obtained from a FFT, β_0 is an unknown intercept parameter, β_1 is an unknown slope parameter that is proportional to the magnetic-field strength, m is the mass of an ion, z is the number of charges, m_{charge} is the mass of the charge carrier (~ 1.0073 Daltons) and ε is the error term. The unknown calibration parameters β_0 and β_1 are most commonly estimated using either internal or external calibrants. Typically, the mass of the charge carrier is included in the mass of an ion and equation 5 is simplified to $f = \beta_0 + \beta_1(z/m) + \varepsilon$. However, we will always separate the mass of the charge carrier from the mass of the ion.

A more accurate form of equation 5 was recently proposed [12],

$$f = \beta_0 + \beta_1 \left(\frac{z}{m + zm_{charge}} \right) + \beta_2 A_{Total} + \beta_3 A_{Ion} + \varepsilon, \quad (6)$$

which incorporates additive effects for ion abundance (A_{Ion}) and the total charge content in the ICR cell (A_{Total}). McIver et al. [13] showed experimentally that the total charge content in the ICR cell for all ions exerts a linear decrease in all cyclotron frequencies. Muddiman and Oberg [12] later proposed the transformation defined in equation 6.

2. METHODS

We assume that the errors $\varepsilon(t_i)$ in the raw data, i.e. the digitized transient signal represented in equation 1, are uncorrelated and have constant variance over the time of acquisition. Both are reasonable assumptions given our knowledge of the physical

system that creates the transient signal. Because the discrete Fourier transformation corresponds to multiplication of the data by an orthogonal matrix, un-windowed FFT data are also uncorrelated and equivariant. However, this is not true of windowed FFT data, which have a more complex correlation structure. This is also not true of data after transformation from frequency to m/z scale using any of equations 5-6. Thus, from a statistical standpoint it is advantageous to do data fitting using only the transient data, or the un-windowed FFT data, as the covariance structure remains unaltered. For this reason, the proposed procedure utilizes un-windowed FFT data for mass calibration and peak detection purposes.

Mass Calibration as a Function of Frequency

In the frequency domain we propose that the set of peaks corresponding to a molecular species can be jointly fit. As such, it is not necessary to transform to the m/z domain for peak detection. Specifically, equation 5 can be transformed such that the calibration equation is a function of f_0 , the frequency of one of the peaks that comprise a species, instead of mass. In doing so, it is assumed that the resolving power of the mass analyzer (e.g., FT-ICR) is such that isotopic clusters are resolved and at least a subset of the abundant species exists at multiple charge states, as is the case with ESI. Provided these assumptions hold, consider a single isotopic cluster that has charge z_0 and denotes a molecular species of interest. Let f_0 be the frequency of any identified peak (ion) in the isotopic cluster, m_0 the corresponding mass, and k_0 the number of neutrons present in that ion. Assume that, at least locally, equation 5 holds.

Then, the frequency $f_{k,z}$ of another ion in a cluster, one with k neutrons and charge z , is

$$f_{k,z} = \beta_0 + \beta_1 \left(\frac{z}{m_0 + zm_{charge} + (k - k_0)m_{neutron}} \right) + \epsilon_{k,z}, \quad (7)$$

where $m_{neutron}$ represents the mass difference between ^{12}C and ^{13}C (~ 1.0034 Daltons).

Equation 7 implies that if the mass (m_0) for any single peak in an isotope cluster is known and the corresponding frequency location and charge can be estimated from the spectrum, then the expected frequency locations for all other peaks in the cluster is a simple calculation because the peaks differ only by the mass of a neutron ($m_{neutron}$).

Reformulating equation 7 to be a function of f_0 instead of m_0 results in

$$f_{k,z} = \beta_0 + \left(\frac{z}{z_0} \right) \left(\frac{1}{f_0 - \beta_0} + \frac{(z - z_0)m_{charge} + (k - k_0)m_{neutron}}{\beta_1 z_0} \right)^{-1} + \epsilon_{k,z}. \quad (8)$$

Briefly, to obtain equation 8 we have simply transformed a well-understood linear calibration transformation that is a function of mass (equation 5) into a nonlinear calibration transformation that is a function of frequency. The motivation for obtaining an equation that is a function of frequency was to be able to perform mass calibration without calibration information. Thus, equation 8 utilizes frequency information from individual ions that comprise isotopic distributions (determined by the mass of a neutron) in addition to molecular species that exist at multiple charge states (determined by the mass of the charge carrier) for mass calibration. To verify the utility of the calibration parameters in this setting, we examined the resulting frequency shifts that were associated with changing each of parameter estimates (Table 1). Table 1 shows

that changes in β_1 affect overall peak spacing while changes in β_0 primarily result in charge-state frequency shifts (shifts due to space-charge effects). This implies that β_1 can be estimated from molecules that exist at only a single charge state, whereas β_0 requires the existence of molecules that are present at multiple charge states. Lastly, note that z_0 would be completely aliased with β_1 were it not known that z_0 is an integer ≥ 1 .

