Published in final edited form as:

Proteomics. 2010 January ; 10(1): 164–167. doi:10.1002/pmic.200900570.

The Effect of Interfering Ions on Search Algorithm Performance for ETD Data

David M. Good¹, Craig D. Wenger¹, and Joshua J. Coon^{1,2,*}

¹Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706

²Department of Biomolecular Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706

Abstract

Collision-activated dissociation (CAD) and electron-transfer dissociation (ETD) each produce spectra containing unique features. Though several database search algorithms (*e.g.*, SEQUEST, Mascot, and OMSSA) have been modified to search ETD data, this consists chiefly of the ability to search for *c*- and *z*[•]-ions; additional ETD-specific features are often unaccounted for, and may hinder identification. Removal of these features *via* spectral processing increased total search sensitivity by ~20% for both human and yeast datasets; unique identifications increased by ~17% for the yeast datasets and ~16% for the human dataset.

Recently [1], we reported on the use of post-acquisition spectral processing of ETD spectra for increasing peptide identifications at a controlled false discovery rate (FDR) using the Open Mass Spectrometry Search Algorithm (OMSSA) [2]. Here, we extend this work by examining the effects of performing this processing on the number of identifications, at a set confidence threshold, using two other commonly employed database search algorithms that were originally constructed to search CAD data (SEQUEST [3] and Mascot [4]), as well as one search algorithm (ZCore [5]) created explicitly for the analysis of tandem mass spectra resulting from electron-based fragmentation. We compare the effect of removing different interfering ions on the results of each search algorithm and outline the most effective pre-processing method for each algorithm. In addition, we illustrate the effectiveness of these algorithms for analysis of ETD-generated data.

Experiments were performed using both yeast and human samples. Yeast cells were prepared and analyzed as described previously [6]. Additionally, one sample of human embryonic stem (hES) cells was prepared in the same manner. Detailed descriptions of all sample preparations and mass spectrometry analyses are included in the Supplementary Information. Pre-processing of all resulting spectra was performed as described earlier [1].

MS/MS spectra were searched against concatenated target-decoy databases [7] of yeast (downloaded 08/21/2007 from *Saccharomyces Genome Database* at <http://www.yeastgenome.org/>) or human (v3.53 from the International Protein Index at <http://www.ebi.ac.uk/IPI/>) proteins. Searches were performed using OMSSA [2] (Open Mass Spectrometry Search Algorithm; version 2.1.4; freely available from the NCBI), SEQUEST [3] (Thermo Fisher Scientific, San Jose, CA; version 28, revision 12), MASCOT [4] (Matrix Science, London, UK; version 2.2.04), and ZCore [5] (beta version acquired from Thermo Fisher Scientific, San Jose, CA).

*To whom correspondence should be addressed. Prof. Joshua J. Coon, Department of Chemistry, University of Wisconsin – Madison, 1101 University Avenue, Madison, WI 53706. jcoon@chem.wisc.edu. Tel: (608) 890-0763.

The parameters common to all search algorithms were: static modification of +57 Da on cysteine residues (carbamidomethylation), variable modification of +16 Da on methionine residues (oxidation), precursor mass tolerance of ± 4.0 Da, fragment mass tolerance of ± 0.5 Da, maximum of 3 missed cleavages, and Lysine-C as the protease. Search results were filtered to an approximately one percent FDR [7] using software written in-house for optimization of unique peptide identifications based on expectation value score and precursor mass error thresholds.

The overall results are summarized in Figure 1. Here, we display the effects of each pre-processing step on the number of both total and unique peptide identifications at an $\sim 1\%$ FDR. In general, both total and unique identifications increase as more interfering peaks are removed from the spectra. Pre-processing has little to no effect on the number of identifications generated when using ZCore, which was anticipated due to ZCore inherently carrying out many of these same pre-processing steps. The largest increases were for OMSSA, with 8271 more total identifications and 4170 more unique identifications in the yeast dataset and 2379 more total identifications and 1384 more unique identifications in the human dataset. SEQUEST and Mascot also showed considerable increases, with SEQUEST gaining 6279 total and 2731 unique identifications for the yeast datasets and 2261 total and 1265 unique identifications for the human data, and Mascot improving by 4450 total and 1688 unique identifications in yeast and 1154 total and 515 unique identifications in human.

