Charger: Combination of Signal Processing and Statistical Learning Algorithms for Precursor Charge-State Determination from Electron-Transfer Dissociation Spectra

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2008

Impact Factor: 5.64 · DOI: 10.1021/ac071332q · Source: PubMed

3 AUTHORS, INCLUDING:

Rovshan G Sadygov
University of Texas Medical Branch at Galves…

51 PUBLICATIONS **3,815** CITATIONS

SEE PROFILE

Anal. Chem. **2008,** *80,* 376−386

Charger: Combination of Signal Processing and Statistical Learning Algorithms for Precursor Charge-State Determination from Electron-Transfer Dissociation Spectra

Rovshan G. Sadygov,* Zhiqi Hao, and Andreas F. R. Huhmer

ThermoFisher Scientific, 355 River Oaks Parkway, San Jose, California 95134

Tandem mass spectrometry in combination with liquid chromatography has emerged as a powerful tool for characterization of complex protein mixtures in a high-throughput manner. One of the bioinformatics challenges posed by the mass spectral data analysis is the determination of precursor charge when unit mass resolution is used for detecting fragment ions. The charge-state information is used to filter database sequences before they are correlated to experimental data. In the absence of the accurate charge state, several charge states are assumed. This dramatically increases database search times. To address this problem, we have developed an approach for charge-state determination of peptides from their tandem mass spectra obtained in fragmentations via electron-transfer dissociation (ETD) reactions. Protein analysis by ETD is thought to enhance the range of amino acid sequences that can be analyzed by mass spectrometry-based proteomics. One example is the improved capability to characterize phosphorylated peptides. Our approach to charge-state determination uses a combination of signal processing and statistical machine learning. The signal processing employs correlation and convolution analyses to determine precursor masses and charge states of peptides. We discuss applicability of these methods to spectra of different charge states. We note that in our applications correlation analysis outperforms the convolution in determining peptide charge states. The correlation analysis is best suited for spectra with prevalence of complementary ions. It is highly specific but is dependent on quality of spectra. The linear discriminant analysis (LDA) approach uses a number of other spectral features to predict charge states. We train LDA classifier on a set of manually curated spectral data from a mixture of proteins of known identity. There are over 5000 spectra in the training set. A number of features, pertinent to spectra of peptides obtained via ETD reactions, have been used in the training. The loading coefficients of LDA indicate the relative importance of different features for charge-state determination. We have applied our model to a test data set generated from a mixture of 49 proteins. We search the spectra with and without use of the charge-state determination. The charge-state determination helps to significantly save the database search times. We discuss the cost associated with the possible misclassification of charge states.

An integrated system of liquid-chromatography (LC) and mass spectrometry is widely used as a fast and efficient tool for identification and characterization of proteins in biological samples.[1,2] In such configurations, LC separates peptides in a complex mixture by their hydrophobicity and other physicochemical properties.[3] The tandem mass spectrometry then selects a single mass-to-charge ratio from the eluting peptides and subjects them to fragmentation, providing product ions that are detected by the second mass analyzer. The product ions are representative of the precursors. The peak list information of product ions and precursor mass-to-charge ratio values are used to search protein and nucleotide databases to identify the amino acid sequence represented by the spectrum and thus identify the protein from which the peptide was derived. The identification procedure is commonly done in a high-throughput manner by database search algorithms.[4−10] With this approach, thousands of peptides are identified and this information can be used together with quantification analyses to infer biologically important information.

A conventional method of peptide fragmentation employed in high-throughput proteomics studies has been low-energy collision-induced dissociation (CID).[11] In this technique, gas-phase peptides

* To whom correspondence should be addressed. Phone: (408) 965-6039. Fax: (408) 965-6190. E-mail: rovshan.sadygov@thermofisher.com.

(1) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. 3. *Nat. Biotechnol.* **1999**, *17*, 676−82.
(2) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212−7.
(3) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat. Biotechnol.* **2004**, *22*, 214−9.
(4) Eng, J. K.; A. L. M.; Yates, III, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.
(5) Sadygov, R. G.; Cociorva, D.; Yates, J. R., III. *Nat. Methods* **2004**, *1*, 195−202.
(6) Sadygov, R. G.; J. R. Yates, I. *Anal. Chem.* **2003**, *75*, 3792−8.
(7) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−67.
(8) Geer, L. Y.; Markey, S. P.l Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. OMSSA: An Efficient MS/MS Peptide Spectra Search Algorithm with Explicit Probability Scoring. Nashville, TN. 2004 (conference proceeding).
(9) Craig, R.; Cortens, J. P.; Beavis, R. C. *J. Proteome Res.* **2004**, *3*, 1234−42.
(10) Sadygov, R.; Wohlschlegel, J.; Park, S. K.; Xu, T.; Yates, J. R., III. *Anal. Chem.* **2006**, *78*, 89−95.
(11) Papayannopoulus, I. A. *Mass Spectrom. Rev.* **1995**, *14*, 49−73.

10.1021/ac071332q CCC: $40.75
Published on Web 12/15/2007

are charged in electrospray ionization by protonation. The fragmentation is induced by collision of the protonated peptides with atoms of an inert gas. Translational energy of inert gas atoms is imparted onto the peptide and is converted into vibrational energy that is then distributed throughout the peptide bonds. When the vibrational energy exceeds the activation energy for the bond cleavage, the dissociation can occur. Low-energy CID generates mainly series of b- and y- type fragment ions.[11] The presence of basic amino acids may impede random protonation and lead to prevalence of only few fragments (containing arginine and lysine) in the tandem mass spectra.[12] Besides, it has been observed that other bonds, especially those of side chains containing posttranslationally modified groups (e.g., phosphorylation), can be lower energy pathways for dissociation. As a result, mass spectra of these peptides are often dominated by ions corresponding to neutral group losses from the peptides, rather products of peptide backbone fragmentations. These problems of fragmentation may often impede successful peptide sequencing.[13]

The advent of the new fragmentation methodology, electron-transfer dissociation (ETD),[13−18] creates new opportunities for mass spectrometry-based proteomics research. It is believed to be similar to electron capture dissociation (ECD),[19,20] often used in top-down proteomics studies.[21,22] Similar to ECD, ETD is thought to occur via recombinant dissociation with a characteristic nonergodic nature allowing amine bond ($N-C_\alpha$) cleavage before randomization of excitation energy. The main ion types generated in ETD are c- and z-ions. Like collision-induced dissociations, ETD reactions can be implemented in a high-throughput mode.