To estimate the calibration parameters in equation 8 requires the use of a numerical procedure (e.g., Nelder-Mead Simplex, Gauss-Newton, Newton-Raphson) [14]. With respect to estimation, equation 8 is less problematic than using equation 7. Equation 7 is a function of mass (m_0); yet, mass is a function of the calibration parameters. As a consequence of the poorly parameterized model, the numerical procedure will have difficulties converging to the correct parameter estimates. For poorly parameterized models, Seber and Wild [14] state that changing the parameterization of the model can have a marked effect on the performance of the algorithm. Thus, equation 8 is a re-parameterized version of the model, which is a function of frequency instead of mass.

Provided that there is justification, additional effects are easily incorporated into the reformulated calibration equation defined in equation 8. For example, the linear calibration equation defined in equation 6 is also easily reformulated to be a function of f_0 instead of m_0 ,

$$f_{k,z} = \beta_0 + \left(\frac{z}{z_0} \right) \left(\frac{1}{f_0 - \beta_0 - \beta_2 A_{Total} - \beta_3 A_{f_{k_0, z_0}}} + \frac{(z - z_0)m_{charge} + (k - k_0)m_{neutron}}{\beta_1 z_0} \right)^{-1} + \beta_2 A_{Total} + \beta_3 A_{f_{k,z}} + \varepsilon_{k,z} \quad (9)$$

where $A_{f_{k,z}}$ denotes the abundance of peak $f_{k,z}$, and, z_0 and k_0 is the charge and the number of neutrons for peak denoted as f_0 .

Modeling Isotopic Distributions

To model isotopic distributions, a template is built based on *averagine*, an average amino acid that was developed specifically for modeling isotopic distributions [15]. Table 2 presents the isotopic decomposition, based on *averagine*, for a molecular species of mass 906.6723 Daltons. The expected number of extra neutrons for an ion is a function of the number of isotopes present and the corresponding natural abundance. Because the sum of Poisson densities is itself a Poisson density, the joint isotopic distribution for all single (+1) isotopes follows a Poisson distribution with shape parameter $\lambda_{(+1)}$, where $\lambda_{(+1)}$ denotes the expected number of extra neutrons for all +1 isotopes. Likewise, the joint isotopic distribution for the +2 isotopes is two times a Poisson random variable with shape parameter $\lambda_{(+2)}$ and the joint distribution for the +3 isotopes is three times a Poisson random variable with shape parameter $\lambda_{(+3)}$. The theoretical isotopic distribution for the combination of +1, +2, and +3 isotopes is the convolution of the three aforementioned Poisson random variables.

Rockwood and Van Orden [16] developed an algorithm to compute the expected isotopic distribution based on a Fourier convolution. For the data presented in this paper, we used an approach that is less computationally intensive. Table 2 suggests that the isotopic distribution is dominated by single isotopes, and in particular, by ^{13}C . Thus, the joint distribution is close to a single Poisson density with shape parameter $\lambda_m = \lambda_{(+1)} + 2\lambda_{(+2)} + 3\lambda_{(+3)}$. Figure 3 displays observed FFT data (vertical lines), the theoretical fits (x) based on the convolution, and an approximation (o) using the Poisson approximation. The Poisson approximation has somewhat too narrowed a distribution in comparison to the theoretical fit; however, the observed data is even narrower. For the data presented, low-abundant ions are not as well detected in the ICR cell as high-abundant ions and thus a single Poisson distribution provides an adequate fit to the data. In fact, λ_m is a nuisance parameter in the proposed algorithm and adjusts to the local variations in ^{13}C frequencies for each molecule. Only a reasonable starting estimate is needed for λ_m in the algorithm, as the final estimate is obtained from the fitting routine that provides the best fit to the observed data for each molecular species.

Simultaneous Mass-Calibration & Peak-Detection Algorithm

By taking advantage of the reformulated calibration transformations described in equations 8-9, mass calibration and peak detection are approached simultaneously in the frequency domain. For an identified peak, f_0 , that denotes an ion that belongs to cluster s_0 of charge z_0 , there are four realized parameters in the algorithm: (i) space-charge parameter β_0 , (ii) peak-spacing parameter β_1 , (iii) Poisson parameter λ_m , and

(iv) peak abundances $\alpha_{j(s)}$. It turns out that estimation of the peak abundances requires only a single parameter per charge state and is a simple calculation given the other three parameters. With regard to choosing an f_0 , utilizing a top-down approach where the most abundant peak is chosen is sufficient and computationally appealing. However, the algorithm only requires that a peak be identified and thus any peak in the cluster will suffice.

