ETD Spectra Processing and Database Searching – Performance Evaluation of a new MS/MS Search Algorithm

Zhiqin Hao; Kovshon Sadayov; Roger G Biringer and Andreas FR Hübler

Thermo Fisher Scientific, San Jose, CA, USA

Overview

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

Methods: SEQUEST and a new algorithm were used for data analysis, and results were displayed using a web-based version of the BioTools™ user interface. The XLM 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 significantly reduces the number of spectra that need to be searched. The new algorithm demonstrates higher sensitivity and specificity than either SEQUEST or MASCOT database identification with the performance scores of the new algorithm and SEQUEST being more than 1% higher on average. The results of the new algorithm were compared to those of SEQUEST and MASCOT, and the new algorithm identified more unique peptide identifications as well as more peptides at a lower FPR level, but consistently showed higher sensitivity than the other two search algorithms.

Introduction

Database search algorithms which are widely used with CID spectra, such as MASCOT and SEQUEST, have been applied by researchers to MS/MS data generated by ETD instruments. However, ETD spectra are generated by 3- to 5-time-lower resolution mass spectrometers and contain many fragment ions with the same charge state as intact protein ions. Thus, ETD spectra need to be searched several times in order to cover several potential precursor charge states (7, 12). A new database search algorithm has been recently developed which specifically takes into account the unique mass 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 (8). This pre-processing step reduces the number of entries that need to be searched. The use of this data pre-processing step reduces not only the number of database searches, but also the number of peptide identifications generated from multiple searches of each ETD spectrum. In this study, the combination of a rigorous spectrum data pre-processing and novel ETD database searching algorithms evaluated in comparison with SEQUEST and MASCOT for identification of peptides and proteins.

Data analysis

DTD files were generated using a beta version of BioWorks including the ETD data pre-processing function. Searches were carried out against the uniprot database and a previously developed data pre-processing step that assigns a charge state to precursor ions according to the characteristics of ETD spectra (13). This pre-processing step reduces the number of database searches by significantly reducing the number of false positive identifications. However, this pre-processing step also reduces the number of peptide identifications generated from multiple searches of each ETD spectrum. Figure 1 shows a data set containing 2721 DTA files of ETD spectra. The red bars show the number of DTA files assigned to each charge state for precursor ions. Researchers historically have had to search an ETD spectrum multiple times in order to cover all possible charge states. Prior to this study, the ETD pre-processing function reads, extracts and examines the characteristics of a spectrum and evaluates the charge of the precursor ion (1). This characteristic, and other spectral features provide information on the precursor ion charge state. In some cases, when assignment of a single precursor ion charge state is not possible, the two most likely charge states are determined; e.g., +3 and +4 or +5 and +6. 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 each ETD spectrum before data pre-processing. The results are normalized with respect to the number of DTA files. Figure 3 shows a spectrum of Peptide HPGDFGADAQGAMTK from MH+3 Precursor. As shown in Figure 3, the average deltaCn scores for the new algorithm, MASCOT and SEQUEST were 0.8, 0.6 and 0.2, respectively. Thus, the new algorithm is significantly better than other algorithms in terms of overall performance.

Results

Evaluation of the ETD spectra pre-processing function

Analysis of ETD spectra generated by three MS/MS instruments is challenging in the absence of accurate charge state information for precursor ions. Researchers historically have had to search an ETD spectrum multiple times in order to cover all possible charge states. Prior to this study, the ETD pre-processing function reads, extracts and examines the characteristics of a spectrum and evaluates the charge of the precursor ion (1). This characteristic, and other spectral features provide information on the precursor ion charge state. In some cases, when assignment of a single precursor ion charge state is not possible, the two most likely charge states are determined; e.g., +3 and +4 or +5 and +6. 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 each ETD spectrum before data pre-processing. The results are normalized with respect to the number of DTA files. Figure 3 shows a spectrum of Peptide HPGDFGADAQGAMTK from MH+3 Precursor. As shown in Figure 3, the average deltaCn scores for the new algorithm, MASCOT and SEQUEST were 0.8, 0.6 and 0.2, respectively. Thus, the new algorithm is significantly better than other algorithms in terms of overall performance.

Conclusions

A new ETD MSMS search algorithm was evaluated along with a new data pre-processing function for spectra identification from ETD spectra. Data files from samples of different levels of complexity, and digested using different enzymes were processed. The results are as follows:

1. 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.
2. The new algorithm had significantly better overall sensitivity and specificity than either MASCOT or SEQUEST, and the number of peptide identifications was lower at a given FPR level.
3. Average deltaCn scores generated by the new algorithm were lower than for MASCOT and SEQUEST, indicating better overall discriminatory power.
4. Under the same search parameters used in the study, the sensitivity of the new algorithm for ETD spectra is about 50% that for SEQUEST.

References: