

# Data Processing and Database Search Models for Tandem Mass Spectra Obtained via Electron Transfer Dissociation

Zhiqi Hao; Rovshan Sadygov; Roger G Biringer; Terry Zhang and Andreas FR Hühmer

Thermo Fisher Scientific, San Jose, CA, USA

**Thermo**  
SCIENTIFIC

## Overview

**Purpose:** Determine charge states of peptides from their tandem mass spectra obtained in electron transfer dissociation. To improve identification of peptides from their tandem mass spectra.

**Methods:** Use signal processing and statistical analysis to determine charge states of peptides from their tandem mass spectra. Signal processing transforms the spectra to maximize the sum of the complementary ions. When applied to the whole spectrum, the sum becomes an indication for the position of the precursor mass. Statistical analysis uses Linear Discriminant Analysis (LDA) to learn spectral patterns of different charge states from spectra of peptides with known charge states.

**Results:** The algorithmic approach has been implemented in a program that we refer to here as *Charger*. *Charger* can determine charge states of peptides as high as +7. For spectra with abundant complementary ions, the charge state is determined from signal processing. For spectra for which complementary ion information is insufficient or information is beyond the scan range, charge state is determined from the statistical machine trained on the spectra of known charge states.

## Introduction

Mass spectrometry in combination with liquid-chromatography (LC-MS) followed by a database search is an efficient methodology for identification and characterization of proteins from biological samples. Proteins are first digested using a site-specific enzyme. LC separates a peptide mixture based on their hydrophobicity. Tandem mass spectrometry selects a single mass-to-charge ratio from eluting peptides and subjects it to dissociation. The information-rich mass spectra of subsequent peptide fragments are used in database searching for the identification of peptides. Collision induced dissociation (CID) is a conventional methodology used for peptide fragmentation. The recent introduction of electron transfer dissociation (ETD) is a significant advance in the analysis of peptides and proteins. ETD reactions primarily break N-C<sub>α</sub> bonds and create c and z ions. ETD of highly-charged peptides is thought to provide more complete information on the primary structure of the peptides, including positions of modification sites.

In the instruments with unit mass resolution, charge states of ETD spectra cannot be determined from isotope distribution of precursor ions. Uncertainty of the charge states is a significantly larger challenge for ETD data than for CID data because, unlike CID spectra where ions up to +3 charge constitute the absolute majority of spectra, in ETD a substantial portion of precursor ions are charged higher than +3. Determination of the charge states of precursor ions are important because charge state and precursor ion mass-to-charge ratio are used to identify peptides from a database search. In the absence of the charge state knowledge, all practical charge states are assumed likely and corresponding precursor masses are then searched in the database. This creates problems for the search, itself, and for post-search processing, as the effective number of spectra used for the database search increases. The increase in the size of data set also creates additional demands on CPU processing power and hard disk space.

In this work we present an algorithm to determine charge states of precursor ions from their ETD tandem mass spectra. The algorithm uses signal processing and statistical analysis to determine charge states. Signal processing transforms spectra in a manner to maximize the pair-wise sum of the fragment ions. For spectra with abundant complementary ions, the sum will peak at the precursor ions' mass, providing the mass-to-charge ratio that precisely determines precursor charge.

If the precursor charges cannot be determined by signal processing, it will be predicted by a classifier based on linear discriminant analysis (LDA). Several spectral features serve as predictors for LDA. The classifier is trained on a spectral data set with known precursor charge states. The classifier assigns a score – the Fisher's score – to every spectrum with an assumed charge state. Based on the Fisher score distribution for the given charge and the current value of the score, a decision is made as to whether the assumed charge state is correct or not.

## Methods

ETD reactions in general tend to cause the fragmentation of precursors with charge states +2 or higher. The proportion of different charge states depends on concrete experimental conditions such as the type of enzyme used to digest proteins. It has been suggested that the fragmentation reaction rate is proportional to the square of the charge, thus promoting reactions of highly charged peptides. Our algorithm is trained to work with precursor charge states up to +7.

We employ a signal processing procedure to determine charge states of peptides of which the majority of fragment ions fit into the mass scan range. The main expectation of this approach is that masses of complementary fragment ions of a precursor will add up to the precursor mass + 1 amu. In general, to determine the sum of every peak from all others in a set of N numbers will take NI operations. Some prior knowledge about the precursor ion mass range and fragment ions may help to reduce this complexity. We propose that the fastest and most complete solution of this problem is done by implementing fast Fourier transforms. The autocorrelation of the tandem mass spectrum S will be:

$$\text{Corr}(S, S) = F^{-1}(F(S) * F(S)^*)$$

For our purposes we use a modified autocorrelation. The modification allows for observations of pair-wise sums of fragment ions with large mass differences.