The proposed algorithm consists of a step-wise process that entails three key steps. First, Part 1 chooses a peak, f_0 , and subsequently determines the charge state of the isotopic cluster that contains the identified peak f_0 . Second, Part 2 locates all other peaks in the corresponding isotopic cluster as well as peaks for other charge states of the same molecular species. Using *averagine* to estimate the isotopic distribution, Part 2 estimates peak heights and a set of specie-dependent calibration parameters. Note that by utilizing equation 8, calibrants are not necessary to obtain estimates of the calibration parameters. Remember that β_1 is estimated using peak spacing information and β_0 is estimated from charge spacing information. Parts 1-2 are repeated until all molecular species are located, or similarly, until all peaks that have abundances larger than some pre-determined signal-to-noise threshold have been accounted for. Lastly, Part 3 estimates an overall set of calibration parameters using information from all molecules. The algorithm is described in detail in the Appendix.

3. APPLICATION

The proposed algorithm was evaluated in terms of mass-measurement precision (MMP) and mass-measurement accuracy (MMA) using a series of mass spectra of ammonium-adducted polypropylene glycol that were initially presented in Muddiman and Oberg [12]. Ten analysis samples were prepared (though only nine were utilized), each with varying amount of analyte (i.e., a range of A_{Total}). Muddiman and Oberg reported the theoretical mass for nine of the singly-charged oligomers and these nine oligomers were used to assess the performance of the proposed algorithm. Herein, all analyses were done on un-windowed FFT data that were zero-filled twice.

Equation 8 was used to detect the largest $L=40$ molecular species in each spectrum and the Nelder-Mead Simplex was used to estimate the calibration parameters for each molecular species. Subsequently, taking into account previous calibration work [12-13], an overall set of calibration parameters were estimated using equation 9 that accounts for the fact that the total charge content in the ICR cell exerts a linear decrease in all cyclotron frequencies. The total charge in the ICR cell, A_{Total} , was computed as the sum of all FFT abundances in the spectrum that existed within appropriate ranges of the data (arbitrary units). Utilizing equation 9 resulted in a median MMP of 1.80 parts-per-million (ppm) and a minimum and maximum MMP of 0.01 and 32.98, respectively.

MMP was defined as the standardized difference between the measured mass (\hat{m}) and the average measured mass across the nine spectra (\bar{m}) for each molecular species

($MMP = \left| \frac{\hat{m} - \bar{m}}{\bar{m}} \right| \cdot 10^6$). To estimate the average measured mass across the nine spectra,

species were matched across spectra using complete-linkage clustering [17].

As evidenced by the high MMP, the proposed algorithm has the ability to match equivalent species across spectra, which is essential in studies where the objective is to estimate differential expression across treatment or disease groups. After obtaining a list of molecular species that are differentially expressed across two groups of interest, the second goal is identification, which requires high MMA. MMA was evaluated using the nine known oligomers that were recorded in Muddiman and Oberg [12]. The median absolute MMA for the nine oligomers was 62.51 ppm, with corresponding minimum and maximum MMA of 51.04 and 98.47, respectively. MMA was defined as the standardized difference between the measured mass and the theoretically known mass (m_T) for each of the nine oligomers ($MMA = \left| \frac{\hat{m} - m_T}{m_T} \right| \cdot 10^6$).

Unfortunately, MMA of 62.51 ppm is not adequate for comprehensive identification. Figure 4a displays the mass-measurement errors ($\hat{m} - m_T$) versus the theoretically known masses (m_T). The numbers in Figures 4a and 4b denote spectra, such that 1 denotes the spectrum with the smallest total charge (A_{Total}) and 9 denotes the spectrum with the largest total charge. The negative slope implies that the measurement errors are primarily a function of mass; however, a smaller component of the error is due to total charge. Because the magnetic-field parameter is the only parameter that is a function of mass (similarly, frequency), this implies that the magnetic-field parameter $\hat{\beta}_1$ is not estimated accurately using the proposed approach. In fact, the sensitivity analysis provided in Table 1 provides evidence that obtaining an accurate estimate of

the magnetic-field parameter is difficult when using the proposed nonlinear model that is a function of frequency. However, this is not true for calibration models that are a function of mass.

