research articles Journal of proteome research

Database Search Strategies for Proteomic Data Sets Generated by Electron Capture Dissociation Mass Spectrometry

Steve M. M. Sweet,[†,#] Andrew W. Jones, Debbie L. Cunningham,[†] John K. Heath,[†] Andrew J. Creese, and Helen J. Cooper*

School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

Received May 28, 2009

Large data sets of electron capture dissociation (ECD) mass spectra from proteomic experiments are rich in information; however, extracting that information in an optimal manner is not straightforward. Protein database search engines currently available are designed for low resolution CID data, from which Fourier transform ion cyclotron resonance (FT-ICR) ECD data differs significantly. ECD mass spectra contain both z-prime and z-dot fragment ions (and c-prime and c-dot); ECD mass spectra contain abundant peaks derived from neutral losses from charge-reduced precursor ions; FT-ICR ECD spectra are acquired with a larger precursor *m*/*z* isolation window than their low-resolution CID counterparts. Here, we consider three distinct stages of postacquisition analysis: (1) processing of ECD mass spectra prior to the database search; (2) the database search step itself and (3) postsearch processing of results. We demonstrate that each of these steps has an effect on the number of peptides identified, with the postsearch processing of results having the largest effect. We compare two commonly used search engines: Mascot and OMSSA. Using an ECD data set of modest size (3341 mass spectra) from a complex sample (mouse whole cell lysate), we demonstrate that search results can be improved from 630 identifications (19% identification success rate) to 1643 identifications (49% identification success rate). We focus in particular on improving identification rates for doubly charged precursors, which are typically low for ECD fragmentation. We compare our presearch processing algorithm with a similar algorithm recently developed for electron transfer dissociation (ETD) data.

Keywords: ECD • neutral loss • OMSSA • Mascot • identification • CID • mass spectrometry • FT-ICR • LTQ-FT

Introduction

Electron capture dissociation (ECD) is a radical-driven fragmentation technique which provides an alternative to slow-heating collision induced dissociation (CID).[1] ECD has successfully been applied to the small-scale detailed characterization of various peptides, modified or otherwise.[2,3] These experiments are greatly facilitated by a prior knowledge of the peptide sequence, allowing manual analysis of the ECD data. In contrast, large-scale proteomic experiments utilizing ECD rely on a database search step in order to identify the fragmented peptide.[4,5] The database search engines employed were originally designed to accept low resolution CID data. High resolution ECD data presents a significantly different challenge. The characteristics of FT-ICR ECD data are sub-10 ppm mass accuracy, low noise levels, intense precursor and charge-reduced precursor peaks, and strong neutral loss peaks from the charge-reduced precursor.[6,7] Furthermore, hydrogen transfer can occur between ECD c-prime and z-dot fragments, resulting in c-dot and z-prime products.[8]

The search engines that have been employed for large-scale ECD data analysis are Mascot and OMSSA.[4,5] These search engines have certain limitations, for example, the product ion tolerance cannot be specified in ppm and the benefits of high mass accuracy data are not fully realized. We have analyzed large-scale ECD data sets both manually and using various search engines. It is apparent from these analyses that certain generic aspects of ECD mass spectra are likely to be detrimental to their identification by database search engines. The most obvious of these is the high intensity precursor and charge-reduced precursor peaks. Both search engines tested here already anticipate these peaks, removing them from consideration. For example, Mascot removes peaks within the fragment ion tolerance window about each of the precursor isotope peaks. However, the search engines do not consider coeluting peaks in the precursor isolation window and are, in fact, ignorant of the isolation window size. Another characteristic of ECD is the generation of various neutral losses from the charge-reduced precursors. These peaks are not utilized by currently available search engines. In the case of ECD of doubly

* Address correspondence to: Helen J. Cooper, School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. Telephone: +44 (0)121 414 7527. Fax: +44 (0)121 414 5925. E-mail: H.J.Cooper@bham.ac.uk.

† CRUK Growth Factor Group.

Current address: Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

10.1021/pr9008282 CCC: $40.75

Published on Web 10/13/2009

charged precursors, these neutral loss peaks occupy a region of the m/z range which is either free of c, z and y fragment ions or can only contain certain large c, z and y ions. Previous observations in our laboratory reveal that ECD mass spectra from doubly charged precursors typically give lower scores and identification rates than more highly charged precursors. In part this is likely due to the fragmentation process itself: electron capture efficiency increases as the square of the ion charge hence doubly charged precursors are less likely to capture an electron. In addition, only one of the resulting fragments can retain the single positive charge. We test whether the identification rate for ECD mass spectra from doubly charged precursors can be increased through removal of some of the aforementioned uninformative peak types. Recently, Good et al.[9] developed a similar presearch processing algorithm for electron transfer dissociation (ETD) data. The algorithm was tested on a low resolution ETD data set. The Good algorithm differs from that described here: (1) rather than remove the entire neutral loss region from the mass spectrum, our algorithm retains all potential true peaks; (2) our algorithm does not remove neutral loss regions from higher (>2+) charge-state precursor ions; (3) our algorithm allows removal of noise peaks. We compare our algorithm with the Good algorithm for high resolution ECD data.