The autocorrelation based method is very selective and creates almost no false positives. However, for many spectra there are not enough complementary ions for confident charge state determination via this method. In these cases we use a linear discriminant analysis trained on a data set with spectra of known charge states for further classifications. We have determined several spectral properties that help to identify precursor charge states. The goal in LDA is to determine loading coefficients of these features from a spectra of known charge states. The spectra are classified according to the Fisher score, F:

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

where  $x_i$  are the features, and  $c_i$  is the transformation matrix to be determined from the training data set by linear discriminant analysis. If we denote the data matrix as X, and the transformed matrix as Y, then we seek such a transformation matrix w that separates true and false assignments as much as possible. In LDA analysis this is achieved analytically by maximizing the Raleigh quotient, J(w),

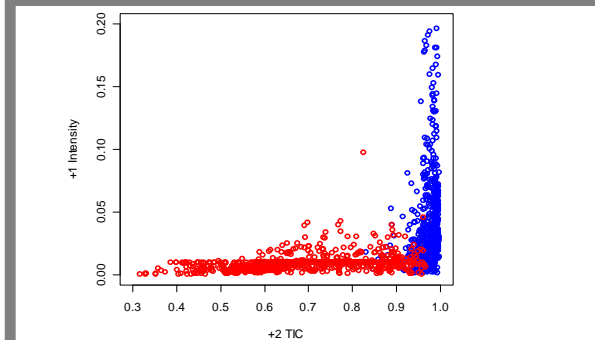
$$\arg \max \left( J(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w} \right)$$

where,  $m_1$  and  $m_2$  are the means of the groups and  $S_B$  and  $S_W$  are the between groups and within groups variances. The transformation matrix, w, is the maximum of the ratio of the between groups variance to the within group variance. The solution is obtained as:

$$w = S_W^{-1}(m_1 - m_2)$$

The groups in this case are construed from true and false charge state assignments. We determine the transformation matrix elements and loading coefficients for every charge state. Several features such as total ion currents, rankings of the reduced precursor ions and their neutral losses have been tested for use as predictors in the linear discriminant. Some of the features are common for all charge states while others are specific for a given charge state. Values of loading coefficients suggest whether a given feature is important for charge state discrimination. An example of two features (total ion current, TIC, and intensity of reduced precursor ion) for +2 charged precursor ions is shown in Figure 1.

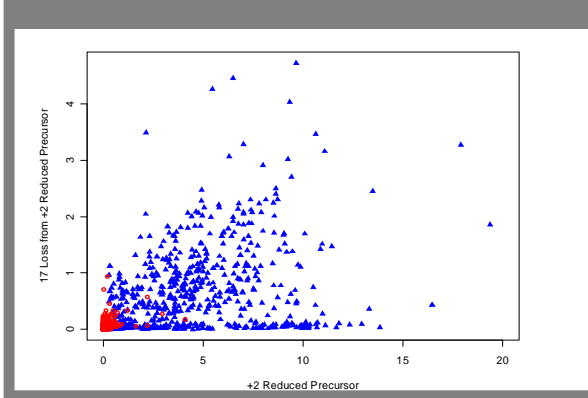
FIGURE 1. Scatter plot of data matrix generated from features from false (red) and true (blue) assignments of +2 charge.



As seen in Figure 1, for most of the false assignments of +2 charge, the corresponding total ion current (TIC) up to the reduced precursor ion is less than 90% percent of the total ion current. For true charge state assignments this value is above 90%. This feature, together with the ratio of the reduced precursor ion to the highest fragment ion, serve as good indicators for

+2 charge state classification. Two other pairs of features used for charge state classifications are reduced ion and neutral losses from this ion. The distribution of these features for true and false charge state assignments from the training set is shown in Figure 2.

FIGURE 2. Feature distributions from the training data set.



Overall we use several features for every charge state. As we will show below, the value of the corresponding loading coefficients indicates the level of importance of that specific feature for the classification. For +2 charged peptides, the LDA computation gives the highest positive loading coefficient (6.65) to the portion of spectral TIC explained by the charge model. The most negative coefficient is the TIC immediately following the reduced precursor of the +2 charged peptide – the model penalizes +2 spectra highly for any ions beyond precursor mass. To our surprise, the value of the reduced precursor itself does not have a high loading coefficient. One explanation for this is that ions with twice the original precursor ion are observed in +4 and +6 spectra. Therefore this ion on its own does not serve a diagnostic role. *Charger* determines loading coefficients for all charge states up to +7 from the training set. The data for other charge states are not shown here.

## Results

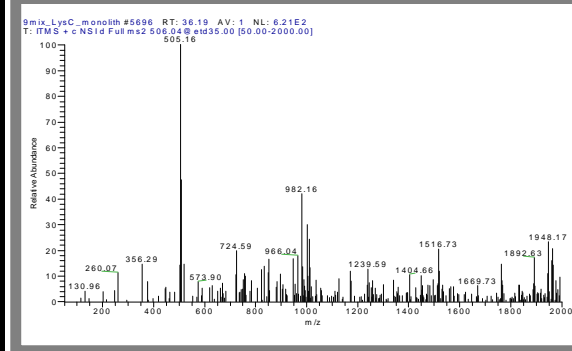
### Signal processing for charge state determination.