ETD fragmentation is thought to provide more extensive characterization of the primary structure of polypeptides (e.g., posttranslational modifications) than conventional methods of peptide fragmentation. The peptide bond cleavage rate is thought to be faster than the rate of the intramolecular energy redistribution. Specifically, for one important class of modifications, phosphorylation, peptide backbone cleavage is faster than the rate of phosphor group dissociation. This allows for observing fragment ions containing phosphor groups and potentially enables locating phosphorylation sites.[23]

ETD also poses some challenges to processing of mass spectral data. For example, precursor ions are often highly charged and fragment ions often contain products of hydrogen abstraction. In instruments with unit mass resolution, charge states of peptides cannot be determined from precursor ions' isotopic envelope. The charge state of a peptide is important to know in order to determine the precursor mass. Database search algorithms often use precursor mass in the identification procedure. Without a charge-state determination, all potential charge states are assumed. In the case of the ETD, these could be charge states up to (depending on proteolytic digest) +7 or higher. This could lead to large increase in the storage of mass spectral data and drastically increased requirements on computer resources for database searching and peptide identification.[24] Postsearch processing will also be affected by the large number of spectra. The importance of charge-state determinations for spectra obtained with unit mass accuracy has been well recognized.[24,25] Several previously developed algorithms to address the problem dealt with tandem mass spectra from CID fragmentation.[24−28] It has been observed that one of the predictive features for charge-state determination is the series of complementary ions. Other features such as ion density in different mass ranges, neutral losses of small molecular groups such as water, ammonia, and carbon monoxide, were suggested for use in a machine learning approach.[25,27] The charge-state determination problem in ETD is similar to charge-state determination from collision-induced dissociation.[25] The main differences between these two problems are due to differences between spectral features. ETD tends to produce peptides of higher charge states (almost no peptides with +1 charge) with a significant portion of +4 and +5 charge states. While in CID most of the obtained spectra are those from peptides between charge states +1 and +3. Often in ETD spectra, the charge-reduced intact precursor ions are very prominent. This feature is empirically known to be of diagnostic value in charge-state prediction.[29] In contrast, in CID spectra, this feature may often be absent.

In this work, we report on an algorithm to determine charge states of peptides. The algorithm uses a combination of signal processing and statistical analysis. The signal processing is based on correlation analysis of the spectrum. The correlation analysis is an often used technique in signal processing of mass spectral data.[30] In applications to tandem mass spectra, the technique has been used for peptide identification, for example. The well-known cross-correlation score determines peptide identification score as a correlation between preprocessed tandem mass spectrum and a model spectrum of an amino acid sequence.[4] Self-correlation of tandem mass spectra carries useful information as well, even though it has not been widely used so far. We show here that a modified self-correlation can serve as a method for charge-state determination of peptides from their tandem mass spectra. In its essence, the approach is similar to the charge-state determination technique based on the complementary ions.[24,26] However, the

(12) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399−406.
(13) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E.; Coon, J. J. *Anal. Chem.* **2007**, *79*, 477−85.
(14) Coon, J. J.; Shabanowitz, J.; Hunt, D. F.; Syka, J. E. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 880−2.
(15) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528−33.
(16) Chrisman, P. A.; Pitteri, S. J.; Hogan, J. M.; McLuckey, S. A. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1020−30.
(17) Liang, X.; Hager, J. W.; McLuckey, S. A. *Anal. Chem.* **2007**, *79*, 3363−70.
(18) Pitteri, S. J.; Chrisman, P. A.; Hogan, J. M.; McLuckey, S. A. *Anal. Chem.* **2005**, *77*, 1831−9.
(19) Kelleher, N. L.; Zubarev, R. A.; Bush, K.; Furie, B.; Furie, B. C.; McLafferty, F. W.; Walsh, C. T. *Anal. Chem.* **1999**, *71*, 4250−3.
(20) Zubarev, R. A. *Curr. Opin. Biotechnol.* **2004**, *15*, 12−6.
(21) Kelleher, N. L. *Anal. Chem.* **2004**, *76*, 197A−203A.
(22) Taylor, G. K.; Kim, Y. B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. *Anal. Chem.* **2003**, *75*, 4081−6.
(23) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2193−8.
(24) Sadygov, R.; Eng, J. K.; Durr, E.; Saraf, A.; McDonald, H. W.; MacCoss, M. J.; Yates, III, J. R. *J. Proteome Res.* **2002**, *1*, 211−5.
(25) Colinge, J.; Magnin, J.; Dessingy, T.; Giron, M.; Masselot, A. *Proteomics* **2003**, *3*, 1434−40.
(26) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327−42.
(27) Aaron, A.; Klammer, C. C.; Wu MacCoss, M. J.; Noble, W. S. *IEEE* **2005**, *8* (11), 175−85.
(28) Hogan, J. M.; Higdon, R.; Kolker, N.; Kolker, E. *OMICS* **2005**, *9*, 233−50.
(29) Hunt, D.; Syka, J. 2006. Personal communication.
(30) Owens, K. G. *Appl. Spectrosc. Rev.* **1992**, *27*, 1−49.

correlation analysis is more powerful, allowing modifications and adoption to tandem mass spectra. This procedure is highly specific and successful for high-quality spectra where many of the complementary ions are present.

For the majority of spectra, we predict charge states from features of tandem mass spectra. For this purpose, we train a linear discriminant on a curated data set of spectra from a mixture of nine known proteins. A set of features that play a role in charge-state determinations are inferred from the training set. The features are combined to produce a single discriminant score—Fishers' score. The coefficients for different feature contributions show that the charge-reduced ion properties are the most important properties for determining charge state. Distributions of Fishers' scores for different charge states are used to predict probabilities of charge states of spectra. The approach has been implemented in software named Charger.

METHODS

Sample Preparation and Data Acquisition. Nine protein digests were prepared as follows: samples of chicken lysozyme, horse cytochrome *c*, rabbit glyceraldehyde-3-phosphate dehydrogenase, chicken ovalbumin, bovine carbonic anhydrase, horse apomyoglobin, bovine β-casein, bovine serum albumin, and bovine α-lactalbumin were individually denatured in 6 M urea and then reduced and carboxyamidomethylated using standard procedures. Residual urea and reduction/alkylation byproducts were removed, and buffer was exchanged for 100 mM ammonium bicarbonate pH 8 with either dialysis (Lysozyme) or 5K VivaSpin ultrafilters (all other proteins). An aliquot of each reduced, alkylated protein sample was digested with LysC at an enzyme/protein ratio of 1:50 by mass at 37 °C for 18 h. Digests obtained with the same protease were combined and diluted with 0.1% formic acid to produce a mixture with the nine proteins (for the list, see Supporting Information).

A total of 49 protein digests were prepared as follows: 200 μL of 6.0 M urea in 100 mM ammonium bicarbonate, pH 8.0 was added to each vial of Universal proteomics standard (49 protein mix, Sigma). The protein mixture was reduced with dithiothreitol (3 mM, 1 h.) and alkylated with iodoacetamide (12 mM, 1 h.). Following the alkylation, the excess alklyating reagent was sequestered with an equal molar amount of DTT (1 h.). Solvent was exchanged for 100 mM ammonium bicarbonate, pH 8.0 using 200 μL (25 fmol/μL for each protein present). The alkylated proteins were proteolytically digested overnight at 37 °C with a 15:1 protein:protease ratio. The list of the proteins is in the Supporting Information.

LC−MS Analysis of Samples. A 2-μL aliqout of the 49-protein digest (50 fmol for each protein) or 9-protein digest was directly injected onto a MC10-C18W-150MS, C18 column, 10 cm × 150 μm (Micro-Tech Scientific, San Diego, CA) and separated with a 30-min 0−60% linear gradient (A, 0.1% formic acid; B, 100% acetonitrile/0.1% formic acid) at a flow rate of 1.0 μL/min using a Finnigan Surveyor HPLC equipped with a Micro AS and nanospray source (Thermo Fisher Scientific, San Jose, CA). The eluted peptides were analyzed by a Finnigan LTQXL with ETD (Thermo Fisher Scientific) operating in the alternating CID/ETD modes with data-dependent MS/MS detection.

