Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra

David L. Tabb,[†] and Manesh B. Shah
Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

Michael Brad Strader,[‡] Heather M. Connelly,* Robert L. Hettich, and Gregory B. Hurst
Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

We report an automated method for determining charge states from high-resolution mass spectra. Fourier transforms of isotope packets from high-resolution mass spectra are compared to Fourier transforms of modeled isotopic peak packets for a range of charge states. The charge state for the experimental ion packet is determined by the model isotope packet that yields the best match in the comparison of the Fourier transforms. This strategy is demonstrated for determining peptide ion charge states from "zoom scan" data from a linear quadrupole ion trap mass spectrometer, enabling the subsequent automated identification of singly- through quadruply-charged peptide ions, while reducing the numbers of conflicting identifications from ambiguous charge state assignments. We also apply this technique to determine the charges of intact protein ions from LC-FTICR data, demonstrating that it is more sensitive under these experimental conditions than two existing algorithms. The strategy outlined in this paper should be generally applicable to mass spectra obtained from any instrument capable of isotopic resolution. (J Am Soc Mass Spectrom 2006, 17, 903–915) © 2006 American Society for Mass Spectrometry

Determination of charge state of an ion from its mass spectrum is relatively straightforward if the spectrum is sufficiently resolved to distinguish peaks in the isotope packet for the ion. The spacing between the isotope peaks is simply the reciprocal of the charge state. For small numbers of spectra, this charge state determination can be performed manually. However, modern mass spectrometry (MS) instrumentation, especially when interfaced with a chromatographic separation, can yield sufficiently large numbers of spectra to render manual charge state determination impractical.

Recent advances have given us a wider array of MS instrumentation capable of providing isotopically resolved data from which the charge state can be determined. The introduction of slower *m/z* scan speeds allowed higher resolution measurements in "3D" quadrupole ion trap mass spectrometers [1]. The increased ion volume available in recently-introduced linear, or "2D" quadrupole ion trap instruments also allows higher resolution spectra to be obtained via slower scan speeds, but with less signal averaging than required for 3D quadrupole ion trap instruments [2]. With the linear quadrupole ion trap, it is possible to obtain isotope-resolved parent ion spectra, for example, of singly- and multiply-charged peptides that have *m/z* values up to a few thousand. The "Ultra Zoom Scan" feature of the ThermoFinnigan LTQ mass spectrometer provides isotope-resolved spectra over user- or software-selected windows of 10 *m/z* width; the central *m/z* values for these windows can be chosen dynamically by instrument control software, based on the most intense ions observed in a conventional full-scan mass spectrum.

FTICR mass spectrometry provides very high-resolution and accuracy because of the accuracy with which it is possible to measure the frequency of ion cyclotron motion in the Penning trap [3]. This resolving power enables acquisition of mass spectra of electrosprayed intact protein ions with resolved isotopologues

Published online May 19, 2006

Address reprint requests to Dr. G. B. Hurst, Chemical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6131, Oak Ridge, TN 37831-6131, USA. E-mail: hurstgb@ornl.gov

* Also at the Graduate School of Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, 1060 Commerce Park, Oak Ridge, TN 37830.

[†] Present address: Department of Biomedical Informatics, Vanderbilt University Medical Center, 465 21st Ave. S., U9211 MRB III, Nashville, TN 37232-8575.

[‡] Present address: Laboratory of Neurotoxicology, National Institute of Mental Health, Bethesda, MD 20892.

1044-0305/06/$32.00
doi:10.1016/j.jasms.2006.02.003

Received December 2, 2005
Revised January 30, 2006
Accepted February 1, 2006

for charge states up to $z = 10$ and higher [4]. In contrast to quadrupole ion traps, no special scan mode is required to obtain high-resolution mass spectra from the FTICR instrument. However, this performance usually is achieved under direct infusion electrospray conditions in which the analyte concentration and ion detection parameters can be optimized. A more challenging scenario is presented for situations in which these carefully controlled conditions are not possible, such as LC-FTICR-MS measurements. In this case, the signal quality is compromised and the direct measurement and resolution of charge states from intact proteins is much more difficult.

In addition to quadrupole ion trap and FTICR instruments, other classes of mass spectrometers are also capable of producing isotopically-resolved mass spectra, including hybrid time-of-flight and orbitrap instruments.

Fourier Transform of a Model for Isotope-Resolved Mass Spectra

The Fourier transform (FT) provides a means to analyze a signal, obtained as a series of measurements that are a function of a variable x, for components that are periodic (i.e., occur at regular increments of x). This approach has been demonstrated for the analysis of mass spectra of synthetic polymers, where x is the mass-to-charge ratio (m/z), and the periodic component sought is the regular spacing of peaks due to different numbers of monomer units in the polymer [5–7]. The FT has also been applied to determining the spacing of peaks in capillary electropherograms used for DNA sequencing [8].

To illustrate the use of the FT for charge determination of peptide and protein ions, we develop a model for isotope-resolved mass spectra of multiply-charged ions, and describe the FT of this model. The relevant periodic component in the mass spectrum is due to regularly-spaced peaks corresponding to the monoisotopic species (all ^{12}C) plus species containing one or more ^{13}C atoms per ion, with a spacing of $\sim 1/z$ (one u divided by the number of charges on the ion, z); the approximation neglects mass defect and contributions from isotopes of other elements. The spacing of this isotope series can be represented as a sum of delta functions [$\delta(x) = 1$ for $x = 0$; $= 0$ for $x \neq 0$; reference [9], p. 70] spaced at intervals of $1/z$ and offset from integral multiples of $1/z$ by the fractional part (Δ_m) of the actual m/z. This sum of delta functions has been called the "shah" function, defined as $\mathrm{III}(x) = \sum_{n=-\infty}^{\infty} \delta(x - n)$, where n is an integer (reference [9], p. 77); for our example, $III\left(\frac{m}{z}\right) = \sum_{n=-\infty}^{\infty} \delta\left(\frac{m}{z} - \frac{(n+\Delta_m)}{z}\right)$. This function is represented in the upper left part of Figure 1 for $z = 1$ and $z = 2$. The shape of each individual peak in a real mass spectrum is not a delta function, but rather is determined by the mass spectrometer's resolution function, which we will approximate as Gaussian, $G_{res} = e^{-x^2/2\sigma_{res}^2}$, with mean $= 0$ and σ_res consistent with widths of experimentally measured peaks (see Figure 1). By convolving the shah function described above with this Gaussian, we produce a set of peaks, equally spaced by $1/z$, whose widths are consistent with the instrument resolution. The number of isotopologue species contributing to the isotope distribution of a peptide or protein is limited; therefore, the model set of peaks spaced by $1/z$ must be truncated on both ends. Although an asymmetric function would be more accurate, especially for smaller peptides, we choose for simplicity a second Gaussian, $G_{iso} = e^{-\left(x-\frac{m}{z}\right)^2/2\sigma_{iso}^2}$, as the truncating function in our model (see Figure 1). This truncating function, which is centered near the average m/z for the isotope distribution and has a width, σ_iso, approximating the expected width of the isotope distribution for species of that m/z, will, when multiplied by the regularly-spaced set of peaks obtained in the previous step, yield a truncated set of peaks. The model can be summarized as $[\mathrm{III}(m/z)*\mathrm{G}_\mathrm{res}]\cdot\mathrm{G}_\mathrm{iso}$, where the * symbol represents convolution. The two plots at the upper right of Figure 1 illustrate this model for $z = 1$ and $z = 2$.