For spectra with a large number of complementary ions, charge state determination is more specific and accurate. This is done by finding the maximum of all pair-wise sums of fragment ions. We illustrate an example of this procedure with a spectrum of a chicken ovalbumin peptide (AFKDEDTQAMPFRVTEQESK) in Figure 3. For ETD spectra, we frequently observe that the most intense ions are those of the reduced precursor – products of electron transfer without fragmentation. Charge state of peptides can be determined using these ions, as well. This information is used in the LDA model.

Table 1. Effectiveness of charge state determination using *Charger*.

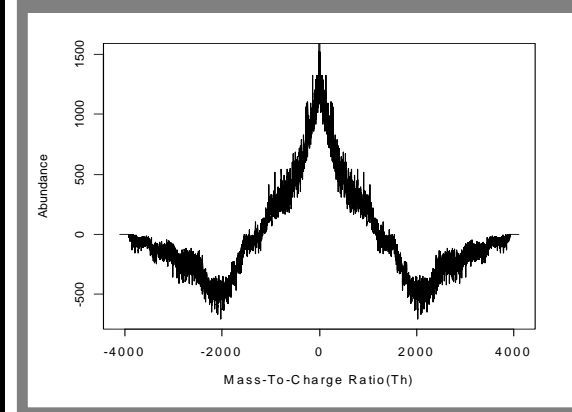
	9 Protein Mix LysC digest	9 Protein Mix	49 Protein Mix
Number of Spectra without <i>Charger</i>	10074	8712	16326
Number of Spectra with <i>Charger</i>	1807	1537	3074

FIGURE 3. ETD spectrum of a bovine albumin peptide with +4 charge, LKPPNPTLCDEFKADEK.



After processing and correlation, the spectrum is transformed as shown in Figure 4. The function in the figure represents the pair-wise sum of all fragment ions from Figure 2. Note that the maximum at the 0 Th corresponds to the self-sum of all peaks. The minimum of the function often indicates the position of the precursor mass. We have used and compared charge state determination via self-convolution (data not shown) with this method because in practice, the approach was proven to be more sensitive.

FIGURE 4. Autocorrelation of the spectrum in Figure 2. The peptide precursor mass matches the minimum of the autocorrelation function at 2021 amu.

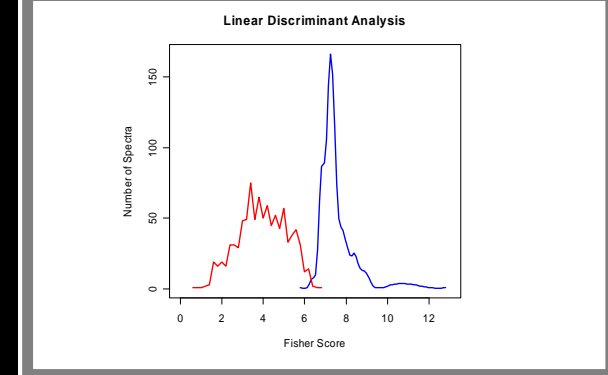


### Linear Discriminant Analysis

In our observations, charge state determination via autocorrelation is successful mostly for spectra with a large number of complementary ions with prominent abundances. The method is very selective and generates almost no false positives. For most peptides where the charge states could be determined via processing of tandem mass spectra, we use linear discriminant analysis. Figure 5 shows an example of the Fisher scores for true and false assigned +2 charged peptides from our data set.

It is apparent in this example that the Fisher score is an effective measure for charge state classifications of +2 charged peptides. The overlap region between +2 charged and other spectra correspond to mixtures of peptides with different charge states.

FIGURE 5. Fisher scores for true (blue) and false (red) assigned +2 charges. The area in the overlap corresponds to spectra which are a mixture of +2 and other charge states.



We have applied this approach for confident precursor charge state determination to several spectral data sets generated from protein samples with known content, with results summarized in Table 1. The first row indicates the number of spectra that are necessary to process if no charge state determination is done, and every charge state from +2 to +7 is assumed to be likely. The second row indicates the number of spectra after processing with *Charger*. It is evident from the data in Table 1 that ETD data preprocessing significantly reduces the effective number of spectra to be processed in a database search.

## Conclusions

- We have developed and implemented a combination of signal processing and linear discriminant analysis to determine charge states of peptides from their ETD tandem mass spectra.
- Signal processing self-correlates the tandem mass spectra to determine the mass-to-charge ratio characteristic of most of the pair-wise complementary ion sums.
- Statistical analysis uses information from a spectral data set of peptides with known charge states to determine features that separate the true and false charge state assignments as much as possible.
- A Fisher score which classifies the charge states is a linear function of the features.
- Application of this novel ETD data preprocessing routine not only reduces the number of spectra for database processing, but also reduces the number of potentially false positive identifications.

## References

- Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, and Hunt DF. . *Protein sequence analysis using electron transfer dissociation mass spectrometry*, P.N.A.S. 2004, 101, 9528.
- R G. Sadygov, J. Eng, E. Durr, A. Saraf, H. McDonald, M. J. MacCoss, and J. R. Yates, III, Code Developments to Improve the Efficiency of Automated MS/MS Spectra Interpretation, *J. Proteome Research*, 2002, v. 1., pp 211 - 215