To obtain an accurate estimate of the magnetic-field effect, we investigated the benefits of a single external calibration sample. Ideally, we would prefer not to include internal calibrants, particularly in biomarker-discovery experiments, due to the possibility of contaminating potential biomarkers of interest and thus external calibrants are favored. Additionally, because the magnetic-field strength remains relatively constant suggests that an external calibration sample only needs to be run as often as the magnetic-field strength is expected to vary, e.g., on a weekly or monthly basis. To verify the benefits of a single external calibration sample, we assigned the first sample that was run by Muddiman and Oberg [12] to be the external calibration sample and the remaining eight spectra were used to estimate MMA. Using the nine known oligomers in our external calibration sample, we used a simple linear regression model to estimate the calibration parameters. The dependent variable was the observed cyclotron frequencies and the independent variables were the known m/z and the observed ion abundances. This resulted in the following regression model,

$$f = -4.3191 + 150,005,961(z/m) + 0.0321A_{ion}.$$

Note that we did not include an effect for total charge because it is completely confounded with the intercept term. The parameter estimates above do not exactly match the estimates found in Table 2 (spectrum #1) of Muddiman and Oberg [12] for two reasons: (1) we model frequency in total cycles, whereas [12] uses cycles-per-

second, and (2) the observed frequencies and ion abundances were measurements and thus will be slightly different across estimation procedures.

An overall set of calibration parameters was estimated a second time using equation 9 and this time the magnetic-field parameter was held constant using the estimate from the external calibration sample ($\hat{\beta}_1 = 150,005,961$). This resulted in a median absolute MMA of 2.80 ppm, an improvement from 62.51 ppm. Figure 4b displays the updated mass-measurement errors versus the theoretically known masses. The negative slope that was present in Figure 4a was eliminated; however, a small error due to total charge still exists. The error is a result of A_{Total} being a measured value and thus an imprecise variable. Measurement error is often disregarded; however, when the goal is to obtain parts-per-million or even parts-per-billion MMA measurement error can cause considerable error.

4. DISCUSSION

We have described a procedure that combines mass calibration and peak detection into a single algorithm and is able to achieve < 2 ppm mass-measurement precision and < 3 ppm mass-measurement accuracy with only a single external calibration sample. This is achieved by changing the parameterization of a well-understood linear calibration equation that is a function of mass into a nonlinear equation that is a function of frequency. Performing fitting routines in the un-windowed frequency domain is appealing because peak shape and peak width are known, thus requiring fewer and more stable parameters in comparison to the m/z domain. Furthermore, un-windowed

FFT data are uncorrelated and equivariant, which are important qualities with respect to model fitting.

Moreover, because we have proposed a modeling approach, familiar model-fitting diagnostics are encouraged and simple to apply. For example, residual diagnostics should be incorporated to assess model fit and to determine if additional effects are necessary to improve the MMP and MMA. Similarly, well known statistical tests can be incorporated to formally test the significance of effects in the nonlinear model. Thus, we have demonstrated that mathematical modeling requires fewer parameters in the unwindowed frequency domain and allows for mass calibration to be combined with the fitting routine.

ACKNOWLEDGMENTS

Funding was provided to Eckel-Passow by NCI grant R25 CA92049 and to Oberg by the Fraternal Order of Eagles Cancer Research Fund. This work was also partially supported by the National Institutes of Health (CA105295-1), the W.M. Keck Foundation, and the Mayo Clinic College of Medicine.

REFERENCES

1. Diamandis, E.P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Molecular & Cellular Proteomics*, 3(4), 367-378.
2. Anderson, N.L. and Anderson, N.G. (2002) The human plasma proteome. *Molecular & Cellular Proteomics*, 1(11), 845-867.
3. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246 (4926), 64-71.
4. Yamashita, M. and Fenn, J. B. (1984) Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.*, 88(20), 4451-4459.
5. Henry, K. D., Williams, E. R., Wang, B. H., McLafferty, F. W., Shabanowitz, J. and Hunt, D. F. (1989) Fourier-transform mass spectrometry of large molecules by electrospray ionization. *Proc. Natl. Acad. Sci.*, 86(23), 9075-9078.
6. Comisarow, M. B. and Marshall, A. G. (1974) Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.*, 25(2), 282-283.
7. Horn, D.M., Zubarev, R.A. and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom*, 11, 320-332.
8. Bruce, J.E., Anderson, G.A., Brands, M.D., Pasa-Tolic, L. and Smith, R.D. (2000) Obtaining more accurate Fourier transform ion cyclotron resonance mass measurements without internal standards using multiply charged ion. *J Am Soc Mass Spectrom*, 11, 416-421.

9. Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews*, 17, 1-35.
10. Savitski, M.M., Ivonin, I.A., Nielsen, M.L., Zubarev, R.A., Tsybin, Y.O. and Hakansson, P. (2004) Shifted-basis technique improves accuracy of peak position determination in Fourier transform mass spectrometry. *J Am Soc Mass Spectrom*, 15, 457-461.
11. Zhang, L-K, Rempel, D., Pramanik, B.N. and Gross, M.L. (2005) Accurate mass measurements by Fourier transform mass spectrometry. *Mass Spectrometry Reviews*, 24, 286-309.
12. Muddiman, D.C. and Oberg, A.L. (2005) Statistical evaluation of internal and external mass calibration laws utilized in Fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry*, 77, 2406-2414.
13. McIver, J.R.T., Hunter, R.L. and Bowers, W.D. (1985) Coupling a quadrupole mass spectrometer and a Fourier transform mass spectrometer. *Int. J. Mass Spectrom. Ion Proc.*, 64, 67-77.
14. Seber G.A.F. and Wild C.J. (1989) *Nonlinear Regression*. Wiley: New York, pp. 91-126.
15. Senko, M.W., Beu, S.C. and McLafferty, F.W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J Am Soc Mass Spectrom*, 6, 229-233.
16. Rockwood, A.L. and Van Orden S.L. (1996) Ultrahigh-Speed Calculation of Isotope Distributions. *Analytical Chemistry*, 68, 2027-2030.

17. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A. and Le, Q. (2004) Sample classification from protein mass spectrometry by “peak probability contrasts”. *Bioinformatics*, 20, 3034-3044.

APPENDIX: Algorithm

PART 1: Charge-State Determination

For a single spectrum, the algorithm to determine z_0 is as follows:

1. To begin, consider equation 8 (later, we will show how to incorporate the improved version shown in equation 9) throughout Part 1. Thus, obtain starting estimates for the calibration parameters β_0 and β_1 , either from previous calibration work, or if available, from the manufacturer's software.
2. Find the frequency location f_0 of the most abundant peak in the FFT signal that is assumed to belong to an isotopic cluster and whose abundance is larger than signal-to-noise threshold. Note that any procedure that is capable of locating a single peak in the spectrum will suffice and the most abundant peak is simply used out of convenience here.
3. For each plausible charge z do the following:
 - a. Estimate m_0 using the starting estimates for the calibration parameters.
 - b. As a starting estimate, set the Poisson parameter equal to $\lambda_m = m_0(\lambda_1 + 2\lambda_2 + 3\lambda_3)$ using the abundance values in Table 2 to estimate λ_1 , λ_2 and λ_3 .
 - c. Approximate cluster s_0 using a Poisson density with mean λ_m and retain all peaks where the Poisson probability is larger than a scaled version of the signal-to-noise threshold. Denote the estimated Poisson probabilities for each of the retained j ions as $\hat{\alpha}_{j(s_0)}^P$.

- d. Estimate $f_{k,z}$ for each retained peak, where z equals the current plausible charge.
 - e. Estimate a predicted FFT signal $\hat{Y}_z^*(f_i)$ for charge z using the *sinc* function in equation 4, where $\alpha_{j(s)} = \hat{\alpha}_{j(s_0)}^P$ are the Poisson probabilities from step (c) above and $\tilde{y}_{j(s)}$ are the $f_{k,z}$ from step (d). Note that the Poisson probabilities provide only the general shape of cluster s_0 . To obtain correct peak heights $\hat{Y}_z^*(f_i)$ is multiplied by a constant; however, only the general shape is required at this stage and thus the derivation of the constant (peak-height parameter) is addressed in the next section.
 - f. Calculate the correlation of the observed FFT abundances $Y(f_i)$ with the predicted abundances $\hat{Y}_z^*(f_i)$.
4. Retain the charge z that produces the largest correlation, here after referred to as z_0 .
 5. Lastly, the algorithm determines if the current molecule exists at any other charge states. All other charge states that produce $f_{k,z}$ in the observed cyclotron frequency range and that have abundances larger than the signal-to-noise threshold are retained. Note that clusters denoting the same molecular species have the same Poisson mean λ_m .

PART 2: Specie-Dependent Mass Calibration & Peak Detection

The simultaneous mass-calibration and peak-detection algorithm estimates β_0 , β_1 , λ_m , and the overall peak-height parameter utilizing a nonlinear optimizer that minimizes mean-squared error (MSE). Switching to matrix notation, let \mathbf{Y} be the $n \times 1$ vector of observed FFT abundances and $\hat{\mathbf{Y}}^*$ the $n \times p$ matrix of predicted signals, such that the p columns of $\hat{\mathbf{Y}}^*$ denote the predicted abundances for each of the p charge states the molecule exists at. Then,

$$MSE = (\mathbf{Y} - \hat{\boldsymbol{\theta}} \hat{\mathbf{Y}}^*)^T (\mathbf{Y} - \hat{\boldsymbol{\theta}} \hat{\mathbf{Y}}^*)$$

where $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{Y}}^{*T} \hat{\mathbf{Y}}^*)^{-1} \hat{\mathbf{Y}}^{*T} \mathbf{Y}$ is the vector of peak-height parameters for each of the associated charge states.

