

# Evaluation of a New MS/MS Search Algorithm for Peptide Identification from ETD Spectra

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

Thermo Fisher Scientific, San Jose, CA, USA

**Thermo**  
SCIENTIFIC

## Overview

### Purpose:

To evaluate the performance of a new MS/MS search algorithm for peptide identification from ETD spectra.

### Methods:

ETD spectra were generated using a Thermo Scientific LTQ XL™ ETD system. SEQUEST® and the new algorithm were applied and results were displayed using a beta version of a new BioWorks™ user interface. The .XML format files were also used to search ETD spectra with MASCOT™ version 2.2 using the ETD-TRAP instrument option.

### Results:

The data pre-processing function of the new algorithm largely reduces the number of spectra that need to be searched.

The new algorithm demonstrates higher sensitivity and specificity than either SEQUEST or MASCOT for peptide identification with ETD spectra.

The new algorithm generates higher deltaCn scores. This larger score difference between the first and the second best hits indicates that this algorithm is more discriminatory than either SEQUEST or MASCOT.

## Introduction

Database search algorithms which are widely used with CID spectra, such as MASCOT and SEQUEST, have been applied by researchers to MS/MS spectra generated using electron transfer dissociation (ETD). However, ETD and CID generate spectra of completely different characteristics. The currently available search algorithms which provide a search option for c and z ion series generated by electron capture dissociation are usually not optimized for ETD spectra. Furthermore, the fact that ETD generates high quality spectra for peptides of higher charge states forces researchers to carry out multiple searches for a single ETD spectrum generated by a low resolution MS instrument. Each low resolution ETD spectrum needs to be searched several times in order to cover several potential precursor charge states (usually from +2 to +7).

A new database search algorithm has been recently developed which specifically takes into account the unique characteristics of ETD spectra. It includes a data pre-processing step that assigns a charge state to precursor ions according to the characteristics of ETD spectra(1). Thus, a pre-processed ETD spectrum no longer needs to be searched multiple times. The use of this data pre-processing step reduces not only the number of data files searched but also the number of false positive identifications generated from multiple searches of each ETD spectrum. In this study, the combination of a rigorous spectrum data pre-processing and novel ETD-specific scoring algorithm is evaluated in comparison with MASCOT and SEQUEST for identification of peptides and proteins.

## Methods

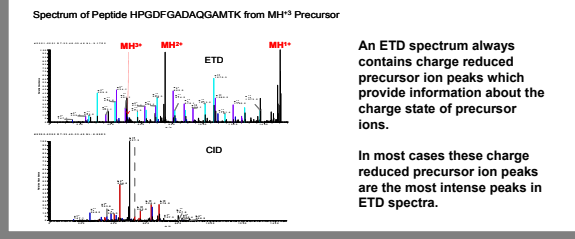
### Sample preparation, data collection and instrument method

Protein samples (standard 9 protein mixture, ABRF sPRG 49 Protein Standard, *E. coli* cell lysate and human CSF) were denatured, reduced and alkylated before being digested with enzymes. The digested samples were fractionated by reverse phase chromatography using a Thermo Scientific Surveyor™ HPLC equipped with a Micro AS and nanospray source (Thermo Fisher Scientific, San Jose). The eluted peptides were analyzed by a Thermo Scientific LTQ XL with ETD (Thermo Fisher Scientific, San Jose) and raw data files were acquired using a Data Dependent™ alternating ETD/CID MS/MS instrument method (1 full MS plus 3 ETD and 3 CID MS/MS on the 3 most intense peaks with dynamic exclusion) using Xcalibur™ 2.0 software.

### Data analysis

DATA files were generated using a beta version of BioWorks including the ETD data pre-processing step. A reversed database was generated based on the forward database. Searches were carried out against the uniprot\_sprot database in both forward and reverse directions using a standard, unmodified version of SEQUEST, MASCOT 2.2, and the new algorithm. The following parameters were used: carboxyamidomethylated cysteine as static modification; fully enzymatic with four missed cleavage sites; 4 AMU for peptide tolerance and 1 AMU for fragment ion tolerance. The filtering criteria used were: probability scores for MASCOT and the new algorithm, Xcorr (+2, 1.7; +3, 3.1; +4, 3.7; +5, 4.3; +6, 4.9) and deltaCn for SEQUEST. The false positive rate (FPR) is defined as the proportion of false positives among all positive identifications which pass a certain criteria and calculated as previously reported:  $FPR = FP / (TP + FP)$ , which can thus be estimated as  $N_{reversed} / (N_{forward} + N_{reversed})$  are the number of peptide identifications derived from the forward and reversed database searches that pass a given set of filtering criteria (2). For each MS/MS scan, only the top scoring hit which passed a certain criteria were

FIGURE 1. ETD and CID spectra are of completely different characteristics



counted as a positive identification. The deltaCn value of a hit for a certain spectrum was calculated as (best hit score - second best hit score) / best hit score.