Supplemental Figure 1 illustrates in more detail the effect of each processing step on the number of unique identifications at an $\sim 1\%$ FDR. For the yeast data, these numbers reflect the average percent change for all three analyses. In both the yeast and human datasets, the greatest increase in unique identifications came from removal of the precursor; an $\sim 10\%$ and $\sim 7.3\%$ average increase for the yeast and human data, respectively. Subsequent removal of charge-reduced precursor (CRP) peaks from the spectrum only slightly increased unique identifications ($\sim 13\%$ for yeast and $\sim 11.3\%$ for human). When fully processing the spectrum prior to database searching, gains of $\sim 17\%$ for yeast and $\sim 15.6\%$ for human were attained by removing m/z peaks resulting from neutral losses. As expected due to removing a larger m/z region (and thus more confounding peaks) from the spectrum, removal of these neutral loss-associated peaks had a greater impact on total identifications than removal of the CRPs.

One probable reason for the large increase in identifications from removing the precursor is the contribution of the precursor to the total ion current (TIC) of the spectrum. Supplemental Figure 2 presents the percentage of the TIC removed from the spectrum for each of the processing steps. The precursor was, on average, the most intense peak within a spectrum, and its removal accounted for a decrease in TIC of $\sim 41\%$ for yeast and $\sim 31\%$ for human. Subsequent removal of CRPs and NLs accounted for a decrease in TIC by $\sim 15\%$ and $\sim 11\%$, respectively. In total, approximately 67% of the TIC in the yeast spectra and $\sim 57\%$ of the TIC in human spectra was removed by the spectral processing. This difference is most likely due to differences in precursor charge—the average charge of an identified peptide was ~ 3.25 for yeast and ~ 3.66 for human. Because the reagent anion, fluoranthene, is singly charged, the dominating factor in precursor-to-product conversion efficiency in ETD is the charge of the peptide cation. This increase of ~ 0.4 charges led to an increased reaction rate between precursor and reagent anion, thus leaving a smaller total amount of unreacted precursor within the spectra from human peptides.

An additional benefit of performing these analyses was the opportunity to compare database search algorithms for ETD-generated data. In contrast to a typical database search algorithm comparison, these data were searched using multiple datasets which had been pre-processed. Because of this, we gained further insight into how each search algorithm deals with, and is affected by, confounding peaks. Figure 2 shows the overlaps in unique identifications for each

of the four algorithms used in this study for both the human and yeast data. The four-way Venn diagrams presented here are formatted similarly to a heat map, with the lighter colored areas containing greater numbers of identifications. As expected, the greatest overlap between any of the search algorithms is with the three other algorithms (6435 and 4357 unique identifications overlap in the yeast and human datasets, respectively). The number of unique identifications which were identified by only a single algorithm followed the overall trends in total and unique identifications, with SEQUEST providing the fewest unique identifications for both datasets (223 for yeast and 202 for human), ZCore the greatest for yeast (845), and OMSSA the most for human (548). Mascot afforded 444 and 310 for the yeast and human data, respectively.

Referring back to Figure 1, one can infer the importance of removing certain interfering peaks on the probability of a successful identification for each of the four algorithms tested here. Perhaps most apparent is the lack of dependence on spectral pre-processing when using the ZCore algorithm. This is due to ZCore inherently performing similar processing steps as those employed here, removing CRPs in addition to peaks resulting from ammonia neutral losses, which are typically one of the largest contributors to the TIC. Further, ZCore searches for a^+ - and y -ions in addition to c - and z^+ -ions. However, what is evident from Figure 1 is the minimal gain achieved by including these additional ion types, with very slight differences in total and unique identifications when searching fully processed data with OMSSA and using only c - and z^+ -ions as compared to ZCore. Removal of the precursor was the single most beneficial step in OMSSA and SEQUEST, but had little effect with Mascot. This makes sense, as Mascot automatically performs this step, SEQUEST does not remove the precursor ion, and while OMSSA does perform this processing step, it does so in a crude manner, simply assuming a +4 precursor and removing a ± 5 Th window from the anticipated precursor. Though the newest release of OMSSA allows for removal of CRPs, the default is to not remove these peaks, and this parameter was not set to explicitly remove the precursor in these searches. As a result, the additional removal of CRPs from the spectra led to only slight increases in identifications for SEQUEST and Mascot, with a somewhat larger increase for OMSSA. Finally, removal of the spectral areas which likely contained m/z peaks resulting from neutral losses led to a marked increase in both total and unique identifications for OMSSA and Mascot, with very little effect on SEQUEST identifications.