Data Processing Model Description. Our algorithm employs two approaches for charge-state determination. In the first approach, the complementary ion information from tandem mass spectra is used. Simple arithmetic shows that the sum of the singly charged complementary ions (the ions that are generated from a dissociation of the same bond) equals to the mass of the precursor peptide plus two protons. Then if a series of singly charged complementary ions are present in a spectrum, it is expected that a transformation which sums every peak with all others will show a maximum at the precursor mass shifted by a mass of two protons. In general, to determine a sum of every peak with all others (two-set matching) will take about $N!$ operations, where N is the number of peaks in the spectrum. Some priori knowledge about the fragment ions and precursor mass range reduces the computational burden. However, it seems to us that the fastest and complete implementation is achieved by the use of fast Fourier transforms (FFTs). A self-correlation of mass spectrum, $S()$, will be (Wiener−Khinchin theorem[31])

$$\mathrm{Corr}(S, S) = F^{-1}(|F(S)|^2)$$

where F denotes a Fourier transform and F^{-1} is a reverse Fourier transform. The straight self-correlation of the spectrum will not lead to informative data on its precursor mass. It is due to the cancellation of real and imaginary parts at high-mass values. However, if one "splits" the transforms in a manner that separates correlations into the imaginary and real parts, they do not cancel each other out for large mass values. This result can be shown analytically from the correlation itself. Examples of the application will be shown in the next section.

We have compared the approach described above to another correlation-based technique employed for determining peptide molecular weights from their CID tandem mass spectra.[32] Venable and colleagues used cross-correlation of spectrum with its "reversed" copy.[32] To our understanding, this technique should be similar to self-convolution of the spectrum. Both techniques look to ways of optimizing the sum of complementary ions. We implemented convolution via FFT as well:

$$\mathrm{Conv}(S, S) = F^{-1}(F(S)^2)$$

To demonstrate how the correlation and convolution methods achieve precursor mass determination, we consider a spectrum that consists of two peaks (generalization to a larger number of peaks is straightforward). The spectrum then can be represented as

$$S = I_1\delta(m - m_2) + I_2\delta(m - m_2)$$

where m_1 and m_2 are mass-to-charge ratios and I_1 and I_2 are the corresponding intensities. Self-convolution of this spectrum at mass-to-charge ratio of M contains the following sum:

$$I_1I_2 \int (\delta(m - m_1)\delta(M - m - m_2) + \delta(m - m_2)\ \delta(M - m - m_1))\ \mathrm{d}m$$

Nonzero values of the integral for the first term (the second term

is analogous, only indices 1 and 2 are interchanged) follow from the system of equations:

$$\begin{cases} m - m_1 = 0 \\ M - m - m_2 = 0 \end{cases}$$

The solution of the system of equation for M is

$$M = m_1 + m_2$$

The above solution is the reason for the self-convolution of a spectrum to peak at the mass shift corresponding to precursor mass (+2 amu). An analogous expression can be obtained for the self-correlation function by separating real and imaginary parts of the Fourier transform of the original spectrum.

We implement only one of the two FFT-based methods as it is computationally too expensive to do many FFTs for every candidate charge state (potentially four FFTs per every charge state from +2 to +7). In our applications, a correlation-based method (without reversing spectrum) performed better than convolution (example provided in the Supporting Information). Therefore, we choose to implement the correlation-based method in Charger.

In practical applications, the correlation-based methods work well for spectra with a large number of complementary ions. In real spectra, however, this is not always the case. Often there are not enough complementary ions in the spectrum or noncomplementary ions can account for the majority of the total ion current (TIC) in the spectrum. This problem is especially astute for +2 charged peptides. Fragmentation efficiency of +2 charged peptides in ETD is occasionally not very high (recently developed supplemental activation technique addresses this problem).[13] At the same time, charge states of these peptides can be relatively simply determined from a pattern classification approach as they exhibit a specific distribution of precursor and fragment ions. Another problem with correlation methods is that for their applicability they require mass scan ranges that are at most 1.5 times smaller than the true precursor mass. If the precursor mass is too large, the complementary fragment ions will not fit into the mass range. In this case, even if the ion series are present they will not be complementary.

For above-noted reasons, we employ a statistical pattern recognition approach to determining charge states where the correlation method is either not applicable or fails. We choose a linear discriminant analysis as our method. We use a spectral data set of peptides from digest of nine protein mix in training the LDA classifier. From this spectral data set, we determine features that are the most relevant for charge-state classification. Then a training program extracts these features from all spectra. The extracted features are input to the linear discriminant classifier (model described below). The charge states of peptides in the training data set were determined from correct, high-quality database matches to the known proteins and validated manually

(31) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++. The art of scientific computing,* 2nd ed.; Cambridge University Press: Cambridge, 2002.

(32) Venable, J. D.; Xu, T.; Cociorva, D.; Yates, J. R., III. *Anal. Chem.* **2006**, *78*, 1921−9.

Scheme 1

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & & x_{np} \end{pmatrix} \begin{matrix} Class1 \\ \hline Class2 \end{matrix}$$

if deviations appeared in feature distributions. To train the classifier, we choose a set of features that we empirically knew had strong impact on distinguishing a particular charge state.

For illustration purposes, we show data of +2 charged peptides. Corresponding data for some other charge states are shown in the next section. For +2 charged peptides, we have determined five features: fraction of the TIC explained by +2 charge state, ratio of the charge-reduced precursor ion abundance to the highest abundance in the spectrum, ratio of the neutral loss peak (ammonia loss from the charge-reduced precursor) to the highest abundance ion in the spectrum, and the ion current of ions in 150 amu mass window following right after the precursor mass. The choice of the features matches our observations on the characteristics of the ETD spectra for +2 charged peptides. Thus, normally it is expected that spectra of +2 charged peptides will have all strong ions up to the charge-reduced precursor. Some ions are observed at mass-to-charge ratios higher than the precursor mass. These ions have been suggested to belong to adduct ions. For true +2 charged peptides, these ions are expected to be in much smaller abundance than the fragment ions or intact precursor ions. Data of the features from all spectra in the training set fill in the data matrix, Scheme 1. Figure 1 shows the scatter matrix view of the data matrix for true and false assignments of +2 charge. In this and other scatter matrix views, true (blue circle) and false positive (red diamond) distributions are colored differently. Also, for the purpose of discernible visualization, we restrict the number of features in the plot to three only. These are the most specific features for +2 charge-state determination. They are TIC explained by +2 charged precursor, the sum of precursor ion intensities, and the ion current of ions in the 150 amu mass range following right after the precursor mass. The expressions for computing of the features follow.

