

Workflow for maximizing proteome coverage using CID and ETD

Martin Zeller; Bernard Delanghe; Torsten Ueckert; Thomas Moehring

Thermo Fisher Scientific, Bremen, Germany



Overview

Purpose: To develop strategies for both data acquisition and data mining for maximum proteome coverage

Methods: Using CID and ETD as acquisition strategies in combination with dynamic exclusion list generation for multiple sample injection, and multiple search engines (Sequest, Mascot, and ZCore) for data analysis

Results: Use of complementary dissociation techniques, exclusions lists and use of multiple search engines increase proteome coverage

Introduction

In a typical large-scale shotgun protein sequencing experiment thousands of peptides are generated using enzymatic degradation. The generated peptides comprise an enormous complexity not only in abundance but also in chain length and amino acid composition resulting in different physico-chemical properties.

In this work we would like to present analytical strategies for maximizing the sequencing outcome using the LTQ Orbitrap XL ETD.

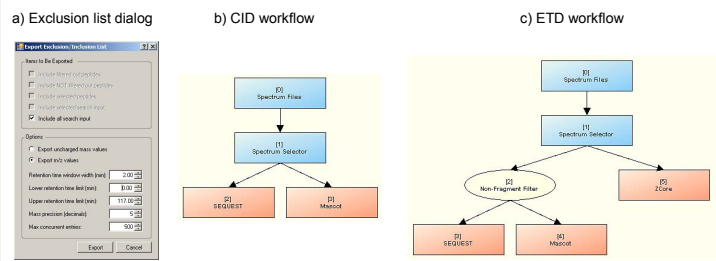
The strategy is based on :

- acquisition strategies
 - use of CID and ETD as complementary fragmentation techniques
 - multiple injections using dynamic exclusions lists of previously triggered precursor ions
- data mining strategies using multiple search engines

Methods

All spectra were acquired on a LTQ Orbitrap XL ETD™ (Thermo Fisher Scientific, Bremen, Germany). The complex *C.elegans* peptide mixtures (2 µl injected) were separated via Surveyor™ LC equipped with MicroAS™ autosampler (Thermo Fisher Scientific) using a reversed phase peptide trap (100 µm inner diameter, 2 cm length) and a reversed phase analytical column (75 µm inner diameter, 10 cm length, 3 µm particle size, both Nanoseparations, NL), at a flow rate of 250 nL/min. A gradient of 5 - 30% acetonitrile in 90 minutes was used. Data was analyzed using Proteome Discoverer Software Suite. After the first run the search input was exported as *m/z* values together with the retention times and a retention time window of 2 minutes and used as exclusion list in the second run (see figure 1a). For the third replicate the exclusion list of the first and the second run was merged. The data was analyzed using the search algorithm Mascot™ and Sequest™ for CID and Mascot, Sequest and ZCore™ for ETD. Before submitting the peaklists of the ETD spectra to Mascot and Sequest, the peak list entries for the precursor, charge reduced species and neutral losses thereof were removed using a 2 Da window (see Figure 1b) and 1c). A decoy database search was performed using 0.2% and 1.0% false discovery rate (FDR).

FIGURE 1. a) Generation of exclusions lists using the search input
b) Workflow in Proteome Discoverer for CID searches
c) Workflow in Proteome Discoverer for ETD searches



Results

In total 932 proteins were identified with a 0.2% FDR and 1168 proteins with a 1.0% FDR. CID identified 1076 proteins and ETD 904 proteins at the 1% FDR. More proteins were identified uniquely by CID compared with ETD. One explanation for this observation is that the standard bottom-up approach used generates relatively small peptides that ionize predominately in a doubly charged form which is unfavorable to ETD. The ratio of the percentage of identified peptides per charge state versus the charge states of precursors selected for MS/MS (search input) shows that CID has strengths at low charge states. 2+ precursor ions are more likely to be identified by CID using Sequest or Mascot. ETD of doubly charged precursor ions does not cause efficient fragmentation, mainly due to the mechanism of ETD where the first step is a charge reduction. ETD appears to be more efficient at fragmentation of higher charge states and therefore more efficient with enzymatic degradations using enzymes that produce peptides with longer amino acid sequences and a higher probability of higher charge states such as LysC.