The Fourier transform provides a tool for determining the frequency of the periodic features (i.e., peaks spaced by $1/z$) in the model mass spectra described above. The lower half of Figure 1 outlines a pictorial approach to the Fourier transform [9] of the components of the model mass spectrum. Two simple concepts guide this approach: (1) Fourier transformation of standard functions yields known counterpart functions in the transform domain; (2) operations combining functions in the x domain have counterparts in the Fourier domain. Although not true for most functions, the FT of the "shah" function is another "shah" function, and the FT of a Gaussian is another Gaussian (reference [9], p. 412). The convolution of two functions in the x domain is a counterpart to multiplication of their transforms in the Fourier domain, and vice versa. For our model of isotope-resolved mass spectra, $[\mathrm{III}(m/z)*\mathrm{G}_\mathrm{res}]\cdot\mathrm{G}_\mathrm{iso}$, the FT will be a convolution of the product of the transforms of $\mathrm{III}(m/z)$ and G_res with the transform of G_iso (see lower half of Figure 1). By the similarity theorem (reference [9], pp 101–104), the FT will exhibit peaks that (1) are spaced by the reciprocal of the spacing of $\mathrm{III}(m/z)$ (see lower left part of Figure 1), (2) have a width determined by the reciprocal of the width of G_iso, and (3) are multiplied by a Gaussian envelope whose width is the reciprocal of G_res. Thus, the FT of the isotopically-resolved mass spectrum from a singly-charged ion will exhibit a series of peaks with the smallest between-peak spacing; the FT of the spectrum from a doubly-charged ion will exhibit a series of peaks spaced by twice the distance of the singly-charged case; the FT of an ion with $z = n$ will exhibit a series of peaks spaced by n times the distance of the singly-charged case. The lower half of Figure 1 shows the transforms of the various components of the model, with the plots at lower right

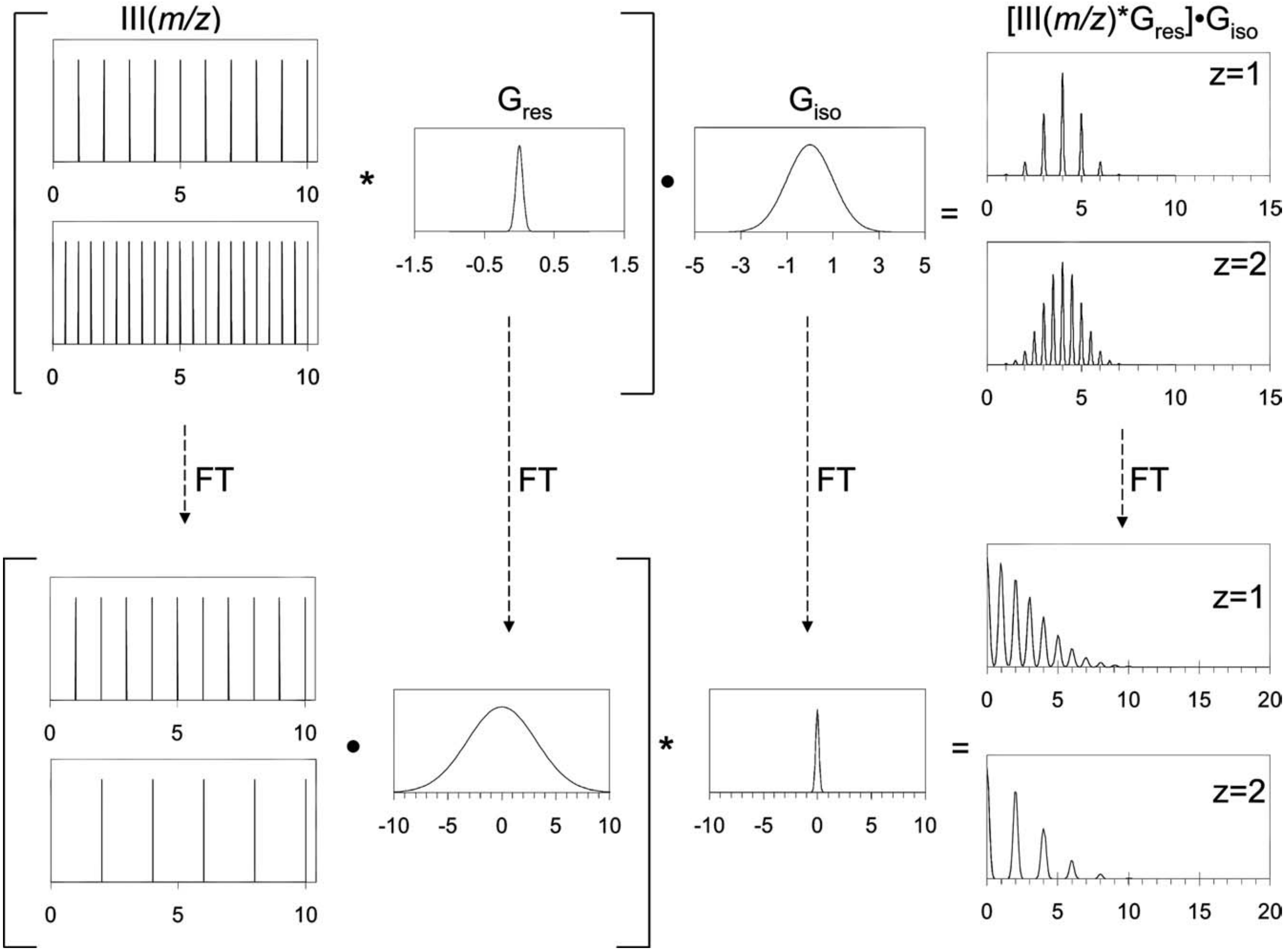

Figure 1. Top half: Model of isotope-resolved mass spectrum of $z = 1$ and $z = 2$ ions. The shah functions at left represent the $1/z$ spacing between peaks. These shah functions are convoluted with a Gaussian function, G_{res}, whose width approximates that of the mass spectrometer instrument function; the convolution operation is denoted by an asterisk. To model the width of the isotopic envelope, the convolution is multiplied by a broader Gaussian function, G_{iso}, yielding the $z = 1$ and $z = 2$ model spectra at upper right. Bottom half: Fourier transforms of the various components of the model from the upper panel. The Fourier transform operation is denoted by the dashed arrows.

showing FTs of model spectra for $z = 1$ and 2. Significantly, the abscissa of the first peak (after the "DC" burst at 0) in the FT of an ion's mass spectrum corresponds to the charge state of that ion.

Charge State Determination for Peptides and Proteins

Peptide-based proteomic mass spectrometry relies upon automated data analysis. The creation of Sequest in 1994 [10] made it possible to match tandem mass spectra to members of peptide databases almost as quickly as the spectra could be acquired. This algorithm and many others comprise a body of bioinformatics that has developed in parallel with advancements in mass spectrometry and analytical chemistry, leading to the phenomenal growth of proteomics during the last decade. Identification of tandem mass spectra can be improved if the charges of the peptide precursor ions can be determined correctly. Identification algorithms generally assume that the peptide charge states they are supplied are correct rather than making a separate determination. For typical operating conditions in quadrupole ion trap mass spectrometers, isotopic resolution is not obtained for parent ions, and charge states are inferred from tandem mass spectra, rather than measured from parent ion spectra. Commonly, multiple charge states are given as possibilities for tandem mass spectra that were produced from peptides for which $z =$ 2 or 3. This results in multiple identifications, some based on an incorrect charge state, from a single spectrum. Identifications based on incorrect charge states increase the number of false positives, thus lowering the overall reliability of the data analysis. In addition, performing multiple identifications for each spectrum increases the time required for identification. Ideally, each tandem mass spectrum would be identified only once, using its correct precursor charge state.

Several efforts have been made to infer charge states of peptide ions correctly on the basis of information in tandem mass spectra obtained from quadrupole ion trap instruments, with limited success. The ZSA algorithm [11] uses eight subtests combined via a neural net to separate spectra by precursor charge into $z = 1$, 2, and 3. The 2 to 3 algorithm [12] uses fragment ion complementarity to discern charge. The PPM-Charger system [13] examines the numbers of putative singly- and doubly-charged fragment ions in the tandem mass spectrum to make this determination. Colinge et al. have also developed algorithms to this purpose [14]. In all of these cases, however, many spectra are left with multiple charge states possible, and these systems may make errors, decreasing the number of correct peptide identifications. Direct measurement of charge from high-resolution parent ion spectra obviates the need for such algorithms based on tandem mass spectra.

Algorithms for charge state determination of intact proteins from high-resolution mass spectra have also been described. Senko et al. described a method that combines a Patterson pattern recognition algorithm with a Fourier transform for charge determination of fragment ions from isotopically-resolved collision-induced dissociation tandem mass spectra of intact proteins in an Fourier transform ion cyclotron resonance (FTICR) mass spectrometer [15]. Zhang and Marshall described an algorithm that uses either isotopically-resolved or lower-resolution spectra from multiple charge states to determine charge states of peptides and proteins from FTICR mass spectra [16].