The TIC is defined as the sum of the all ion intensities in a spectrum S:

$$\mathrm{TIC} = \sum_{i=1}^{N} S(i)$$

where N is the number of peaks in the spectrum. The explained ion current is defined by assuming a charge state for the spectrum and summing all ion intensities up to and including precursor mass of the corresponding charge state. For example, assuming a charge state $+n$, the explained ion current (EIC) will be

$$EIC = \sum_{i=1}^{Mz(S(i)) <= n*(Mz-1)+1.0} S(i)$$

where Mz is the mass-to-charge ratio of the original precursor

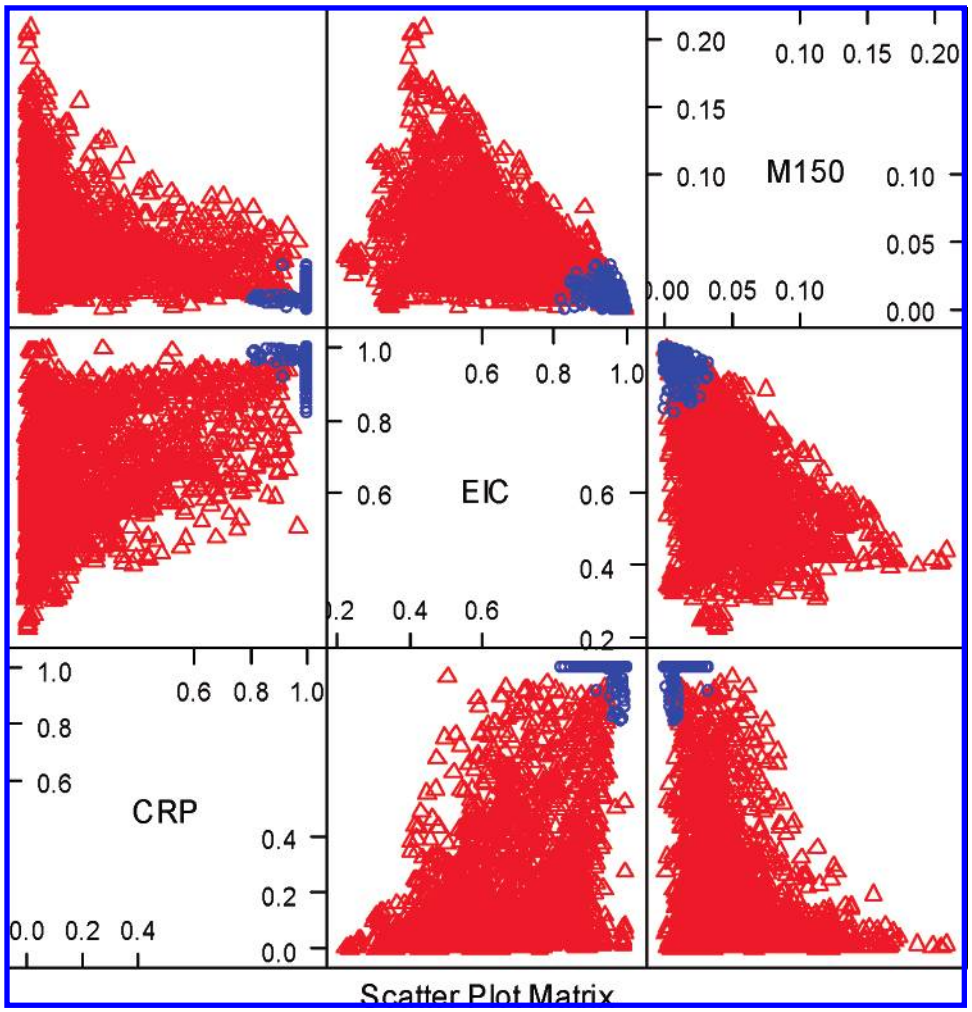


Figure 1. Scatter plot matrix of three most important features for determining charge states of +2 charge peptides. The blue circles denote true positives; red diamonds denote false positives. All variables are normalized to 1. For definitions of CRP, EIC, and *M*150 see the Methods section. All figures in this paper were made in R environment.[35]

and $Mz(S(i))$ is the mass-to-charge ratio of the ith ion in the tandem mass spectrum. The ion current in the 150 amu mass interval immediately following the precursor mass, $M150$, is defined as

$$M150 = \sum_{Mz(S(i))>n^*(Mz-1)+1.0}^{Mz(S(i))<=n^*(Mz-1)+151.0} S(i)$$

The choice of the mass interval width is related to the mass of an average amino acid, 110 amu. On average, more than two fragment ions are expected to fall into this mass range. The ion current of charge-reduced precursor ions (CRP) is

$$\mathrm{CRP} = \sum_{k=n-1}^{1} \sum_{i}^{n^*(Mz-1)/k-1.0<=Mz(S(i))<=n^*(Mz-1)+2} S(i)$$

We have considered other parameters such as neutral loss (ammonia molecule) from +1 and +2 reduced intact precursor ions (except for +2 charge assignments), spectrum sparseness parameter, and what we call a mass balance—ratio of precursor ion abundance to the TIC. The latter two parameters turn out to be important for filtering out low-quality spectra. In our implementation of this ETD data preprocessing application, these spectra are assigned to +2 charge state by default.

A discriminant function generates a single direction (by a linear transformation of the feature space), which separates two classes (Scheme 1), true and false positives, as far as possible in one dimension. This is achieved by a transformation, $\boldsymbol{w}$, that maximizes the Raleigh quotient,[33] $J(w)$:

$$\arg\max\left(J(w) = \frac{|m_1 - m_2|^2}{{S_1}^2 + {S_2}^2} = \frac{w^t S_B w}{w^t S_w w}\right)$$

where m_i are the class means, s_i are the class scatter matrices (i = 1,2), S_W is the sum of the class scatter matrices, and S_B is the between classes scatter matrix.

LDA generates loading coefficients for every feature, which is used in the model. Linear discriminant score, also called Fisher's score, is a dot product of the features with the loading coefficients.

(33) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley: New York, 2001.

$$F = \sum_{i=1}^{p} c_i x_i$$

where x_i are the features, p is the number of features (Scheme 1), and c_i are the coefficients to be determined from the data set by the linear discriminant. We study the distributions of the Fisher's scores for true and false positive charge states of spectra from the training data set. This information, together with the loading coefficients, is used to predict charge states of peptides in new experiments.

The use of the linear discriminant classifier is suitable for charge-state determination. Tandem mass spectra of electron-transfer reactions may contain, just like some CID spectra, a mixture of spectra of two coeluting peptides with different charge states. In these cases, it is not always possible a priori to infer which peptide with its charge state has a better chance of being identified in the database search. Therefore, we assign two charge states to these types of spectra. As we will show below, analysis of loading coefficients from the LDA allows one to infer whether the spectrum is a mixture of spectra of two peptides with different charge states. In these cases, LDA allows us to accurately assign both charge states. Note that in principle it is possible that a spectrum could be a mixture of more than two peptides. But this would be a rare event because of the separations in chromatographic domains and is not considered in our algorithm.

To summarize, the general work flow of the classifier is as follows, Scheme 2: a spectrum is first tested for +2 charge assignment. If it is confirmed to be +2 with a high Fisher score, the program creates +2 charged spectrum and terminates. If it is determined that the spectrum is not a +2 or a +2 with a lower score, the next step is to test +3 charge. The first test is done with the correlation analysis (mass scan range permitting). If the test is successful, the program assigns charge +3 and terminates. When the test fails, then the Fisher score is computed. If the +3 charge can be determined by Fisher score then program terminates. Otherwise, the process is continued untill candidacy of +7 charge is checked. If no charge is determined uniquely, then the two most likely charge states are generated.