FIGURE 2. Proteins identified in total with 0.2% FDR and 1.0% FDR

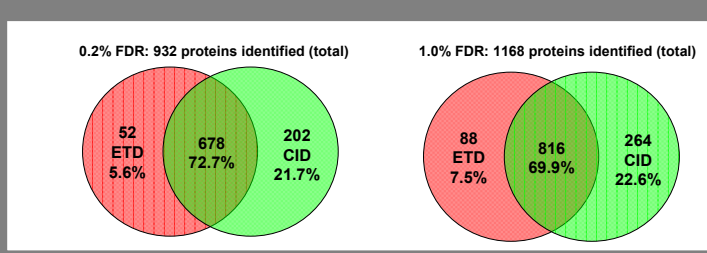


Table 1: Percentage of CID and ETD spectra submitted and identified at 1.0% FDR.

CID:		ETD:	
Search input	Identified peptides	Search input	Identified peptides
Charge 2+: 59.7%	2+: 84.8%	Charge 2+: 55.7%	2+: 32.5%
3+: 30.0%	3+: 14.7%	3+: 32.1%	3+: 54.3%
4+: 8.2%	4+: 0.5%	4+: 8.9%	4+: 10.8%
5+: 1.7%	5+: 0.0%	5+: 2.3%	5+: 1.9%
6+: 0.3%	6+: 0.0%	6+: 0.3%	6+: 0.4%
7+: 0.1%	7+: 0.0%	7+: 0.1%	7+: 0.0%

The second acquisition strategy is based on multiple runs (triplicates in this study) with dynamic exclusion list generation after each run.

FIGURE 3. Number of unique proteins (a) and peptides (b) identified at 1.0% FDR per run

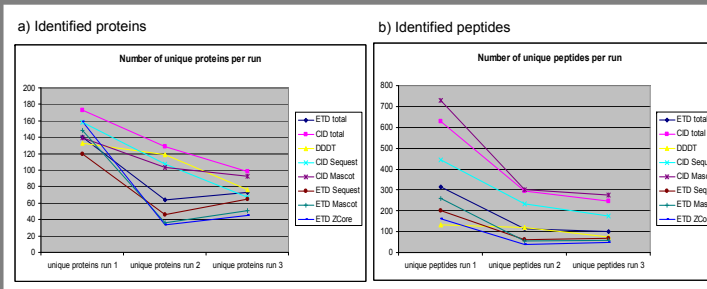


Figure 3 shows the number of identified unique proteins (3a) and unique peptides (3b) per run at 1.0% FDR. The figure also includes the data for the same experiment using the data-dependent decision tree (DDDT) acquisition strategy.

For the number of identified unique proteins, CID and ETD show a different trend: The number of proteins identified with CID decreases in the second and third run almost linearly. Almost the same trend can be seen for the number of unique peptides for CID and predominately proteins with 1 peptide per protein are identified (data not shown), increasing the sensitivity for protein identification. For ETD, the number of identified unique proteins and unique peptides decrease in the second run but increase in the third run. The explanation for this observation is that more likely the higher charge states of the peptides are fragmented that have been missed in the previous runs because of the lower intensity. Maximum protein coverage is not only accomplished by acquisition strategies but also by different post-processing methods.

FIGURE 4. Proteins identified at 1% FDR, total number of proteins identified: 1168

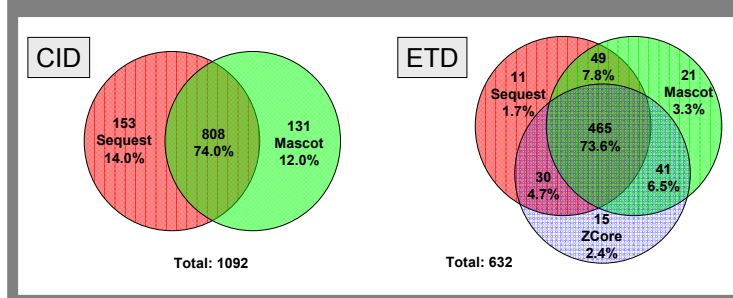


FIGURE 5. Peptides identified at 1% FDR, total number of peptides identified: 3574

