

ETD Spectra Processing and Database Searching Using ZCore - Performance Evaluation of a New MS/MS Search Algorithm

Zhiqi Hao, Rovshan Sadygov, Shijun Li and Andreas FR Hühmer
Thermo Fisher Scientific, San Jose, CA, USA

Overview

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

Methods: SEQUEST®, MASCOT™ and ZCore were used for data analysis, and results were displayed using a new biosoftware platform, Proteome Discoverer.

Results: The data pre-processing function of ZCore significantly reduces the number of spectra that need to be searched. ZCore demonstrates higher sensitivity and specificity than either SEQUEST or MASCOT for peptide identification of ETD spectra. The new algorithm generates higher deltaCn scores. This larger score difference between the first and the second best hits indicates that ZCore is more specific than either SEQUEST or MASCOT. ETD spectra database search using ZCore is less time consuming compared to SEQUEST.

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 (ECD) 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 with a unit resolution MS instrument. Each unit resolution ETD spectrum needs to be searched several times in order to account for several potential precursor charge states (usually from +2 to +7).

A new database search algorithm, ZCore, 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 (1,3). 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 a novel ETD-specific scoring algorithm is evaluated in comparison with MASCOT and SEQUEST for identification of peptides and proteins.

Results

Evaluation of the ETD spectra pre-processing function

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 account for 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). This characteristic, along with 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. (3).

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 no pre-processing function available. 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 2. Precursor Charge State Assignment by Data Pre-processing - Distribution of Precursor Charge State after Processing

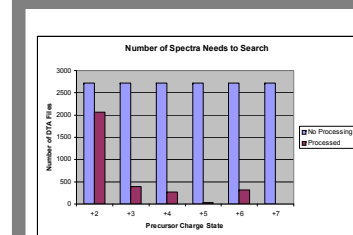


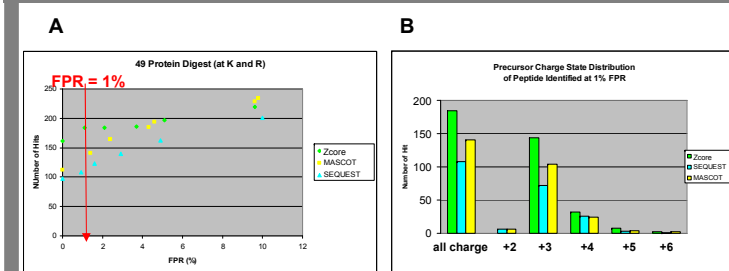
Table 1. Precursor Charge State Assignment by Data Pre-processing - Reduction of the Number of Spectra to be Searched

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

Comparison ZCore with SEQUEST and MASCOT for peptide identification with ETD spectra: sensitivity and specificity

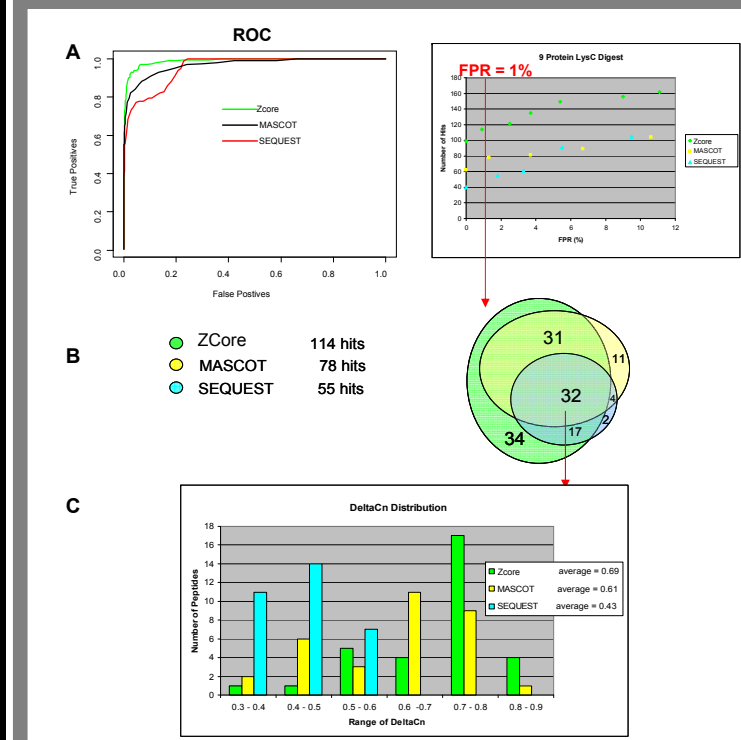
To evaluate the performance of ZCore for peptide identification, raw files containing ETD spectra were searched against the uniprot_sprot database with SEQUEST, MASCOT and ZCore. Searches against the reversed database were used to estimate the number of false positive identifications and this number 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.