To alleviate any concerns of removing a large number of true peaks from the spectra, Supplemental Figure 3 provides the average percentage of potential matching c - and/or z^+ -type fragment ion peaks per peptide which fell within the m/z ranges that were removed during the spectral processing steps. Only a slight percentage of the total possible fragment ions were contained within the removed m/z region(s), even when performing all processing steps ($\sim 2.9\%$ for yeast and 3.6% for human). As anticipated, the greatest number of possible true c - and z^+ -ions fell within the large m/z window in which common neutral loss-associated peaks are present ($\sim 2.3\%$ for yeast and 2.9% for human). At first glance, the fact that there are less c - and z^+ -ions within the regions removed for CRPs than within the precursor region (0.25% vs. 0.37% for yeast and 0.31% vs. 0.38% for human) may seem surprising. However, this is most likely due to the increased number of peptides at charge +3 and below which are generated using Lys-C. While the precursor typically lies in a lower m/z region of the spectrum where there are many true peaks, the CRPs occur at higher m/z values, where there are typically a lower number of informative peaks. Also, removal of the window around the singly charged precursor will not contain any true c - and z^+ -ions, and as the dataset consists mainly of these lower-charged species, on average only one CRP greater than charge +1 will exist within a given spectrum.

As the use of electron-based fragmentation methods has become more frequent in large-scale proteomics experiments, the majority of commonly employed database search algorithms have

failed to adequately adapt their search methodology to account for the unique spectral features these techniques produce. Here, we have detailed the results of removing several of these features prior to submission to a database search on total and unique identifications at an ~1% FDR. These results were also compared to a recently developed database search algorithm adapted specifically for searching peptide spectra produced by electron-based fragmentation. Of the three search algorithms which were originally created to search CAD-generated spectra, all showed an increase in identifications when searching fully pre-processed ETD spectra. However, each algorithm provided a unique response to removal of specific interfering peaks within a spectrum, thus giving insight into how the algorithms are affected by confounding peaks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Danielle Swaney for supplying the yeast and hES cell data. DMG is grateful for support from an NIH pre-doctoral fellowship (Biotechnology Training Program, NIH 5T32GM08349). The University of Wisconsin, the Beckman Foundation, Eli Lilly, and the National Institutes of Health (R01GM080148 and P01GM081629) provided financial support for this work.

References

1. Good DM, Wenger CD, McAlister GC, Bai DL, et al. Post-Acquisition ETD Spectral Processing for Increased Peptide Identifications. *Journal of the American Society for Mass Spectrometry* 2009;20:1435–1440. [PubMed: 19362853]
2. Geer LY, Markey SP, Kowalak JA, Wagner L, et al. Open mass spectrometry search algorithm. *Journal of Proteome Research* 2004;3:958–964. [PubMed: 15473683]
3. Eng JK, McCormack AL, Yates JR. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* 1994;5:976–989.
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
5. Sadygov RG, Good DM, Swaney DL, Coon JJ. A new probabilistic database search algorithm for ETD spectra. *J Proteome Res* 2009;8:3198–3205. [PubMed: 19354237]
6. Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nature Methods* 2008;5:959–964. [PubMed: 18931669]
7. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 2007;4:207–214. [PubMed: 17327847]

Figure 1.

Change in total and unique peptide identifications at an ~1% false discovery rate for each of the four search algorithms tested after varying levels of spectral pre-processing. For all search algorithms except ZCore, spectral processing increased both total and unique identifications.

Overlap in Unique Identifications

Figure 2.

The greatest overlap in unique peptide identifications for the four algorithms was the overlap between all algorithms, as expected. The leading contributor to unique identifications for the yeast dataset was ZCore (845 identifications), while OMSSA contributed the greatest number of unique identifications for the human dataset (548 identifications).