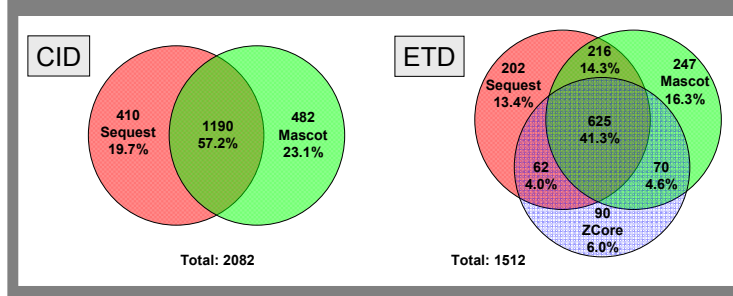
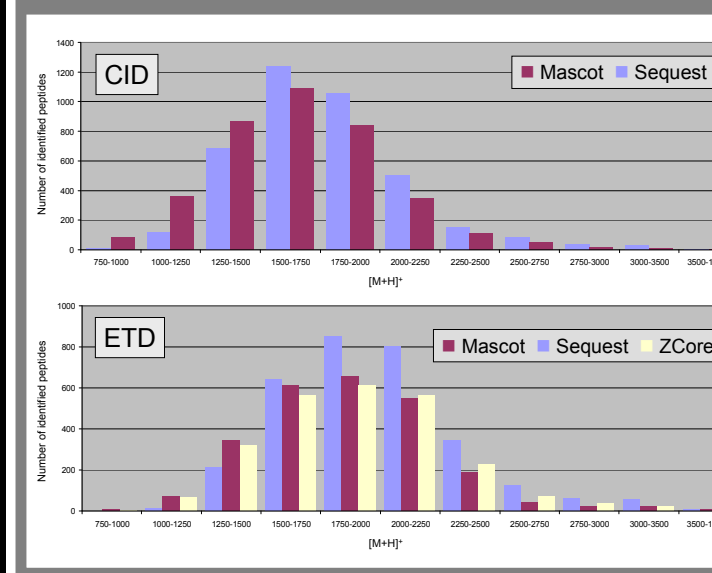


Figure 4 and 5 show the distribution of the unique proteins and peptides identified with Sequest and Mascot for CID and Sequest, Mascot and ZCore for ETD. It can clearly be seen that there is a fair overlap of proteins and peptides identified by all search engines. However, each search engine identifies unique peptides that are missed if only one search engine is used. For ETD it appears that Mascot and Sequest identify peptides and that there is a high overlap of peptides. ZCore does not show a high overlap with Mascot and Sequest and is therefore a complementary search engine.

FIGURE 6. Distribution of identified peptides per search engine



When looking at the number of identified peptides depending on the $[M+H]^+$ in figure 6 it becomes obvious that Mascot identifies more peptides with low molecular weight. At higher MW Sequest has its strengths and identifies more peptides compared to Mascot. The data also suggests that ZCore follows the trend of Mascot but also identifies peptides at higher MW.

This clearly shows the complementary character of the different search engines and the benefit of using them together for in-depths data mining strategies.

Conclusions

- ETD and CID are complementary fragmentation techniques
- Sequest, Mascot and ZCore are complementary search engines
- Enzymes generating longer peptides are more suited for ETD
- Combination of multiple search engines increase proteome coverage

References

McAlister et al., A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer, *J. Proteome Res.*, 7, 3127 (2008)
Swaney et al., Decision tree-driven tandem mass spectrometry for shotgun proteomics, *Nat. Methods*, 5, 959 (2008)

Acknowledgements

We thank Mike MacCoss (University of Washington, USA) for the *C.elegans* digest.

Mascot is a trademark of Matrix Science. All other trademarks are the property of Thermo Fisher Scientific and its subsidiaries.