FIGURE 3. Performance of ZCore, SEQUEST and MASCOT on Peptide Identification from ABRF sPGR 49 Protein Digest - Sensitivity



As shown in Figure 3, from the ABRF sPRG 49 protein digest (at Lys and Arg), ZCore 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 effect 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 ZCore 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, ZCore 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 ZCore has superior 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, ZCore 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.

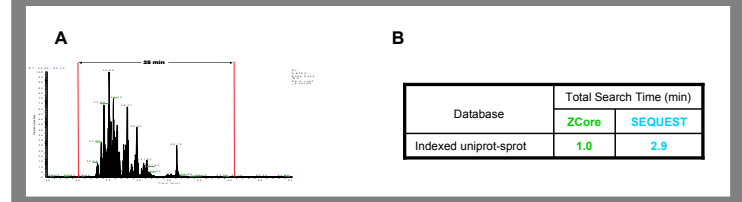
FIGURE 4. Performance of the Zcore, SEQUEST and MASCOT on Peptide Identification from 9 Protein LysC Digest - Sensitivity and Specificity



The discriminatory power of an algorithm can be measured by the difference of scores between the best and the second best match for a MS/MS spectrum. A larger score difference between the first and the second match indicates an increased chance that the first match 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 ZCore 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, ZCore had more peptides with a deltaCn in the range of 0.5

FIGURE 5. Comparison of Search Time for Peptide Identification from 9 Protein Lys C Digest.

- Chromatography of an alternating CID-ETD raw file of 9 protein Lys C digest. Searches were performed for ETD scans from scan 4723 (29.93min) to scan 8639 (55.32min).
- Elapsed search time for ZCore or SEQUEST.



– 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 ZCore, 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.

Evaluation of ETD spectra database search speed of ZCore

An LC MS/MS alternating CID/ETD raw file of 9 protein digest was used to evaluate the search speed of ZCore in comparison with SEQUEST. The searches were performed against an indexed uniprot-sprot database using a personal computer. The search parameters used for the two algorithms were identical. As shown in figure 5, for total number of scans acquired in 25 min (1807 scans), the total search time of ZCore was 1 min, while that of SEQUEST was 2.9 min. Thus ZCore runs significantly faster compared to SEQUEST.

Conclusions

A new ETD MS/MS search algorithm, ZCore, 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:

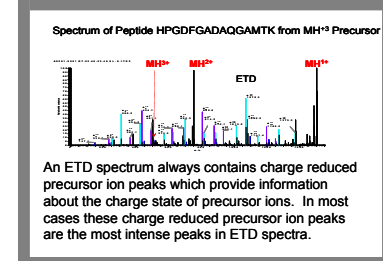
- The charge state assignment of ETD spectra reduced the number of spectra for a database search more than five-fold compared to ETD data analysis without pre-processing.
- The ZCore had significantly better overall sensitivity and specificity than MASCOT and SEQUEST for ETD spectra for all the data 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 ZCore were larger than for MASCOT or SEQUEST, indicating that the new algorithm has better discriminatory power.
- Under the search criteria used for this study, search time of ZCore for ETD spectra is about one third of that of SEQUEST.

References

- Sadygov R.G. 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.
- Sadygov R.G. et al, Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. Anal Chem. 2008 Jan 15;80(2):376-86.

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.

FIGURE 1. ETD Spectrum Characteristics



Methods

Sample preparation, data collection and instrument method

Protein samples were denatured, reduced and alkylated before being enzymatically digested and then fractionated using reverse phase chromatography. The eluted peptides were analyzed using a Thermo Scientific LTQ XL™ with ETD (Thermo Fisher Scientific, San Jose) using a Data Dependent™ alternating ETD/CID MS/MS instrument method (1 full MS followed by 3 ETD and 3 CID MS/MS on the 3 most intense peaks with dynamic exclusion).

Data analysis

ETD data pre-processing step and 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 ZCore in Proteome Discoverer. 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 ZCore, 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: $N_{reverse} / (N_{forward} + N_{reverse})$. $N_{forward}$ and $N_{reverse}$ 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 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 ZCore. Results were normalized with respect to the particular number of correct identifications that every search engine determined.