In addition, ROC curves were generated using the results from the forward database search. Criteria for generating combined (for all peptide charge states) ROC curves were: XCorr for SEQUEST and probability scores for Mascot and the new algorithm. Results were normalized with respect to the particular number of correct identifications that every search engine determined.

FIGURE 2. Precursor Charge State Assignment by Data Pre-processing - Distribution of Precursor Charge State after Processing

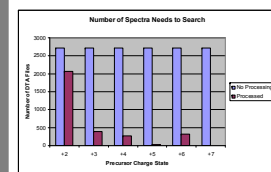


Table 1. Precursor Charge State Assignment by Data Pre-processing - Reduction of the Number of Spectra to be Searched

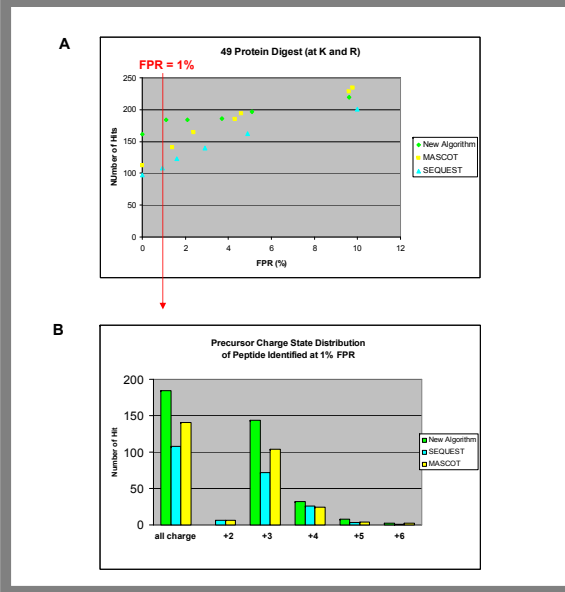
Raw File Name	No Processing	Processed
9 protein lysc digest	10074	1807
9 protein digest (at K and R)	8712	1537
ABRF sPRG 49 protein digest (at K and R)	16326	3074

## Results

### Evaluation of the ETD spectra pre-processing function

Since high-quality ETD spectra are usually observed for higher peptide precursor charge states (+3 and up), analysis of ETD spectra generated by low resolution MS instruments is challenging in the absence of correct charge state information for precursor ions. Researchers historically have had to search an ETD spectrum multiple times in order to cover all the possible charge states. Prior to a database search, the ETD pre-processing function reads, extracts and examines the characteristic of a spectrum and evaluates the precursor ion charge state. One of the most important features of ETD spectra is the series of peaks of charge-reduced precursor ions (Figure 1). Another feature is the loss of NH2 from the charge-reduced precursors. These characteristics and other spectral features provide information on the precursor ion charge state. In some cases, when confident assignment of a single precursor charge state is not possible, the two most likely charge states are determined; e.g., +3 and +6 or +4 and +6 etc. The charge state assignment function was evaluated using several of our standard data sets. Figure 2 shows a data set containing 2721 DTA files of ETD spectra. The red bars show the distribution of the precursor charge state for all ETD spectra after data pre-processing. The blue bars show that each of these 2721 files would have to be searched for each charge state if they were not pre-processed. Table 1 shows that pre-processing of ETD spectra from different raw files reduced the number of files for database searching more than 5-fold. Without this processing function, researchers would either have to rely on the automatic generation of +2 and +3 precursor ion selection and thus miss all the precursors ions of charge state above +3, or have to search 5 times more spectra to capture all the relevant information in the dataset.