The Charger algorithm is written with C++ in .NET2 environment of Windows XP. The input and output of the program are in the dta[34] file format.

RESULTS AND DISCUSSION

Figure 2 shows a tandem mass spectrum of +4 charged horse cytochrome *c* peptide "KTEREDLIYALK". The spectrum exhibits some typical characteristics of ETD tandem mass spectra, mainly prevalence of reduced charge states of intact precursor at 371 Th original +4 charged precursor, 494 Th +3 charge, 741 Th +2 charge, and 1481 Th +1 charge. The spectrum is preprocessed before it is used in "split" correlation. The results of the correlation are shown in Figure 3. The peak at zero is the zero-shift self-correlation of all product ions in the spectrum. The negative peak at 1483 Th corresponds to the precursor mass of the peptide (the details of the correlation around this mass are shown in Figure S4, Supporting Information). Once the precursor mass is known, charge-state determination becomes trivial as the precursor mass-to-charge ratio is also known. Analogues result is obtained in convolution analysis as well, Figure 4. In this case, the overall peak of the convolution corresponds to the precursor mass. A method similar to our convolution analysis approach has previously been shown to accurately determine precursor mass of peptides from their CID tandem mass spectra.[32] However, in general, we find that for ETD spectra the split correlation method works more often than the convolution method. An example of this is provided in the Supporting Information to this paper. The Charger algorithm implements the split correlation approach, as using both methods would have been computationally expensive.

For successful determination of charge states using correlation-based analysis spectrum needs to have enough complementary ions and corresponding mass scan range (not less than ~0.60% of the precursor mass). In practice, for many spectra, especially for +2 charged peptides (due to poor fragmentation in ETD), it is not possible to determine their charge states via correlation analysis. For other spectra, mass scan range could be the limiting factor. Charge states for these spectra were determined from a statistical classifier. To train the classifier, we used a list of spectra whose charge states had been manually curated. This data set contains about ~5700 spectra obtained from a digest mixture of 9 known proteins. Of these, 2400 spectra are +2, 2500 spectra are +3, 530 spectra are +4, 254 spectra are +5, 27 spectra are +6, and 9 spectra are +7 charged. The machine learning algorithm extracts predetermined features from the validated data set and generates coefficients for the linear discriminant loading coefficients. The coefficients are stored and later used to determine a charge state of a new spectrum. Overall we have used 10 features for every charge-state specification. The computational overhead of computing redundant coefficients is not high. We have also experimented with different configuration of features. We found that only few features are actually important for charge-state classification. Only these features are used for later predictions.

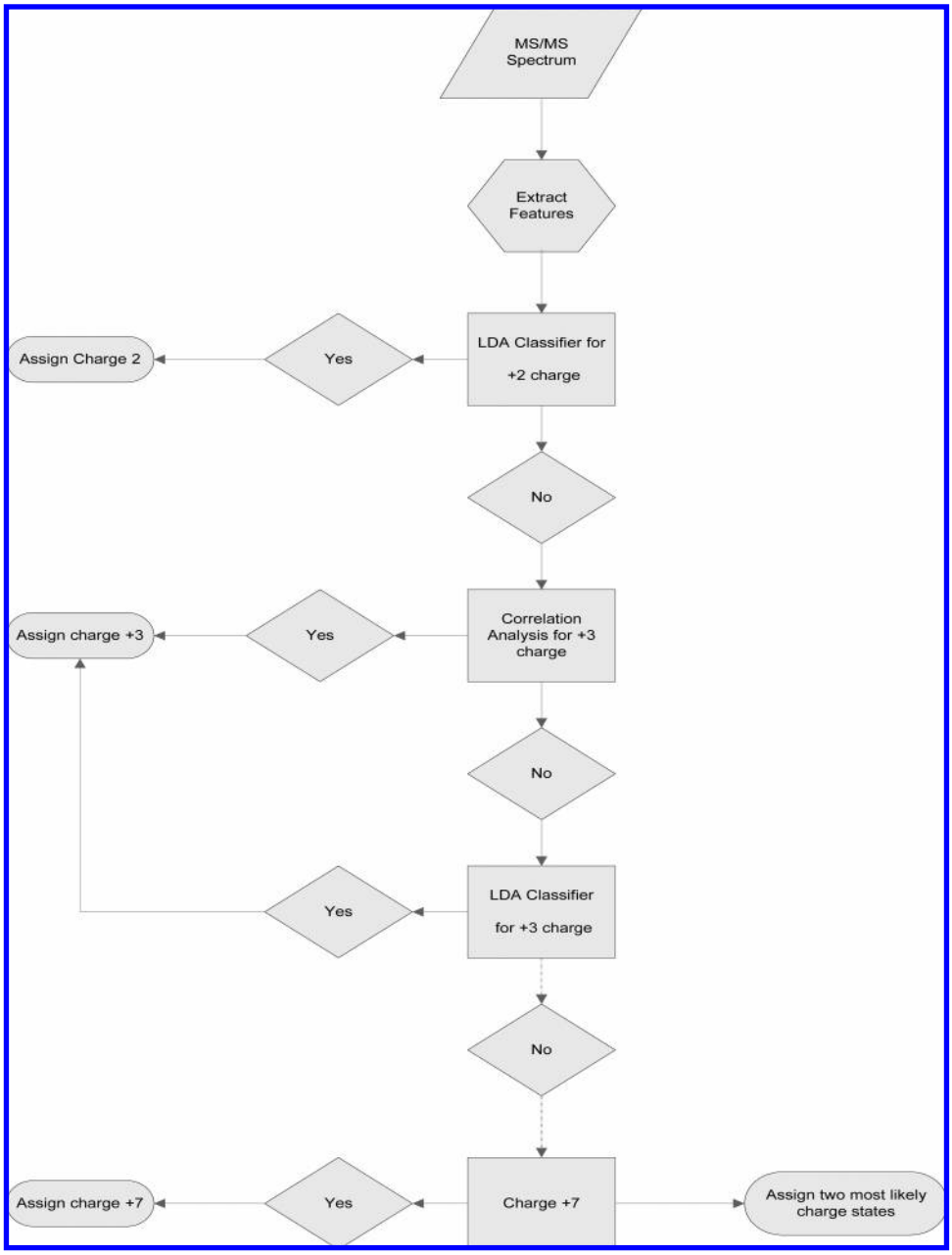
One of the main characteristics of ETD spectra is the observation of charge-reduced precursor ions. This is especially true for the +1 and +2 reduced ions if they fall within mass range and they are present. Figure 1 shows the frequency of the (normalized) CRP in our training data sets. For true positive +2 charged peptides (blue circles) in ETD spectra, their precursor ion current will be highest and normalize to 1.0. Another feature presented in Figure 1 is the percentage of the TIC that is explained by +2 charged species, EIC. For false +2 spectra (red diamonds), this feature will most of the time be less than complete as there are expected to be many fragment ions past the twice the original precursor mass-to-charge ratio. For true +2 charged peptides, this value, on the other hand, most of the time is close to 100%. As seen from the figure, the measures serve as good diagnostics for charge-state determination. LDA analysis confirms this observation. The loading coefficients are presented in Table 1.

The coefficients show that the most important features in determining +2 charged peptides are the normalized reduced ion intensities, CRP, and the portion of ion current of ions passed the proposed +2 charge precursor mass, *M*150. Clearly, these

(34) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., III. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2162−68.