In the present paper, we explore the use of the Fourier transform approach as a general tool for determining charge states from isotopically-resolved mass spectra. We demonstrate this approach for two specific examples: mass spectra of peptides from a linear quadrupole ion trap mass spectrometer, and mass spectra of intact proteins from a FTICR mass spectrometer. For peptides, we demonstrate that the FT charge determination approach reduces ambiguous charge state assignments and false positive identifications relative to a system for assigning precursor charges to tandem mass spectra. For proteins, we show that the FT charge determination approach provides improved performance for spectra of low signal-to-noise ratio compared to vendor-supplied charge state determination packages.

Experimental

All proteins, salts, buffers, iodoacetamide, and guanidine HCl, were obtained from Sigma Chemical Co. (St. Louis, MO). Tris(2-carboxyethyl) phosphine hydrochloride (TCEP) was purchased from Pierce (Rockford, IL) Sequencing grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (affiliate of Merck KGaA, Darmstadt, Germany). HPLC-grade acetonitrile and water were used for all LC-MS-MS analyses (Burdick and Jackson, Muskegon, MI). Ultrapure 18 MΩ water used for sample buffers was obtained from Millipore Milli-Q system (Bedford, MA). Fused silica capillary tubing was purchased from Polymicro Technologies (Phoenix, AZ).

Preparation and LC-MS-MS Analysis of Protein Standard Mixtures and Rhodopseudomonas palustris *Ribosomal Proteins*

A 10 ng/mL protein standard mixture (PSM) was generated from equal amounts of six proteins: bovine serum albumin (MW 69 kDa), yeast alcohol dehydrogenase I (MW 37 kDa), bovine carbonic anhydrase II (MW 29 kDa), horse myoglobin (MW 17 kDa), bovine hemoglobin (MW 15 kDa), and chicken egg lysozyme C (MW 14 kDa). Hemoglobin includes α and β polypeptides, and the isomer yeast alcohol dehydrogenase II was found to be a component of yeast alcohol dehydrogenase I, giving a total of eight polypeptides in the mixture. The mixture was first reduced with 10 mM TCEP for 20 min at ambient temperature followed by alkylation with 15 mM iodoacetamide in the dark for 45 min. Afterwards, the standard mixture was digested with 200 ng of trypsin for 1 h at 37 °C in 100 μl of 80% acetonitrile/20% 50 mM Tris HCl/10 mM $CaCl_2$ (pH 7.6). For training the decision tree (vide infra), the trypsin digest of an "extended" PSM was used [17].

70S ribosomes from *R. palustris* were purified and fractionated using a high salt sucrose cushion and sucrose density fractionation as previously described [18]. Acid extracted [19] ribosomal proteins were denatured and reduced in 6 M guanidine HCl, 50 mM Tris-HCl (pH 7.6), with 10 mM DTT at 60 °C for 45 min. Afterward, the proteins were digested with 1 μg trypsin overnight at 37 °C. Remaining disulfides were reduced with 10 mM DTT at 60 °C for 45 min.

LC-MS-MS analyses were performed with a Famos/Switchos/Ultimate HPLC system (LC Packings, a division of Dionex, San Francisco, CA) coupled to a Finnigan LTQ linear ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) equipped with a nanospray source. A 300 μm i.d. × 5 mm nano C_{18} precolumn (LC Packings) preconcentrated and desalted the samples on-line. The flow rate for reverse-phase liquid chromatography was 0.15 μl/min with a 105 min linear gradient from 100% Solvent A (95% H_2O/5% CH_3CN /0.1% formic acid) to 100% Solvent B (95% CH_3CN/5% H_2O /0.1% formic acid). A 100 μm picoFrit Tip (15 μm i.d. at the tip, New Objective, Woburn, MA) was packed via a pressure cell (New Objective) with ~15 cm of C_{18} reverse phase (Jupiter C_{18} 5 μm particles, 300 Å pore size, Phenomenex, Torrance, CA). Flow from the HPLC system was directed through this column via a tee, which also provided a connection for the electrospray voltage. The LTQ was operated in the data-dependent mode with dynamic exclusion enabled, where the three most abundant peaks in the 400–2000 *m/z* range were subjected to both ultra zoom scan and MS-MS analysis. To minimize space charging and improve mass accuracy for zoom scans, the automatic gain control target for zoom scans was decreased from a default setting of 3000 to 1000.

LC-FTICR-MS of Intact Proteins

Five proteins (ubiquitin, chicken lysozyme C, bovine ribonuclease A, bovine carbonic anhydrase II, and bovine beta lactoglobulin-B) were dissolved in HPLC grade water to give a final concentration of 1 mg/mL of each protein, and diluted as required for the analysis. Unlike the samples described above, the mixture was neither reduced nor digested. All capillary HPLC-FTICR experiments were performed with an Ultimate HPLC (LC Packings) coupled to an IonSpec 9.4 T FTICR-MS (Lake Forest, CA) mass spectrometer equipped with an Analytica electrospray source. A Vydac 214MS5.325 (Grace-Vydac, Hesperia, CA) C4 reverse phase column (300 μm i.d. × 250 mm, 300 Å with 5 μm particles) was directly connected to the Analytica electrospray source with 100 μm i.d. fused silica tubing. Injections of 5.0 μg of total protein were made onto a 100 μl loop. The flow rate was ~4 μl/min, with a 75 min gradient going from high water (95% water, 5% acetonitrile, 0.5% formic acid) to high organic (95% acetonitrile, 5% water, 0.5% formic acid). All mass spectra were acquired with a 2 s hexapole ion accumulation time; 2 scans were signal averaged, 1024 K data points were acquired, and 2 zero fills were performed. The Hann window was used for apodization. Mass resolving powers of 35,000 to 120,000 FWHM were achieved. Mass calibration was performed externally using an ubiquitin protein standard, providing approximately ~10–50 millidalton accuracy.

Data review was performed via the Omega 8 instrument control software provided by IonSpec. The most abundant isotopic mass (MAIM) for each protein was computed [20], and a spreadsheet calculated the *m/z* ratio corresponding to each charge state. Three scans of the data, containing charge state packets for the five proteins, were chosen for charge state analysis. The FTDocViewer "Isotope Clusters" feature displayed the isotopic packets from each spectrum along with the assigned charge state(s). A beta-version of IonSpec's PeakHunter algorithm (version 0.0.24) was then used to assess charges for the same spectra. Scripts were developed as described below for examination of the data in external software.

Charge Determination for Peptide Ions from Ultra Zoom Scans

Code to determine peptide charge states was implemented in two software packages: Raw2MS2 and MS2ZAssign. Raw2MS2 is a Visual Basic application designed to transcode tandem mass spectrometry data from ThermoFinnigan RAW instrument capture files into tab-delimited text files. Mass spectra are exported to MS1 files, and tandem mass spectra are written to MS2 files [21]. We modified Raw2MS2 to write zoom scan spectra to MSZ files, using the same format as the MS1 files. While XCalibur was configured to record centroided data for full-scan mass spectra and tandem mass spectra, it stored zoom scans in "profile" mode, storing the intensity at ~0.01 *m/z* intervals.

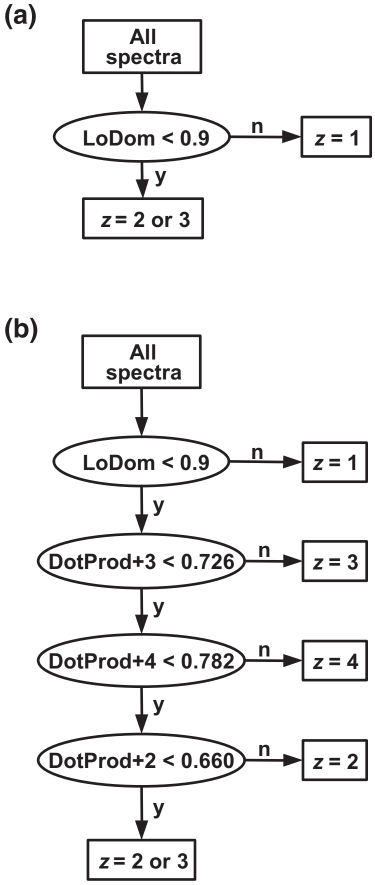

Figure 2. The charge discrimination process that is traditionally employed (**a**) can be modeled as a single step, making no effort to discriminate between spectra from doubly- and triply-charged peptides ions. MS2ZAssign uses a decision tree (**b**) to separate spectra by their precursor charges.