FIGURE 3. Performance of the New Algorithm, SEQUEST and MASCOT on Peptide Identification from ABRF sPRG 49 Protein Digest - Sensitivity

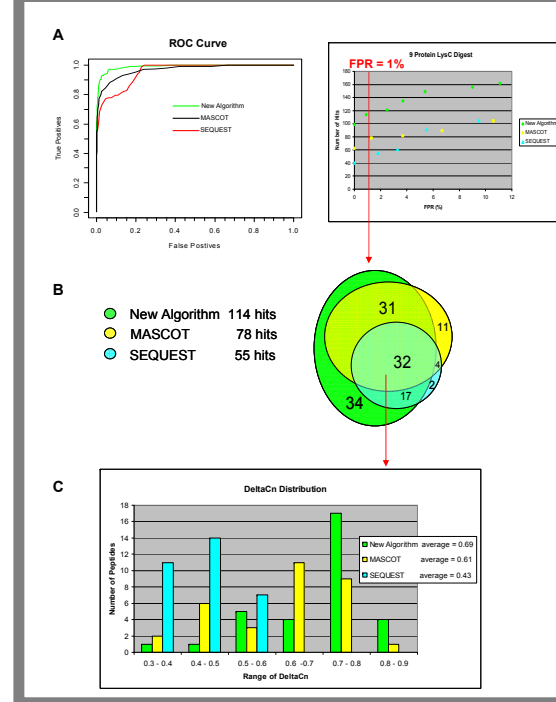


Further evaluation using our standard data sets indicates that the data pre-processing function of the new algorithm created a single correct charge state 90% of the time. For the remaining 10% of the spectra, two most likely charge states were assigned and one of the two were correct assignments (data not shown).

### Comparison of the new algorithm with SEQUEST and MASCOT for peptide identification with ETD spectra: sensitivity and specificity

To evaluate the performance of the new ETD scoring algorithm for peptide identification, SEQUEST and MASCOT were used in comparison to search raw files containing ETD spectra against the uniprot\_sprot database. Searches against the reversed database were used to estimate the number of false positive identifications and this number was used to calculate false positive rate as described in the methods section. The comparison of the search results are presented in Figure 3 and Figure 4. As shown in Figure 3, from the ABRF sPRG 49 protein digest (at Lys and Arg), the new algorithm identified more peptides than SEQUEST or MASCOT at low FPR (<4%). At 1% FPR it determined 42% more positive peptide identifications than SEQUEST and 23% more than MASCOT. The affect on sensitivity is less significant at a higher FPR (Figure 3A). The distribution of the precursor charge state for all the identifications at 1% FPR indicates that the new algorithm was more sensitive than the other two for charge state +3, +4, +5 and +6, but less sensitive for ETD spectra of +2 precursor ions (Figure 3B). For the analysis of larger peptides from a Lys-C 9 protein digest, the new algorithm identified not only more peptides at a low FPR level, but consistently showed higher sensitivity than the other two search algorithms (Figure 4A, right). In both cases, MASCOT demonstrated better sensitivity than SEQUEST at a low FPR (Figure 3A, 4A right). An ROC curve generated from the Lys-C 9 protein digest indicated that the new algorithm has better sensitivity as well as specificity than MASCOT or SEQUEST (Figure 4A, left). Figure 4B is a diagram presenting the overlap among all the three algorithms of identifications at 1% FPR from the 9 protein Lys-C digest. Of all the 131 peptides identified, the new algorithm covered 114 of them while SEQUEST and MASCOT identified 55 and 78 peptides, respectively. The increase in identification was 46% more than MASCOT and more than doubled compared with SEQUEST. 32 peptides were identified by all three algorithms and 84 were identified by at least two. 34 peptides were identified uniquely by the new algorithm, 11 by MASCOT and 2 by SEQUEST.

FIGURE 4. Performance of the New Algorithm, SEQUEST and MASCOT on Peptide Identification from 9 Protein LysC Digest - Sensitivity and Specificity

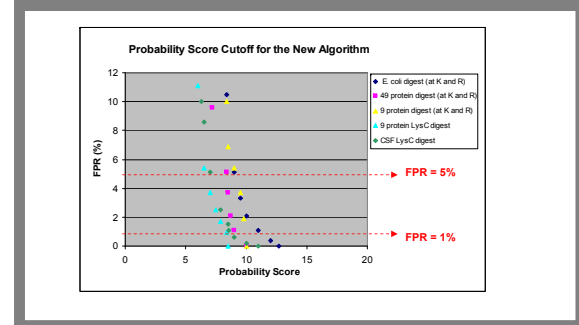


The discriminatory power of an algorithm can be measured by the difference of scores between the best and the second best hit for a MS/MS spectrum. A larger score difference between the first and the second hit suggests a better chance that the first hit is correct. To evaluate the discriminatory power of these three algorithms, the deltaCn of the scores were calculated for the 32 peptides identified by all the three algorithms from the 9 protein Lys-C digest. The distribution of the number of peptides over the deltaCn range is shown in Figure 4C. SEQUEST scores have a deltaCn range of 0.3 to 0.6 while the new algorithm and MASCOT scores have a higher deltaCn range of 0.3 to 0.9, with most of the deltaCn scores in the 0.5 to 0.9 range. Compared to MASCOT, the new algorithm had more peptides with a deltaCn in the range of 0.5 - 0.6 and 0.7 - 0.9, while MASCOT had more peptides with a deltaCn from 0.3 - 0.5 and 0.6 - 0.7. The average deltaCn scores for the new algorithm, SEQUEST and MASCOT are 0.69, 0.43 and 0.61 respectively. This result indicates that this newly developed ETD scoring algorithm has a higher discriminatory power on average than either MASCOT or SEQUEST.

### Score range for confident identification using the new algorithm

To investigate the probability score range and the related confidence level of identification, ETD raw files were collected using samples of different complexity and digested using different enzymes. The raw files were subsequently processed using the new algorithm. The results from reverse database searches were used to calculate false positive rates as described in the Methods section. Figure 5 shows the relationship between the false positive rate and the probability score cutoff value for each data file. For 1.0% FPR the probability scores span from 8.6 to 11.0, and for 5% FPR the range of probability scores is from 6.5 to 9.0 for different samples sets used in this study. This cutoff value range can be used to filter out false positive identifications. It can also be observed that the two Lys-C (cleaving at K) digested samples have relatively lower probability scores for the same level of FPR when

FIGURE 5. Probability Scores for the New Search Algorithm as a Filtering Criteria for Confident Peptide Identification with ETD Spectra



compared to the other samples which were digested using an enzyme cleaving at both K and R. More research is needed to evaluate whether this represents a general trend.

## Conclusions

A new ETD MS/MS search algorithm was evaluated along with a new data pre-processing function for peptide identification from ETD spectra. Data files from samples of different levels of complexity, and digested using different enzymes were processed. The results in this study indicate that:

- For ETD spectra obtained from low resolution MS instruments, the data pre-processing function of the new algorithm assigns charge state to precursor ions according to unique ETD spectral characteristics. 90% of the time, a single correct charge state was created, while 10% of the time when confident assignment of precursor charge state was not possible, the two most likely charge states were assigned.
- The charge state assignment of ETD spectra reduced the number of spectra for a database search more than 5-fold compared to ETD data analysis without pre-processing.
- The new algorithm had significantly better overall sensitivity and specificity than MASCOT and SEQUEST for ETD spectra for all the data files used for this evaluation. For proteins cleaved at both K and R, it showed higher overall sensitivity at low FPR range. For proteins digested by Lys-C, it consistently showed higher overall sensitivity. For precursor charge states from +2 to +6, the new algorithm was more sensitive for all the charge states except for +2.
- Average deltaCn scores generated by the new algorithm were larger than for MASCOT or SEQUEST, indicating that the new algorithm has better discriminatory power.
- The probability scores of the new algorithm can be used as a cutoff filter to remove false positive identifications. In this study, the FPR calculated from a reverse database search suggested a probability score range of 6.5 to 9.0 for 5% FPR and 8.5 to 11 for 1% FPR.

## References

- Sadygov R, et al. Data Processing and Database Search Models for Tandem Mass Spectra Obtained via Electron Transfer Dissociation. poster MPK194, ASMS 2007
- Qian W.J., et al. Probability-based Evaluation of Peptide and Protein Identifications from Tandem Mass Spectrometry and SEQUEST Analysis: The Human Proteome. Journal of Proteome Research, 2005, 4, 53-62

SEQUEST is a registered trademark of the University of Washington. MASCOT is a trademark of Matrix Science Ltd. All other trademarks are the property of Thermo Fisher Scientific Inc. and its subsidiaries.