The algorithm as described is essentially a three-step procedure; however, the first two parts could be combined. Theoretically, z_0 could be estimated in part 2 from the nonlinear optimizer as well and simply rounded to the nearest integer. However, the nonlinear optimizer will encounter local maxima in determining z_0 and thus will have difficulty converging. Furthermore, nonlinear optimizers are computationally time consuming and hence it is more efficient to determine z_0 first and then set it as a fixed parameter in the nonlinear fitting routine.

Lastly, the mass-calibration and peak-detection algorithm involves a fitting routine and as a result, errors are expected and residual peaks are inherent. To reduce the chance of obtaining residual peaks, we suggest the following solution. Choose two thresholds

τ_1 and τ_2 , where τ_1 denotes machine uncertainty and τ_2 denotes uncertainty in the peak-detection algorithm. Let $\hat{\mathbf{Y}}_{cum} = \sum_r \hat{\boldsymbol{\theta}}_r \hat{\mathbf{Y}}_r^*$ denote the $n \times 1$ vector of accumulative predicted signals across the detected molecular species. Then, in step 2 of the charge-state determination, find the frequency location f_0 of the most abundant peak subject to the following constraints, $|\mathbf{Y} - \hat{\mathbf{Y}}_{cum}| > \tau_1$ and $\left| \frac{\mathbf{Y} - \hat{\mathbf{Y}}_{cum}}{\hat{\mathbf{Y}}_{cum}} \right| > \tau_2$.

PART 3: Overall Set of Calibration Parameters

The ultimate goal is to estimate a single set of calibration parameters across all molecular species from which to obtain high mass-measurement accuracy. An overall set of calibration parameters are obtained using the most abundant L molecular species in every spectrum. To do so, parts 1 and 2 as described above are initially applied to only the L largest species in each spectrum, which includes all charge states that correspond to the L species. Equation 9 is used to estimate a single set of calibration parameters and the estimated frequencies are assumed to be the “true” frequencies and are included as the dependent variable in equation 9. Peak locations and abundances from these largest species in each spectrum will be sufficient to estimate an overall set of calibration parameters provided that they represent an adequate range of frequencies, charge states and abundances. After estimating a single set of calibration parameters, the simultaneous mass-calibration and peak-detection algorithm is implemented once again on all spectra to identify all species that are above the signal-to-noise threshold; however, this time setting the calibration parameters as fixed values in the nonlinear fitting routine. As a result, in this phase the only realized

parameters in the nonlinear routine are the peak-shape parameter and a peak-height parameter for each charge state.

Table 1: A sensitivity analysis of the calibration parameters in equation 8. The data presented represents a hypothetical molecular species that exists at two charges states ($z=1,2$) and each charge state is comprised of four isotopic peaks ($k=0,1,2,3$). The bolded column denotes the reference column, where $f_{k,z}$ was calculated using $f_0 = 10^5$, $z_0 = 1$, $k_0 = 0$, $\beta_0 = -3$, and $\beta_1 = 1.45 \times 10^8$. The objective was to verify the utility of the calibration parameters. The first column following the reference column reflects a 100% increase in β_0 and a comparison with the reference columns shows that changes in β_0 primarily result in charge-state frequency shifts. The second column following the reference column reflects a 0.7% increase in β_1 and a comparison with the reference column shows that changes in β_1 result in overall peak spacing shifts.

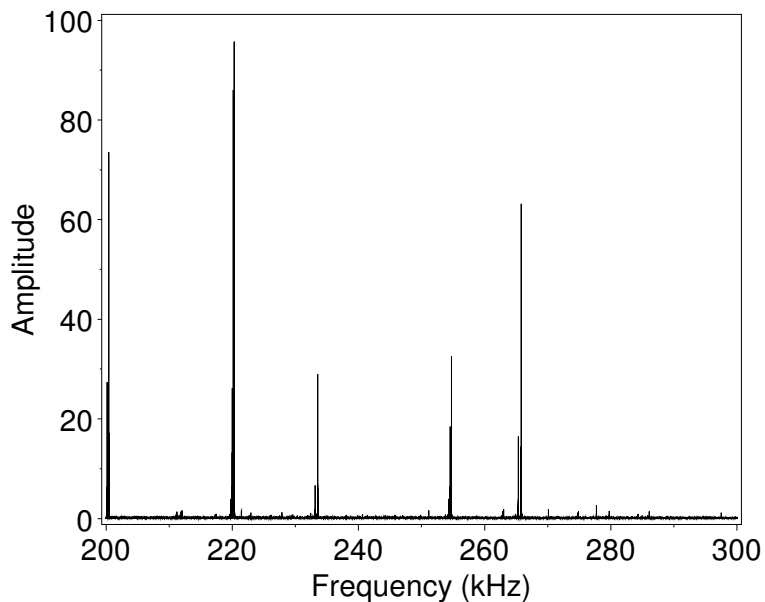
z	k	$f_{k,z} - f_0$	$(f_{k,z} \beta_0 = -6) - f_0$	$(f_{k,z} \beta_1 = 1.46 \cdot 10^8) - f_0$
1	0	0	0	0
1	1	-69.16	-69.16	-68.68
1	2	-138.22	-138.23	-137.27
1	3	-207.18	-207.19	-205.77
2	0	99,864.15	99,867.14	99,865.10
2	1	99,726.03	99,729.01	99,727.92
2	2	99,588.10	99,591.07	99,590.94
2	3	99,450.36	99,453.33	99,454.13