MS2ZAssign (developed in C++) marks appropriate precursor ion charge states within a supplied MS2 file. In its simplest mode, it separates spectra into two types: those from singly-charged peptide ions and those from multiply-charged peptide ions (see Figure 2a.) A "low dominance" score is computed for each tandem mass spectrum indicating the percentage of a spectrum's intensity that falls below the precursor's *m/z*. Tandem mass spectra from singly-charged peptides are those in which the low dominance is 90% or greater. All others are marked as coming from either doubly-charged or triply-charged precursor ions, and two identifications are performed for these spectra. In this way, MS2ZAssign's simple mode mimics the behavior of the DTA extraction software within ThermoFinnigan's BioWorks program.

MS2ZAssign also includes a more elaborate mode, which examines the zoom scan associated with each tandem mass spectrum to determine the charge(s) to be recorded for the appropriate precursor ion. The software begins by generating theoretical zoom scans for a 4 *m/z*-wide range in which a precursor ion's isotopic variants appear. The fast FTs (FFTs) of these charge models are generated and stored for later reference. The software employs the "Fastest Fourier Transform in the West" library (http://www.fftw.org). Once zoom scans have been read into memory and associated with the appropriate tandem mass spectra, the 4 *m/z* unit range in the middle of each 10 *m/z* unit wide zoom scan is extracted, and intensities are distributed among 256 bins of equal width. The average intensity is subtracted from each bin, and the resulting dataset is subjected to a 256-point FFT. For each bin, the FFT yields a complex number that is then multiplied by its complex conjugate to give the modulus, which is a real number. These FFT results from the zoom scans can then be compared to the FFTs generated from the charge model FFTs by computing normalized dot product scores:

$$s = \frac{\sum \mathrm{AB}}{\sqrt{\sum A^2 \sum B^2}}$$

where s is the score, A is the intensity in a bin of the charge model FFT, and B is the intensity in a corresponding bin of the zoom scan FFT. The charge model FFT that is the best match to the observed zoom scan FFT will produce the highest score, indicating the appropriate precursor charge for the spectrum.

Ultra Zoom Scans and identifications from an "extended" PSM containing 20 proteins [17] were used to generate a decision tree for assigning precursor charge states (see Figure 2b). This process, conducted in the R statistical environment (www.r-project.org), determined the thresholds that would most reduce the Gini impurity [22] of mixed spectra with known charge states. Spectra from singly-charged precursors are separated in the first stage on the basis of their low dominance scores. Next, spectra featuring zoom scan FFTs that are highly similar to the $z = 3$ model FFT are labeled as coming from $z = 3$ precursor ions. Similar steps then set aside spectra from quadruply- and doubly-charged precursor ions. The remaining spectra are identified twice: once as coming from doubly-charged precursors and once as coming from triply-charged precursors.

Database Identification of Peptides

Tandem mass spectra were matched with peptide sequences using the DBDigger algorithm [23] including the MASPIC scorer [24]. For the PSM samples, a sequence database was constructed that contained chicken lysozyme C, bovine serum albumin, carbonic anhydrase II, hemoglobin α and β chains, superoxide dismutase, ubiquitin, yeast alcohol dehydrogenase I and II, horse myoglobin, porcine trypsin, and 14 keratin sequences. These sequences were then repeated in the database in reversed orientation (C-terminal to N-terminal). For the "extended" PSM samples used to train the decision tree, the database contained sequences for each component of the mixture [17], plus >4800 sequences for proteins encoded by the *R. palustris* genome [25]. For ribosomal proteins, the database contained all *R. palustris* proteins, several keratin sequences, and trypsin in both forward and reverse orientations. Database searches were conducted in semi-tryptic mode, requiring only one end of each candidate peptide sequence to conform to a trypsin cutting site. Post-translational modifications included: oxidation of methionine, loss of ammonia from glutamine (only when Gln was the N-terminal residue), loss of ammonia from asparagine (only when Gly was the next residue in the sequence), cysteine to thioproline conversions (only when Cys was the N-terminal residue), and carboxyamidomethylation of cysteine. These chemical modifications were chosen because they are common results from sample handling; the ammonia losses can happen spontaneously, thioproline can result from the use of formic acid, and carboxyamidomethylation results from reduction and alkylation of proteins. In addition to these modifications, database searches on ribosomal proteins also included beta methylthiolation on aspartic acids and single, double, or triple methylations on lysines and arginines.

After each database search, the DBDigger scores for spectra that were matched to peptide sequences that could only be found in reversed protein sequences were set aside for analysis by SQTRevPuller, a C++ program used at ORNL for score threshold determination. Score thresholds for each charge that would remove 97% of these reverse hits were determined from the score distributions. Likewise, a threshold that would remove 75% of the reverse hits was determined for DeltCN, which is the fractional difference in DBDigger score between the two highest scoring peptide matches for a given tandem mass spectrum. These thresholds were used to filter the identifications by DTASelect, version 1.9 [26]. DTASelect also employed the −e rev option to remove all reversed proteins from the displayed list.

Charge Determination for Protein Ions from FTICR Spectra

The process by which intact protein ion charges can be measured is essentially the same as for peptides. The mass spectra from the IonSpec instrument are extracted to an MS1 file [21] by "MakeMS1", a Visual Basic Script for FTDocViewer in the Omega8 instrument control software. The isotopic packets for proteins ranging in charge from $z = 5$ to 30 are modeled, and FFTs of these charge models are stored (see Figure 3). The observed mass spectra are read into memory by the "Tact"

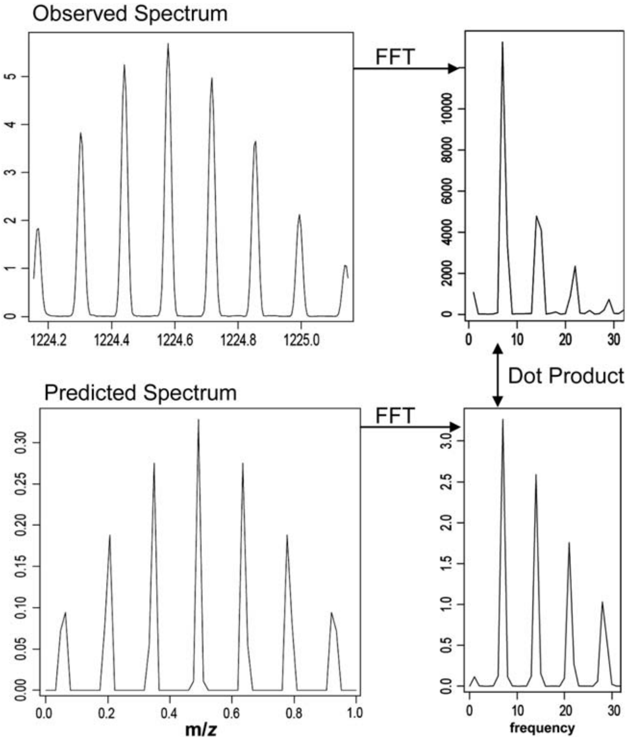

Figure 3. Procedure for determining charge state from FTICR data. Observed isotopic packets from FTICR mass spectra are subjected to FFT, thus identifying their key frequencies (top 2 panels). A modeled isotopic packet ("Predicted Spectrum") is computed for each charge state and subjected to FFT (bottom 2 panels). The observed isotopic packet's FFT is compared to the FFTs of the models by normalized dot product, and high scores indicate the most likely charge states. In this example, the ubiquitin charge state is $z = 7$. Its isotopic peaks are spaced 1/7 *m/z* apart, and the FFT of its isotopic packet shows a strong peak at a frequency of seven peaks per *m/z*.

algorithm, C++ software created at ORNL for analysis of FTICR data from intact proteins. The software identifies the set of nonoverlapping one *m/z*-wide windows containing the highest intensity within each mass spectrum. The FFT of each one *m/z*-wide window is computed, and the charge model FFT that best matches the FFT of the observed spectrum (in terms of normalized dot product score) is stored as the charge state for that packet.

Binning the peaks from FTICR mass spectra is somewhat different than this operation for zoom scan mass spectra. Zoom scan spectra are "profile" data comprised of intensity samples taken at small *m/z* intervals. The MakeMS1 script, however, can generate either centroided peak lists or *m/z* profiles from the FTICR mass spectra. If a peak centroid is added to an intensity bin array for FFT, small *m/z* errors may cause the centroid to contribute to a bin at a higher or lower *m/z* than it should (thus distorting the frequencies discovered by FFT). To mitigate this problem, Tact adds intensity to multiple bins for each centroid rather than just one bin. When profile data are available, the software increments only one bin for each point on the profile.

Results and Discussion

FT of Ultra Zoom Scan Spectra for Peptide Charge Determination

Figure 4 shows examples of the Fourier transforms of observed mass spectra from linear ion trap ultra zoom scans. Note the similarity of the appearance of these FTs of experimental spectra to those shown for model spectra in the lower half of Figure 1. The FTs of the experimental spectra show a peak corresponding to the charge state of the ion, in agreement with the model described in the Introduction. Especially for the $z = 4$ mass spectrum shown in Figure 4, some high-frequency noise is observed. However, the low-frequency range of the corresponding FT, which contains the charge state information, is not significantly affected by the noise, which would appear in higher-frequency regions in the FT. To obtain results such as those shown in Figure 4, it was necessary to decrease the number of ions allowed into the ion trap via the automatic gain control (AGC) feature of the LTQ. The default AGC settings for the zoom scan mode allowed a sufficient excess of ions into the trap such that the spacing between isotope peaks in ultra zoom scans was slightly less than $1/z$ increments, perhaps due to space charge effects.

The protein standard mixture yielded 4984 spectra in a single LC-MS-MS elution. The simpler charge determination mode of MS2ZAssign (Figure 2a) separated them into 1389 spectra from singly-charged precursors and 3595 spectra from multiply-charged precursors. The latter were analyzed by DBDigger twice, assuming both $z = 2$ and $z = 3$. The total number of identifications performed was 8579. When zoom scans were used to determine charge state through the decision tree shown in Figure 2b, the number of identifications changed considerably. One thousand three hundred eighty-eight spectra were targeted for identification as singly-charged peptides, while 689, 457, and 201 were marked for identification as doubly-, triply-, and quadruply-charged peptides, respectively. Of the 4984 tandem mass spectra, 2249 were associated with zoom scans that did not allow the software to assess a unique charge state; for these spectra, precursor ions with both $z = 2$ and 3 were used for identification with DBDigger. By using the information in zoom scans, MS2ZAssign left 37% fewer spectra (2249 versus 3595) with ambiguous ($z = 2$ or 3) charge state assignment, and reduced by 16% the number of identifications performed by DBDigger.

DBDigger identified the spectra using the charge states assigned by each of these two MS2ZAssign modes, and DTASelect filtered the resulting identifications. Seven hundred fifty-seven identifications passed thresholds using the simple charge state assignment system, and 747 passed with the new zoom scan-based discriminator. For the simple system, the identifications were split among 135 ions with $z = 1$, 392 with $z = 2$, and 230 with $z = 3$. For the zoom scan system, the split

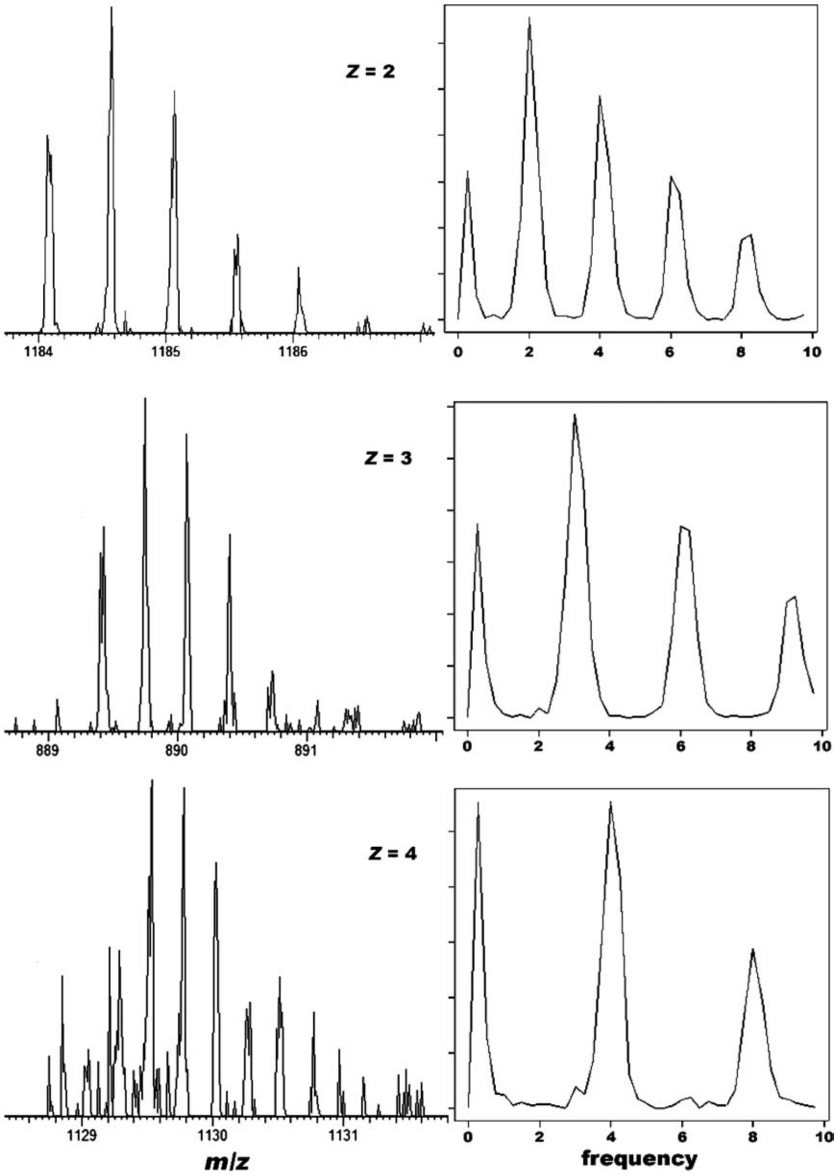

Figure 4. Ultra zoom scans from peptides with $z = 2, 3$, and 4. The Fourier transforms of these signals are shown to the right. The frequencies dominating the FTs correspond to the charge carried by the precursor ions, enabling charge assignment before sequence identification.

was 135 ions with $z = 1$, 357 with $z = 2$, 200 with $z = 3$, and 55 with $z = 4$. While it may appear that the simple system produced more identifications, it is worth noting that 29 of the spectra from the simple system gave high scoring identifications for precursors with both $z = 2$ and 3. Because the precursor ions bear one particular charge, only one of the two identifications for each spectrum can be correct. The number of double identifications dropped from 29 to 4 for the new zoom scan-based charge assignment system. The corrected numbers of passing identifications were 728 for the simple system and 743 for the zoom scan system.