(35) R: A language and environment for statistical computing. [2.5]; Vienna, Austria, 2006.

Scheme 2



coefficients enter into Fisher's score with different signs. Thus, the coefficient of CRP is large and positive, the coefficient of *M*150 is large and negative. The other important coefficients are the normalized abundances of singly charged precursor ions, which is relatively large. We found that the other two features in our study, normalized +1 precursor ion intensity and intensity of ammonia loss from this ion, are only marginally important. Both of these coefficients are positive, thus confirming empirical observations that these ions play a diagnostic role in determining precursor charge states. In the cases when charge-reduced species are absent (supplemental activation), the charge state is determined from EIC. The smallest charge state for which EIC is equal to 1 (100% of total ion current and if the proposed precursor mass is less than mass scan range) is assumed to be the charge state.

Table 1 also presents loading coefficients for charge states up to +5. We did not have enough statistics to train the model for +6 and +7 charged peptides. Based on the limited information, the training program we have generated heuristic rules for these charge states. The main feature for them is the CRP value.

Figure 5 depicts the distribution of Fisher scores for true (blue color) and false (red color) assignments of +2 charge. The peak for true assignments is a score of ~8. From Table 1, this corresponds to a case when all of the TIC is explained by the +2 charge and either +1 reduced precursor or its neutral loss is present. The false assignments of +2 charge state generate a Fisher score that most of the time is less than 6. Again, from Table 1, it is seen that this will most likely happen only when EIC is less than 1.0 or *M*150 is a large value. The difference between these cases is important since if EIC is 1.0 and *M*150 is a large value as well, it is an indication that the spectrum is a sum of spectra of coeluting peptides (see Figure S6 in Supporting Information for an example). Thus, when the Fisher score is in

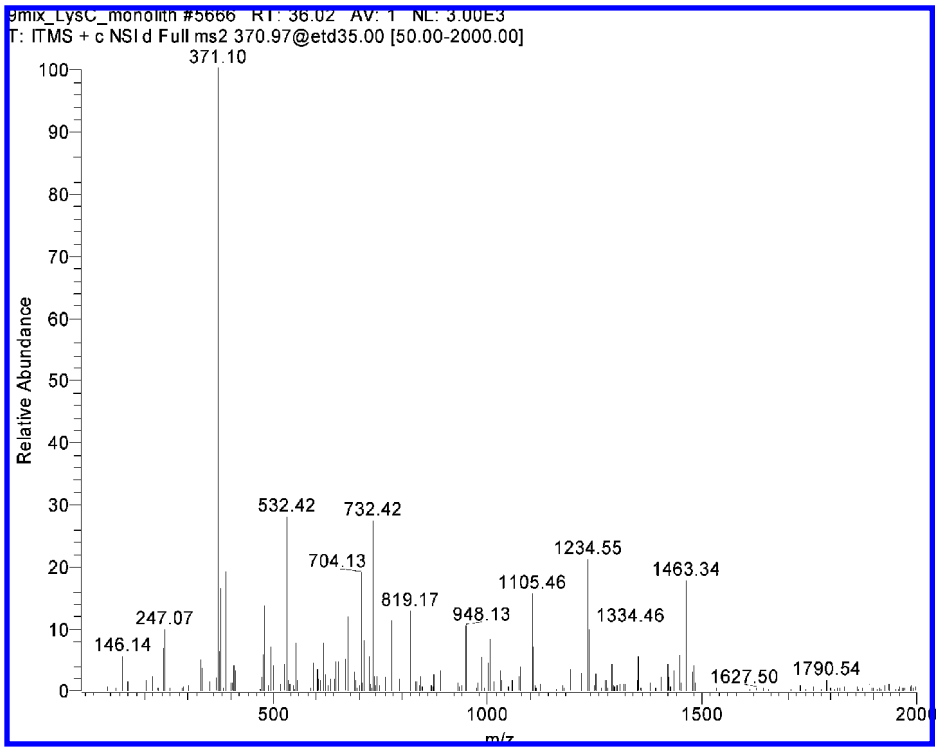


Figure 2. ETD tandem mass spectrum of +4 charged "KTEREDLIYALK" peptide of horse cytochrome *c*.

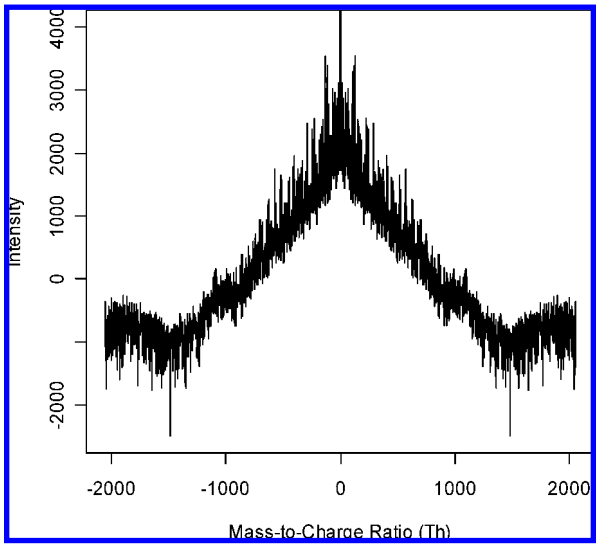


Figure 3. Correlation analysis of +4 charge spectra of "KTEREDLIYALK" from Figure 1. The minimum of the correlation corresponds to the peptide precursor mass plus mass of two protons. For good-quality spectra (containing complementary ions), the correlation analysis-based charge/mass determination is preferable, because this method is highly specific.

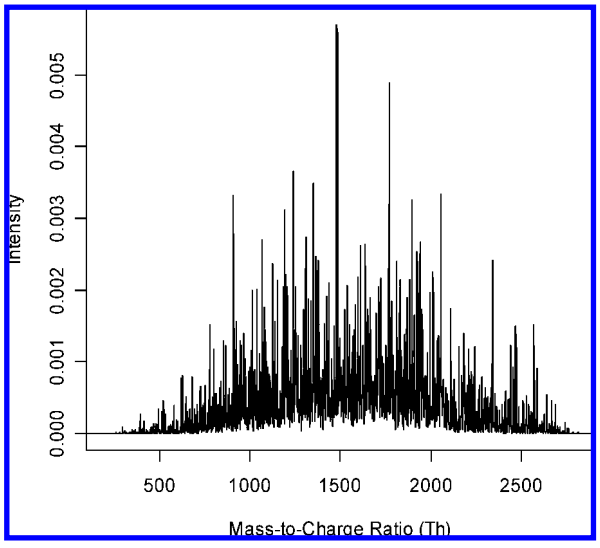


Figure 4. Self-convolution of the spectrum from Figure 1. The maximum of the convolution, in this case, uniquely determines the precursor mass of the peptide. For the spectrum in Figure 1, both correlation and convolution methods work equally well. However, in general, in our experience correlation methods work in many cases when the convolution fails.

the bordering region, we check to determine whether the spectrum is a mixture of +2 charge and a coeluting peptide with a different charge state.