Table 2: Expected number of isotopes, natural abundances, and expected number of extra neutrons for a protein of mass 906.6723 Daltons (oligomer 16_A in Muddiman and Oberg [12]). The abundance distributions are derived from Senko et al. [15] using *averagine*, an average amino acid that is defined to have a molecular formula of C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417} and an average molecular mass of 111.1254 Daltons.

Element	Stable Isotope	Number	Natural Abundance	Expected number of + <i>k</i> isotopes ($\lambda_{(+k)}$) (= Number × Natural Abundance)
² H	+1	63.3	0.000115	0.0073
¹³ C	+1	40.3	0.010700	0.4312
¹⁵ N	+1	11.1	0.003680	0.0408
¹⁷ O	+1	12.1	0.000380	0.0046
³³ S	+1	0.4	0.007600	0.0030
				$\lambda_{(+1)} = \mathbf{0.4869}$
¹⁸ O	+2	12.1	0.002050	0.0248
³⁴ S	+2	0.4	0.042900	0.0172
				$\lambda_{(+2)} = \mathbf{0.0420}$
³⁵ S	+3	0.4	0.000200	0.0001
				$\lambda_{(+3)} = \mathbf{0.0001}$

Figure 1: FT-ICR spectrum for (a) the 200 to 300 kHz range and (b) expands the 219.5 to 220.5 kHz range. The isotopic cluster in (b) represents an ionized-molecular specie of charge $z = 2$, where A denotes the monoisotopic mass, $A+1$ has one ^{12}C replaced by one ^{13}C isotope, $A+2$ has two ^{12}C replaced by two ^{13}C isotopes, etc.

(a)



(b)

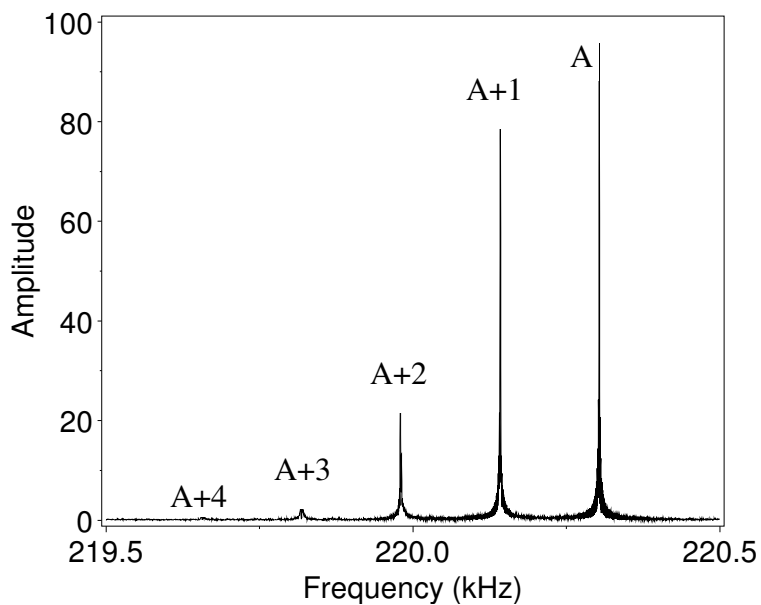


Figure 2: An expanded view of the largest peak in Figure 1b. Circles represent the observed DFT frequency data (kHz) and the line represents the fitted *sinc* function from equation 4. The frequency of the largest peak is denoted as $\tilde{\gamma}_{j(s)}$ in equation 4.