It should be noted that the protein database used for these PSM searches was small by proteomics standards, containing some 25 sequences, each in forward and reverse orientations. The use of small databases is generally not advisable due to increased difficulty in estimating error rates in identifications. For this reason, we included the reversed sequences in the database and used SQTRevPuller to provide cutoff thresholds tailored to each raw file, to reduce artifacts introduced by using a small database. Nonetheless, a slightly larger number of false positive identifications resulting from using a small database is to be expected. However, this less stringent search also provided the opportunity to evaluate the performance of the charge state determination systems on, for instance, low-abundance peptides, which would yield zoom scans and tandem mass spectra with low signal-to-noise ratios. The use of a smaller database was therefore justified for our pur-

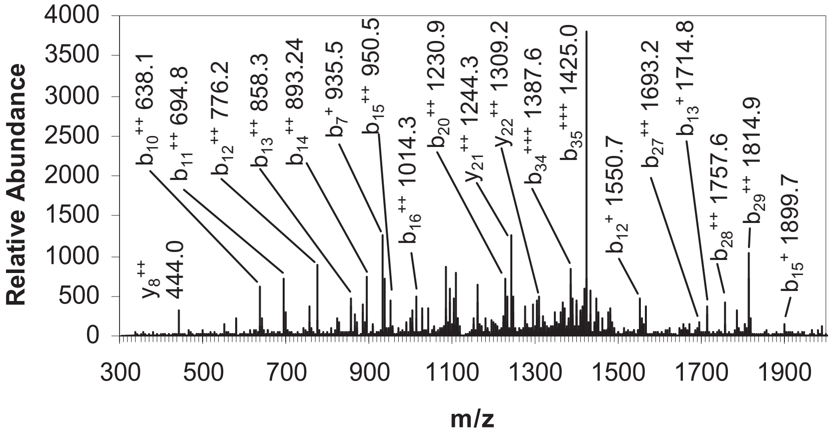

Figure 5. Tandem mass spectrum from the quadruply-charged sequence RHPYFYAPELLYYANKYNGVFQECCQAEDKGACLLPK. The dominance of b ions observed for this peptide can be explained by the presence of both arginine and histidine at the peptide N-terminus. The most intense ion resulted from the loss of the C-terminal dipeptide at a labile peptide bond. DBDigger reported that 74.7% of the fragment ions sought were matched to observed peaks. DBDigger, however, generates only $z = 1$ and 2 fragment assignments for quadruply-charged precursor ions.

pose, which was evaluation of the charge state determination systems.

Using FFT to analyze zoom scans enables us to identify spectra from higher charge states than previously possible, as Figure 5 shows for a $z = 4$ ion. Of the 55 confident identifications resulting from $z = 4$ peptide ions, 26 represented sequences that were also identified from other spectra, generally at lower charge states. Of the remaining identifications examined, 17 had N- and C-termini that each matched to other identifications, and all 12 remaining sequences had at least one terminus matching to other identifications. These similarities between $z = 4$ identifications and those of lesser charged peptide ions increase the confidence that can be placed in these identifications. The ability to recognize tandem mass spectra from peptide ions with $z = 4$ by their accompanying zoom scans may make it easier to tune database search tools to identify them. The greater length of peptides with $z = 4$ may enable researchers to improve observed sequence coverage for protease-resistant proteins.

When charge state assignment neglects the possibility of $z = 4$ spectra, false identifications result. All of the 55 successfully identified $z = 4$ spectra in the protein standard mixture were incorrectly labeled as $z = 2$ and $z = 3$ spectra under the simple charge assignment system. A comparison of the correctly identified $z = 4$ sequences to the sequences identified assuming $z = 2$ and $z = 3$ gave the surprising result that 12 identified sequences from lower charges matched part of the $z = 4$ sequences. For example, scan 9270 matched LASHLPSDFTPAVHASLDKFLANVSTVLTSK with $z = 4$ and LDKFLANVSTVLTSK with $z = 2$. In eight cases, this partially matching sequence resulted from the assumption that $z = 3$, but in four cases the partially matching sequence resulted from the $z = 2$ identification. In three of the 12 partially matching sequences, the N-terminus of the sequence was matched rather than the C-terminus of the sequence. These partial matches resulted from the fact that a subsequence (drawn from either the N- or C-terminus of a peptide sequence) will yield many of the same fragment ions as the full sequence. These two sequences matched to scan 9270 above would produce y ions at the same *m/z* ratios (though the sequence for $z = 2$ would produce fewer ions). The database search employed was semi-tryptic, meaning that peptide sequences were required to match to a Lys or Arg cleavage site on one terminus, giving a higher chance that partial sequence matches could result. These results suggest that proteomic identifications that do not explicitly handle charge assignment for peptides with $z = 4$ probably contain some incorrect identifications at $z = 2$ and $z = 3$ that are actually from quadruply-charged peptides.

Examination of spectra collected for an *R. palustris* ribosomal sample produced similar findings. Identification was conducted using searches configured for a variety of post-translational modifications that have been previously observed in these proteins. The total number of confidently identified spectra was roughly the same for both charge inference systems: 2152 spectra for the zoom scan-based system and 2096 for the simple system. After subtracting 1 for each double identification ($z = 2$ or 3), the zoom scan charge inference system identified 2100 spectra while the older system resulted in 1988 identifications. The identifications from spectra of precursor ions with $z = 4$ comprised a similar percentage of the confident identifications: 7% of the standard mixture identifications and 6% of the ribosome identifications.

An examination of proteins that gained and lost confident identifications when the FFT-based charge inference algorithm was used is instructive. Eight ribosomal proteins gained identified spectra from this system, and 20 lost them, compared to 19 proteins that stayed the same. Ribosomal proteins L21 and L25 each gained four more spectra. L21's sequence coverage remained the same because the four additional spectra from $z = 4$ precursors that it gained occupied the same regions of sequence that its $z = 3$ precursors covered. One of the gained peptides with $z = 4$ was identified in three replicates of the same spectrum. L25's additional spectra increased its sequence coverage marginally (from 78.3 to 80.4%) by adding a $z = 4$ peptide that stretched 57 residues through a region repeating the motif PAAAAK. By reinforcing other identifications in this region, the $z = 4$ peptides help confirm that this prokaryotic protein contains a repetitive low-complexity sequence [18, 25]. In each of these two cases, quadruply-charged peptides were identified that corresponded well to identifications at lower charges.

At first glance, it appeared that the losses of five spectra each for ribosomal proteins S2 and S6 represented a failure in the FFT charge state inference system. A closer look, however, reveals that many of these losses were actually reductions in false positives by the new system. Using MS2ZAssign and DBDigger, a peptide from S6 in scan 8317 was identified as an ion with $z = 2$ from an internal tryptic peptide, but it was also identified as a $z = 3$ ion from an N-terminal peptide of S6 that extended to the same C-terminus as the internal peptide. One of these identifications must be false, and the FFT charge state inference system unambiguously assesses this peptide with $z = 2$. In the traditional charge assignment system, scan 2966 is identified as a peptide from S6 with $z = 2$, but it is also identified as a peptide from another protein with $z = 3$. The FFT charge state system identifies the charge state of this peptide exclusively as $z = 3$. All of the "lost" spectral identifications for S6 are actually duplicate identifications that were successfully screened out by the FFT charge assignment system; analysis of peptide identifications from S2 shows similar results. By reducing the average number of identifications produced for each spectrum, this system reduces false positive identifications from tandem mass spectral collections.

Charge Measurement for Intact Protein Isotope Clusters in FTICR Spectra

The resolution of FTICR mass spectrometry makes it possible to apply our technique to small sections of normally-acquired mass spectra rather than requiring special high-resolution scans, such as the ultra zoom scans mentioned above for the LTQ instrument. The Tact software for intact protein identification from FTICR data was adapted to find isotopic packets in collections of mass spectra and perform charge state assignments by FT. The FFTs of these packets were compared to FFTs of modeled isotopic packets to determine charges. Because most proteins adopt multiple charge states under electrospray conditions, multiple isotopic packets of known charge are available to test charge state detection algorithms. Three mass spectra from a liquid chromatographic separation interfaced via electrospray with the FTICR were examined; scan 10 included charge packets for ribonuclease A, scan 23 showed the presence of ubiquitin and lysozyme, and scan 42 gave evidence for beta lactoglobulin and carbonic anhydrase. These three spectra are included as text files in Supplementary Material (which can be found in the electronic version of this article.). Table 1 compares the performance of IonSpec's "FTDocViewer" and "PeakHunter" software to that of Tact for charge state inference. Each charge determination reported from Tact is the top-scoring match. While Table 1 also lists the Tact score for each top-ranking assignment, it is important to emphasize that these scores are not used in an absolute sense, but rather for ranking matches for each isotope packet. That is, there is no absolute threshold score above which a charge state assignment is accepted. Instead, the reported charge state assignment is simply that with the highest Tact score. While Tact reported only one charge assignment for each peak packet, FTDocViewer and PeakHunter can report multiple charge assignments for each set of isotopic peaks, giving them a better chance of randomly hitting the correct charge but reducing their specificity.