Figure 6 is a scatter plot matrix for true and false assignments of +3 charge states. The same features (as for +2 charge) are used since they are computed to be most important. The +3 charge contributions from *M*150 are even higher. Fisher score distributions for +3 charge analysis are depicted in Figure 7. There is a good separation of the true and false assignments. The loading coefficients for +4 and +5 charge states are presented in Table 1 as well.

The reduced precursor ions are important diagnostic measures for determining charge states of the original precursor. When they are present and prominent in a spectrum, the charge-state determination is largely simplified. However, there can be spectra that are of good quality and readily identifiable in a database search, but they either miss the reduced precursor ion peaks or the reduced ions show very small abundances—smaller than

Table 1. Loading Coefficients Obtained in Training LDA Model.

spectra	ion current[a] CRP	EIC	+1 reduced precursor/loss[b]	+2 reduced precursor/loss[b]	*M*150
+2	6.53	1.261	0.5/0.3	-/-0	−4.68
+3	9.58	4.73	0.12/0.34	0.7/0.51	−15.68
+4	6.9	5.22	0.93/0.61	0.58/0.12	−10.94
+5	7.3	6.4	0.82/0.3	0.63/0.41	−7.5

[a] Ion current is the ion current of charge-reduced intact precursor ion species. [b] Loss refers to neutral loss of ammonia molecule from the intact precursor.

fragment ions. In these cases, separation of reduced precursor ions from fragment ions becomes problematic. It is possible that a certain amount of the total ion current of the ETD spectrum is due to the ion adducts (for an example, see Figure S5 in Supporting Information). Use of the ion current measure only would not produce a correct charge state. However, the correlation analysis correctly assigns the charge as +3.

As can be seen from Table 1 and Figures 5 and 7, no single feature classifies charge states uniquely. Rather, a combination of features is able to predict a charge state.

One of the problems with determining charge states when using reduced ions is the ambivalence of several overtones for +2, +3, +4, and +6 charge states. Thus, a false +6 charge candidate will explain both +2 and +1 overtones of a true +3 charge, as well as +1 overtone of a true +2 charged precursor. An analogous relationship exists between overtones of +2 and +4 charge-state candidates. As seen from Table 1, LDA does not solve this problem. There is not enough statistics for +6 charge state, but coefficients of +2 of +4 charged peptides are not too different from the coefficient of a +1 charge precursor of +2 charge states. A different analysis is needed to accurately separate (if possible) contributions from these ions.

We applied Charger to a test data set of spectra from a mixture of 49 proteins. There are 6910 spectra in the data set, Table 2. Charger assigns 4964 + 2, 1993 + 3, 206 + 4, 19 + 5, and 66 + 6 charge states and generates the corresponding spectra. The number of charge states assigned in split correlation analysis was 156. The rest of the assignments were made by the linear discriminant classifier. The majority of assignments made by split correlation method are +3 and +4 charged peptides.

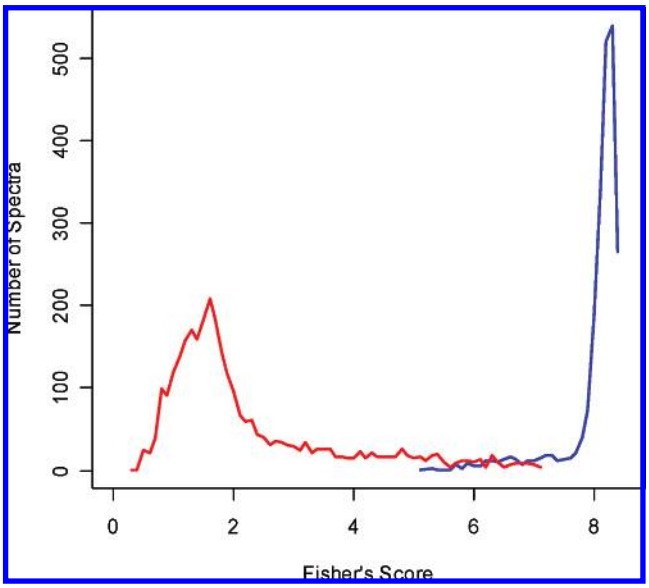


Figure 5. Distributions of Fisher's scores for true (blue) and false (red) positive assignments of +2 charge states.

As expected, the majority of spectra are assigned as +2 charge. This includes low-quality spectra and true +2 charged spectra. Another factor for the large number of low charge states (<+4) is the enzyme used for digesting this mixture—trypsin. The artificially large number of +6 charged spectra is due to ambiguity in determining this charge state. In our experience, peptides with +3 and +4 charge states often are difficult to differentiate from +6 charge peptides when the precursor mass-to-charge ratio is large for the mass scan range. For the sake of reducing possible cost to peptide identification, we utilize relaxed criteria on determining +6 charge states. In our experience, the computational overhead associated with this is not high. In this case, for example, less than 1% of all spectra is classified as +6 charged. There were no +7 classified spectra in the 49-protein mix.

For some spectra, the charge state cannot be determined uniquely. Alternatively, they could be a mixture of coeluting peptides with different charge states. In these cases, two copies of the spectra are created—one corresponding to each charge state. In the test data set, there were 158 (~2%) such spectra.

We have used the charge states assigned spectra in a database search to identify peptides in the mixture. This task was aided by the knowledge of the content of the protein mixture (Supporting Information). The search was carried with a newly developed database search algorithm, which will be described in detail elsewhere. The protein database was uniprot-swissprot (10/2006). The results are summarized in Table 2. The model accurately predicts charge states with very little loss in peptide identifications, less than 2%. The program (Charger) affords substantial time savings, ~350%, in database search time. Note that even though split correlation assigned charge states of 2% of all spectra, for ~30% of confidently identified spectra, charge states were determined via split correlation. Also, it is important that the split correlation assigns a single charge state to spectra to which the LDA method potentially may assign two charge states.

To confirm the above results, we have also searched the 49-protein mix data with another database search algorithm—SEQUEST.[4] At first the data were searched by assuming all charge states from +2 to +6 for every spectrum. Then the search results were sorted, and we kept only spectra that matched any of the 49 proteins with Xcorr and delta Cn larger than 2.5 and 0.1, respectively. This resulted in 350 spectra with unique assignments that we assumed are correct hits. The charge states of these hits served as a basis for our test of the Charger's performance. Next,