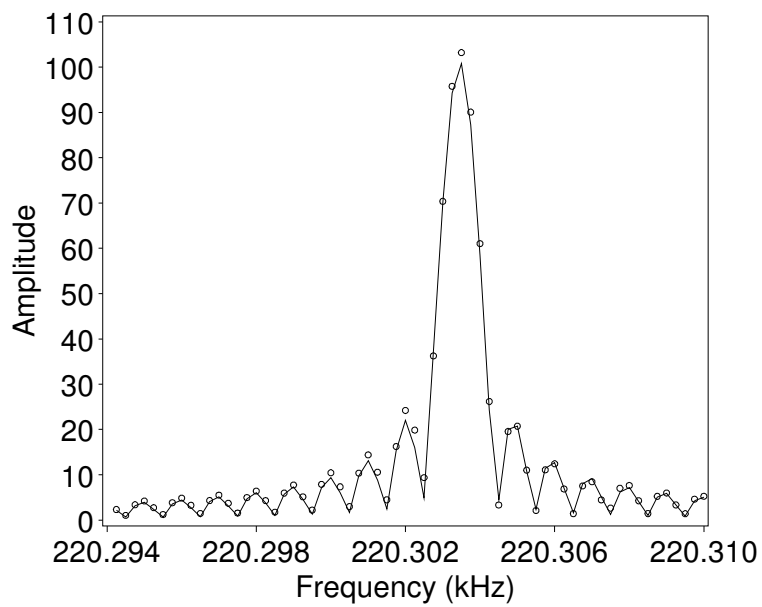


Figure 3: Isotopic distribution for isotope ^{15}A in Muddiman and Oberg [12], which has a mass of 906.6723 Daltons, where vertical lines represent the observed distribution, (\times) represents the theoretical fit that is a convolution of Poisson distributions, and (o) represents the single Poisson approximation.

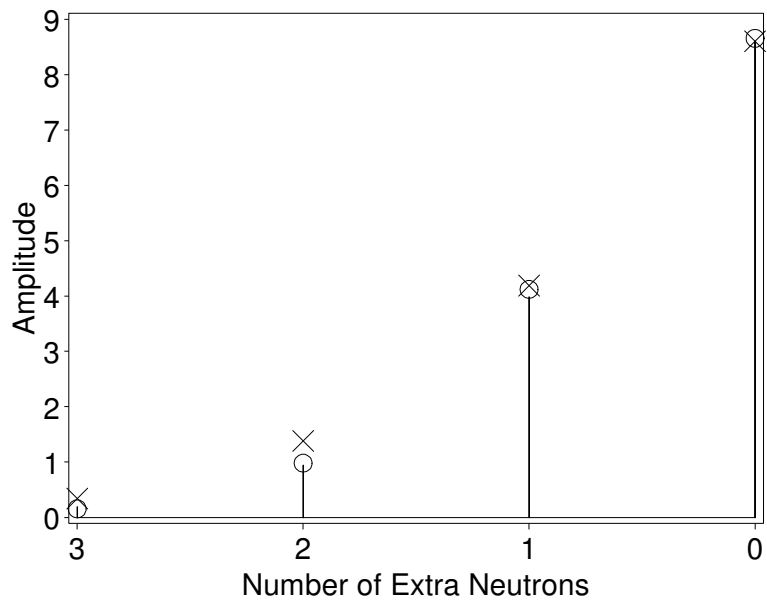
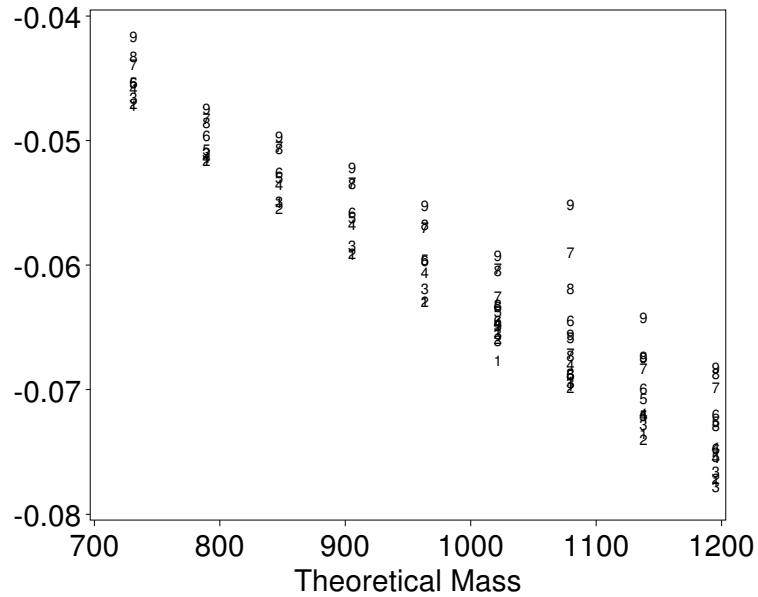


Figure 4: Mass-measurement error ($\hat{m} - m_T$) versus the theoretical mass (m_T) after applying the simultaneous mass-calibration and peak-detection algorithm (a) without incorporating calibrant information and (b) using a single external calibration sample. The numbers in the figures denote spectra, where 1 denotes the spectrum with the smallest total charge (i.e., A_{Total}) and 9 denotes the spectrum with the largest total charge. Number 2 does not exist in Figure 2a because it was used as the external calibration sample. Furthermore, a number will appear more than once if the species was present at more than one charge state in the corresponding spectrum. In these data, species were present at charge states of one or two.

(a)



(b)