The isotopic packets in scan 10 for ribonuclease A were intense, but also contained additional isotopic packets near the most intense packets, suggesting that other forms of the protein were also present. Perhaps because of these additional packets, the two most intense packets, corresponding to the $z = 8$ and $z = 9$ charge states of the protein, resulted in multiple charge state calls by FTDocViewer and PeakHunter and an incorrect charge assignment by Tact. For less intense isotope packets corresponding to higher charge states, Tact and PeakHunter yielded correct results, while FTDoc returned multiple possible charges for the $z = 11$ and 12 states.

Scan 23 included isotopic packets for ubiquitin and lysozyme. Ubiquitin's packets for $z = 7$ through 9 were the most intense, and they were more than an order of magnitude more intense than lysozyme's sole isotopic packet at $z = 9$. All of these packets were assigned the correct charge by all three algorithms. The $z = 5$ charge state for ubiquitin, however, was called correctly by only the Tact algorithm. This isotopic packet was the least intense to be assigned a correct charge in this collection of mass spectra.

Scan 42 comprised a much greater challenge. β-Lactoglobulin and carbonic anhydrase both contributed isotopic packets, but β-lactoglobulin's $z = 14$ charge state was approximately 10-fold more intense than the most intense carbonic anhydrase isotopic packet. FTDocViewer and PeakHunter both yielded multiple charge state calls for many isotopic packets. In several cases, the packets for carbonic anhydrase were not assigned charges by these algorithms, presumably because these low-intensity peaks were not easily centroi-

Table 1. Automated protein charge state assignments from FTICR data

			Charge State Assignment[a]			
Charge	MAIM m/z	Intensity	FTDoc Z	PeakHunter Z	Tact Z	Tact Score
			Ribonuclease A, Scan 10			
7	1956.04	0.05	No Call	No Call	5	0.46
8	1711.66	11.35	10, **8**	10, 4, 4	9	0.58
9	1521.59	21.14	10, **9**	1, 2, **9**	11	0.49
10	1369.53	7.15	**10**	**10**	**10**	0.75
11	1245.12	5.80	**11,** 1	**11**	**11**	0.87
12	1141.44	0.83	3, **12**	**12**	**12**	0.91
			Ubiquitin, Scan 23			
5	1713.92	0.06	2, 4	No Call	**5**	0.76
6	1428.44	4.80	**6**	**6**	**6**	0.86
7	1224.52	26.74	**7**	**7**	**7**	0.83
8	1071.58	19.30	**8**	**8**	**8**	0.84
9	952.62	15.98	**9**	**9**	**9**	0.97
10	857.46	5.53	**10**	**10**	**10**	0.97
11	779.60	2.67	**11**	**11**	**11**	0.92
12	714.72	0.49	**12**	**12**	**12**	0.89
			Lysozyme, Scan 23			
9	1590.87	1.16	**9,** 2	**9,** 4	**9**	0.76
			Beta Lactoglobulin, Scan 42			
11	1662.67	0.09	8	No Call	18	0.47
12	1524.20	0.10	8, 6, 2	**12,** 3, 3	6	0.37
13	1407.03	0.40	2, **13**	14	**13**	0.49
14	1306.60	3.61	2, **14**	1, **14**	**14**	0.50
15	1219.56	2.31	1, 14, **15**	**15,** 2	**15**	0.59
16	1143.40	0.09	2	16	**16**	0.65
17	1076.20	0.14	No Call	29, 7	27	0.34
			Carbonic Anhydrase, Scan 42			
21	1383.08	0.08	4	3	30	0.31
22	1320.26	0.11	5	4	**22**	0.31
23	1262.90	0.38	No Call	8, 8	**23**	0.26
24	1210.32	0.30	1	12, 12	8	0.29
25	1161.95	0.17	1	10, 5	**25**	0.48
26	1117.30	0.16	2, 4	**26**	**26**	0.38
27	1075.95	0.14	No Call	29, 7	**27**	0.34
28	1037.56	0.07	No Call	No Call	**28**	0.54
29	1001.82	0.10	No Call	**29**	**29**	0.63
30	968.46	0.07	No Call	No Call	**30**	0.40

[a] Correct charge assignments are shown in **bold font.**

ded. Tact, however, was able to assign four consecutive correct charge states for β-lactoglobulin. For carbonic anhydrase, only Tact was able to achieve any consistency, assigning eight of the 10 charge states correctly. Scan 42 demonstrates that FFT is particularly powerful for inferring charge states from noisy signals of low intensity.

Overall, Tact performed comparably to FTDocViewer and PeakHunter in determining charge states from ion packets of moderate intensity and signal-to-noise ratio, while providing an improvement for low-abundance, noisy isotope packets. It is important to keep in mind that, in the context of this LC-MS experiment, the proteins were available during only a limited time for MS data acquisition during elution of a peak in a liquid chromatography separation. As such, the optimal performance factors for high-resolution mass measurement must be compromised somewhat to accommodate the shorter time frame for ion detection (i.e., few scans and fewer data points for the transient signal). The exquisite resolution possible for FTICR instruments, along with FTDocViewer and PeakHunter processing algorithms, enables accurate determination of charge states for proteins up to at least 60 kDa under direct infusion conditions, where protein concentrations and ion accumulation and detection parameters can be optimized. Accurate charge state determination is a critical component for computational programs such as THRASH [27] which seek to combine this information with isotopic abundance to permit comparisons between measured and predicted mass spectra for molecular identifications.

Other Approaches to Charge State Determination

One alternative approach to that described in this paper would be to calculate dot products between model spectra for the various charge states with the experimental isotope packets, without performing the FT. However, this dot product, which is simply the value of the cross-correlation of the two functions evaluated at zero, does not account for the fact that the peaks in the experimental spectra do not occur at integral multiples of $1/z$, while peaks in the model spectra would occur at integral multiples of $1/z$. That is, this dot product would not necessarily correspond to the maximum of the cross-correlation, so that a more complex scoring scheme would be required. One advantage of using the FT approach is that the "shift" of experimental peaks in the mass spectrum by the fractional part of the mass does not affect the locations of peaks in the Fourier domain. This "shift" information is contained in the phase of the Fourier transform (reference [9], pp 104–107), which we eliminate by calculating the modulus of the FT (i.e., multiplying the ordinate of each point by its complex conjugate.) In this way, the maxima of peaks in the FT appear at integral multiples of z, so that a simple dot product of the model FT and experimental FT accurately evaluates the similarity between the two.

An alternative approach would be to locate maxima corresponding to each peak in the isotope packet of an ion, and determine the charge from the reciprocal of the separation of adjacent maxima. Peak detection remains an active area of research both for chromatographic [28] and mass spectrometric [29, 30] data, which suggests that a fair amount of optimization of such a routine would be required for our application. The number of individual isotope peaks that could be detected with a peak detection algorithm would vary with both signal-to-noise ratio and with the width of the isotopic packet. Successful detection of some, but not all, peaks in the packet would cause problems for charge state determination, especially if "undetected" peaks occur between "detected" peaks. In this case, the peak separation in *m/z* space would give rise to an incorrect charge. Although beyond the scope of this paper, a systematic comparison of the performances of a peak detection routine with the FT approach, especially for spectra with low signal-to-noise ratios, would be informative. It should also be possible to apply a peak detection algorithm to the FT of a charge state packet, for the purpose of locating the first peak (which gives the charge state) in the FT.

Conclusions

Fourier transforms of sections of isotopically-resolved mass spectra appear to be a robust tool for determining ionic charge states from periodic features in these spectra. We demonstrated charge state measurement from ultra zoom scan spectra of peptides obtained from a linear-quadrupole ion trap mass spectrometer and from high-resolution spectra of intact proteins obtained from an FTICR instrument, although our approach should be applicable to isotopically-resolved spectra from any mass spectrometer. For peptides, the direct determination of charge states offers advantages of decreased false positives as well as increased speed for subsequent database identification algorithms. However, the increased time required to obtain ultra zoom scans offsets somewhat the ability to obtain tandem mass spectra. Future work will seek to optimize the accuracy and speed advantages of direct charge state determination against acquisition of more tandem mass spectra. For high-resolution spectra of proteins obtained by FTICR, no such tradeoff need to be considered, as the mass spectra normally obtained are sufficiently resolved for a direct charge state determination. Existing charge state assignment algorithms for FTICR data appear to require centroiding before charge determination, and errors in this process can lead to errors in assessed charges. Use of FFT for charge determination does not require centroiding and appears to achieve superior sensitivity and noise suppression than algorithms of this type, especially for LC-FTICR-MS measurements.