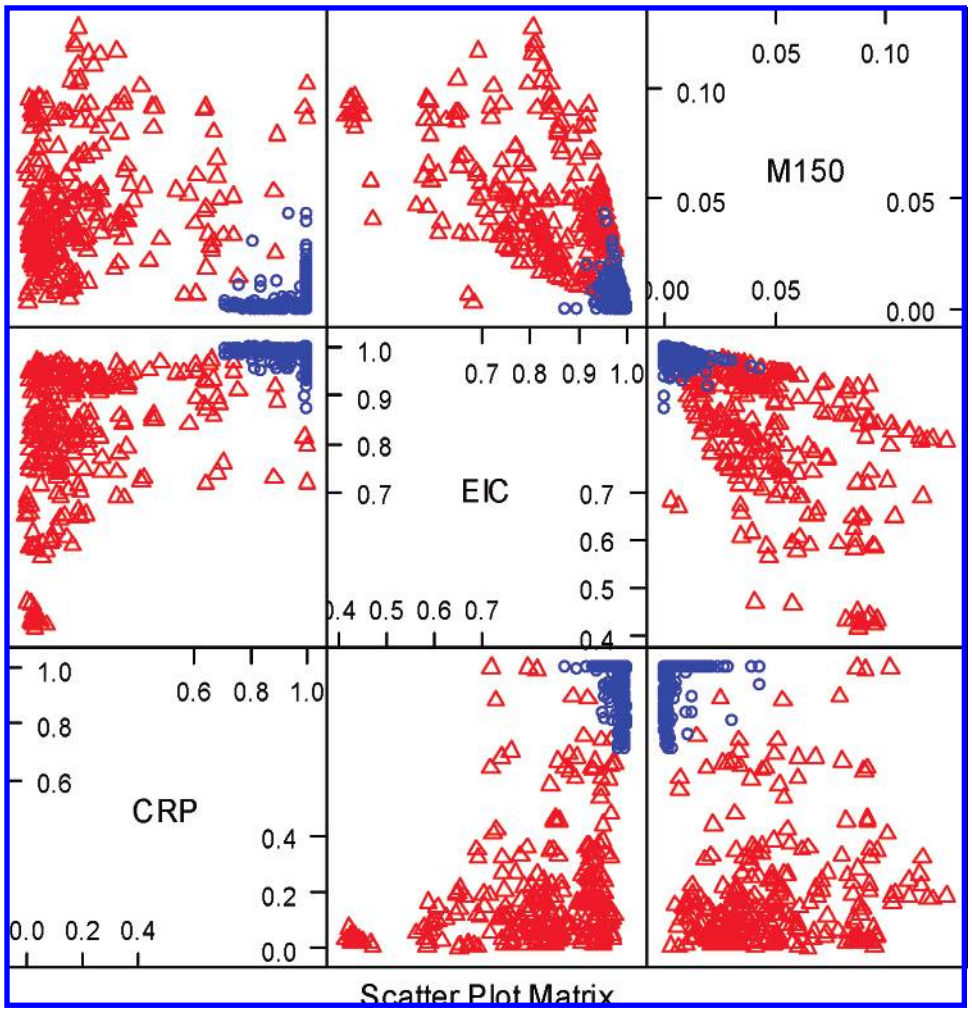


Figure 6. Scatter plot of the data matrix for training to determine features of +3 charged spectra. The notations are similar to ones used in Figure 1 (the blue circles denote true positives; red diamonds denote false positives).

we determined charge states of peptides of the 49-protein mix with Charger. From the comparison of the charge states of the 350 spectra, we generate a receiver operating characteristics (ROC) curve of Charger's performance, Figure 8. For 10 spectra out of 350, Charge made wrong charge-state assignments. These are mainly +4 and +5 charged peptides coeluting with +2 charged species. In these spectra, the charge-reduced precursor of +2 charge species tend to be a dominant peak. Charger tends to assign +2 charge states to these spectra. One of way of solving this problem would be relaxing the criteria for assigning two charges states. This will generate more spectral copies but increase the sensitivity of charge assignment.

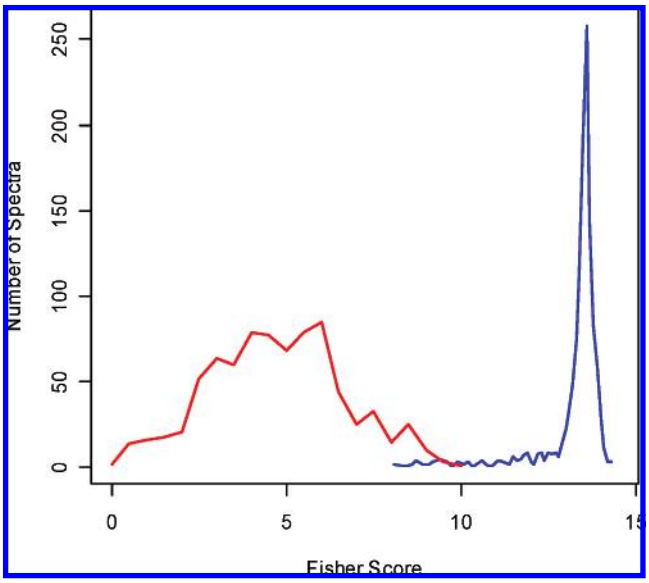


Figure 7. Fisher's score distributions for true and false positive +3 charges.

CONCLUSION

We present our algorithmic development for charge-state determination of peptides from their electron-transfer dissociation tandem mass spectra. We use two approaches. At first, we determine charge states of peptides via self-correlation of mass spectra. The method optimizes the sum of the complementary ion masses and determines precursor mass from which the charge state is inferred. The correlation analysis is highly effective for spectra where many complementary ions are present.

When the self-correlation test fails, the program uses a linear discriminant analysis to predict charge states. The program uses a differentiated approach for every charge state. Spectra of +2 charged peptides are relatively simple and easily characterized. Four features are most sufficient for distinguishing +2 charged

Table 2. Results of Charger in Database Search As Applied to 49-Protein Mixture[a]

	number of spectra	database search time[b] (min)	no. of ID (probability $<10^{-7}$)
without charge determination			
+2/+3/+4/+5/+6	6910/6910/6910/6910/6910	900	307
assigned charge states			
+2/+3/+4/+5/+6	4964/1993/206/19/66	250	302

[a] Originally there are 6910 spectra in the data set. [b] Database search times refer to a search on a single 3.4 GHz Pentium 4(R) CPU computer running under Windows XP.

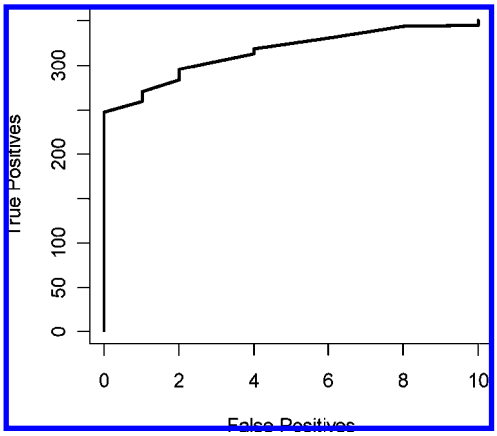


Figure 8. A ROC of charge-state determination by Charger for 49 protein mix data.

spectra from other charge states. Two of these parameters determine the ratios of the explained total ion current to the total ion current of the spectrum, and the other two are related to the intensity of the reduced precursor ion and neutral loss from it. For higher charged spectra, we tested some additional features such as ion current inbetween reduced intact precursor species, ratio of precursor intensities to ion currents, and spectral density. No single feature is found to be deterministic alone. Results show that rather a combination of several features can serve as a diagnostic measure. We found that the most important features for charge-state determinations are explained total ion current, normalized intensities of charge-reduced precursor ion species, and neutral losses derived from them. We trained our statistical model on set of spectra obtained from a mixture of proteins of known identity.

ACKNOWLEDGMENT

We are thankful to Don Hunt and members of his laboratory at the University of Virginia for their discussions on interpreting ETD spectra. We thank our colleagues from the Bioworks team for their work on incorporating Charger into the Bioworks. We are thankful to our colleague Roger Biringer for preparing samples of peptides whose mass spectra were used in this work.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review June 22, 2007. Accepted September 24, 2007.

AC071332Q