In this research, each isotopic packet was treated as an independent charge assignment problem. If, however, the algorithm were configured to exploit sets of related isotopic packets, errors in charge determination could be reduced [16]. For example, Tact was able to assign charges of $z = 22$, 23, and 25 through 30 for carbonic anhydrase (see Table 1), but it failed to assign the $z = 24$ state correctly. If the software were designed to look for coherent series of charge states, this type of error would occur less often. This approach is also applicable to peptide ions, where ions representing different charge states of the same peptide are often present in a single mass spectrum.

While some peptides yield ultra zoom scans of the quality observed in Figure 4, most contain higher levels of noise and feature less resolution. The decision tree system employed (see Figure 2b) reverts to assigning both $z = 2$ and 3 charge states to spectra that have FFTs that do not resemble the charge model FFTs. It may be the case that the set of spectra with indiscernible charge states contains a large proportion of unidentifiable spectra. Systems that are designed to evaluate spectral quality before identification could benefit by incorporating information from high-resolution precursor ion scans in conjunction with data from the tandem mass spectra to assess the probabilities of successful identification.

There are several ways in which this strategy for charge assignment could be improved. First, the charge models were constructed rather empirically, with an emphasis on optimum function rather than on capturing isotopic peak width and relative isotopic intensity accurately. If these models resembled the observed data more faithfully, their FFTs would likely result in superior discrimination by dot product scores. It is also possible that these scores could be interpreted more effectively by a Support Vector Machine or neural net rather than a decision tree. The computation of FFTs requires 2^n equally-spaced data points, and this requirement must be reconciled with the density and spacing of intensity samples from the mass spectrometry data. Optimizing these aspects of the algorithm could potentially make this charge assignment approach even more effective. For these reasons, Tact and MS2ZAssign are currently research tools that are undergoing further development and testing. It is the authors' intent that this report might lead to the use of, and improvements upon, this charge state assignment approach.

Acknowledgments

The authors acknowledge Hayes McDonald for advice on high-resolution measurements in quadrupole ion trap mass spectrometers. Research was funded by the United States Department of Energy's Office of Biological and Environmental Research through the Genomics: GTL Center for Molecular and Cellular Systems. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract number DE-AC05-00OR22725.

References

1. Schwartz, J. C.; Syka, J. E. P.; Jardine, I. High resolution on a quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **1991,** *2,* 198–204.
2. Schwartz, J. C.; Senko, M. W.; Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 659–669.
3. Marshall, A. G.; Hendrickson, C. L. Fourier transform ion cyclotron resonance detection: Principles and experimental configurations. *Int. J. Mass Spectrom.* **2002,** *215,* 59–75.
4. Loo, J. A.; Quinn, J. P.; Ryu, S. I.; Henry, K. D.; Senko, M. W.; McLafferty, F. W. High-resolution tandem mass spectrometry of large biomolecules. *Proc. Natl. Acad. Sci. U.S.A.* **1992,** *89,* 286–289.
5. Prebyl, B. S.; Cook, K. D. Use of Fourier transform for deconvolution of the unresolved envelope observed in electrospray ionization mass spectrometry of strongly ionic synthetic polymers. *Anal. Chem.* **2004,** *76,* 127–136.
6. Wallace, W. E.; Guttman, C. M. Data analysis methods for synthetic polymer mass spectrometry: Autocorrelation. *J. Res. Natl. Inst. Stand. Technol.* **2002,** *107,* 1–17.
7. Danis, P. O.; Huby, F. J. The computer-assisted interpretation of copolymer mass spectra. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 1112–1118.
8. Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res.* **1998,** *8,* 175–185.
9. Bracewell, R. N. *The Fourier Transform and its Applications*, 2nd ed.; McGraw-Hill: New York, 1986, pp 70, 77, 412, 101–104.
10. Eng, J. K.; McCormack, A. L.; Yates, J. R. III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
11. Perez, R. E.; Asara, J. M.; Lane, W. S. ZSA: Peptide precursor charge state determination directly from MS/MS spectra. *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*; Orlando FL, June, 2002.
12. Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R. Code developments to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome Res.* **2002,** *1,* 211–215.
13. Fridman, T.; Day, R.; Razumovskaya, J.; Xu, D.; Gorin, A. Probability profiles—Novel approach in tandem mass spectrometry de novo sequencing. CSB 2003. Proceedings of the 2003 IEEE Bioinformatics Conference; Stanford, CA August, 2003.
14. Colinge, J.; Magnin, J.; Dessingy, T.; Giron, M.; Masselot, A. Improved peptide charge state assignment. *Proteomics* **2003,** *3,* 1434–1440.
15. 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.
16. Zhang, Z. Q.; 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.
17. Tabb, D. L.; Thompson, M. R.; Khalsa-Moyers, G.; VerBerkmoes, N. C.; McDonald, W. H. MS2Grouper: Group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J. Am. Soc. Mass Spectrom.* **2005,** *16,* 1250–1261.
18. Strader, M. B.; VerBerkmoes, N. C.; Tabb, D. L.; Connelly, H. M.; Barton, J. W.; Bruce, B. D.; Pelletier, D. A.; Davison, B. H.; Hettich, R. L.; Larimer, F. W.; Hurst, G. B. Characterization of the 70S ribosome from *Rhodopseudomonas palustris* using an integrated "top-down" and "bottom-up" mass spectrometric approach. *J. Proteome Res.* **2004**, *3,* 965–978.
19. Hardy, S. J. S.; Kurland, C. G.; Voynow, P.; Mora, G. Ribosomal proteins of *Escherichia coli*. I. Purification of 30S ribosomal proteins. *Biochemistry* **1969**, *8,* 2897–2905.
20. Kubinyi, H.Calculation of isotope distributions in mass spectrometry—A trivial solution for a nontrivial problem. *Anal. Chim. Acta* **1991**, *247,* 107–119.
21. McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R. MS1, MS2, and SQT—Three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **2004**, *18,* 2162–2168.
22. Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; John Wiley and Sons: New York, 2001; Chap. VIII.
23. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L. DBDigger: Reorganized proteomic database identification that improves flexibility and speed. *Anal. Chem.* **2005,** *77,* 2464–2474.
24. Narasimhan, C.; Tabb, D. L.; VerBerkmoes, N. C.; Thompson, M. R.; Hettich, R. L.; Uberbacher, E. C. MASPIC: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem.* **2005,** *77,* 7581–7593.
25. Larimer, F. W.; Chain, P.; Hauser, L.; Lamerdin, J.; Malfatti, S.; Do, L.; Land, M. L.; Pelletier, D. A.; Beatty, J. T.; Lang, A. S.; Tabita, F. R.; Gibson, J. L.; Hanson, T. E.; Bobst, C.; Torres, J. L. T. Y.; Peres, C.; Harrison, F. H.; Gibson, J.; Harwood, C. S. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nat. Biotech.* **2004,** *22,* 55–61.
26. Tabb, D. L.; McDonald, W. H.; Yates, J. R. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002,** *1,* 21–26.
27. 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.
28. Shackman, J. G.; Watson, C. J.; Kennedy, R. T. High-throughput automated post-processing of separation data. *J. Chromatogr. A* **2004,** *1040,* 273–282.
29. Jarman, K. H.; Daly, D. S.; Anderson, K. K.; Wahl, K. L. A new approach to automated peak detection. *Chemom. Intell. Lab. Syst.* **2003,** *69,* 61–76.
30. Wallace, W. E.; Kearsley, A. J.; Guttman, C. M. An operator-independent approach to mass spectral peak identification and integration. *Anal. Chem.* **2004,** *76,* 2446–2452.