Methods

Cell Culture and Sample Preparation. Mouse fibroblast NIH 3T3 cells were cultured and lysed as previously described.[5] Proteins were reduced and alkylated (DTT and iodoacetamide), digested using trypsin and the resulting peptides separated by SCX chromatography, again as described previously.[5]

Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS). Online liquid chromatography was performed by use of a Micro AS autosampler and Surveyor MS pump (Thermo Fisher Scientific, Bremen, Germany). Peptides were loaded onto a 75 μm (internal diameter) Integrafrit (New Objective) C8 resolving column (length 10 cm) and separated over a 40 min gradient from 0% to 40% acetonitrile (Baker, Holland). Peptides eluted directly (~350 nL/min) via a Triversa nanospray source (Advion Biosciences, NY) into a 7 T LTQ FT mass spectrometer (Thermo Fisher Scientific), where they were subjected to data-dependent CID and ECD.

Data-Dependent CID and ECD (DD-CID-ECD). The mass spectrometer alternated between a full FT-MS scan (m/z 400−1600) and subsequent CID and ECD MS/MS scans of the most abundant ion above a threshold of 40 000. Survey scans were acquired in the ICR cell with a resolution of 100 000 at m/z 400. Precursor ions were subjected to CID in the linear ion trap. The width of the precursor isolation window was 6 m/z. Only multiply charged precursor ions were selected for MS/MS. CID was performed with helium gas at a normalized collision energy of 35%. Automated gain control was used to accumulate sufficient precursor ions (target value 5×10^4, maximum fill time 0.2 s). For the ECD event, precursor ions were isolated in the ion trap and transferred to the ICR cell. Isolation width was 6 m/z. Automated gain control was used to accumulate sufficient precursor ions (target value 1×10^6 per microscan, maximum fill time 1 s). The electrons for ECD were produced by an indirectly heated barium tungsten cylindrical dispenser cathode (5.1 mm diameter, 154 mm from the cell, 1 mm off-axis). The current across the electrode was ~1.1 A. Ions were irradiated for 60 ms at 5% energy (corresponding to a cathode potential of −2.775 V). Each ECD scan comprised 4 coadded microscans, acquired with a resolution of 25 000 at m/z 400. Dynamic exclusion was used with a repeat count of 1 and an exclusion duration of 60 s. Data acquisition was controlled by Xcalibur 2.0 and Tune 2.2 software (Thermo Fisher Scientific, Inc.).

Data Analysis. DTA files were created from the raw data using Bioworks 3.3.1 (Thermo Fisher Scientific, Inc.). The DTA files were either searched directly, or preprocessed using a perl script to remove non-c,z,y peaks as described in the text. The perl script is available as a Supplementary File online. The regions removed from all ECD mass spectra consisted of (1) a prominent noise peak (m/z 101.7−102.1) and (2) the isolation window around the precursor ion (precursor ± 3 m/z). For ECD mass spectra of $[M + 2H]^{2+}$ ions, the region containing neutral loss peaks from the charge-reduced precursor was modified as follows: in the region $[M + 2H]^+ > m/z > ([M + 2H]^+ - 57)$, no peaks were retained. (No c,z,y peaks can fall in this region as the smallest amino acid residue, glycine, has a mass of 57 Da.) Within the region $([M + 2H]^+ - 57) > m/z > ([M + 2H]^+ - 140)$, the peaks that could correspond to c ions, z prime ions, z dot ions, and y ions were retained. The list of masses retained, annotated with their corresponding fragment, is given in Supporting Information Table 1. The window around each retained m/z was set at ±12 ppm.

For comparison, DTA files were also generated using the DTA Generator developed for postacquisition ETD processing by Good, et al.[9] These DTA files were either left unprocessed (1), processed with the Good ETD algorithm (2) or processed as above with our algorithm (3). Options: Fragment ion intensity, 0.7% (relative). Assumed precursor charge state range: 2−8. ETD Cleaning algorithm: none (1 and 3). Output: individual DTAs. Or for (2) ETD Cleaning algorithm: smart. Clean precursor, charge-reduced precursor, clean NL from charge-reduced precursor.

Searches were carried out against a concatenated database consisting of the mouse IPI database (Version 3.40) supplemented with common contaminants (including keratins, trypsin, BSA) and the reversed-sequence version of the same database. The final database contained 107 688 protein entries (53 844 of which were reversed-sequence versions). CID and ECD data were searched separately.

Mascot (version 2.2.; Matrix Science, U.K.) was searched using the following parameters for ECD data: enzyme, trypsin; maximum missed cleavages, 2; fixed modification, carbamidomethyl (C); variable modifications, acetyl (protein N-terminus), Oxidation (M); Peptide tolerance, 1.1 Da (or as stated in the text); MS/MS tolerance, 0.02 Da; Instrument, FTMS-ECD (2+ fragments if precursor 3+ or higher; ion types, c, y, z + 1, z + 2 [corresponds to z-dot and z-prime]). Settings for CID data were as above, except MS/MS tolerance, 0.5 Da; Instrument, ESI-TRAP (2+ fragments if precursor 2+ or higher; ion types, b, b with NH_3 loss if b significant and fragment contains RKNQ, b with water loss if b significant and fragment contains STED, y, y with NH_3 and water losses (as for b)).

Mascot search results were exported ("Format as: export search results") using the following settings: export format, CSV; significance threshold <1; Max. number of hits, 16 000. All other settings were left as default. The number of proteins exported was checked, to ensure it did not reach the number specified in "Max. number of hits", above. This indicates that all identifications, however low-scoring, were exported. A perl script was used to copy the protein identifier into every peptide identification row for each protein identified.

Exported Mascot results were sorted (in Excel) by descending score, and then by protein accession (A to Z, in this case, to ensure ###REV... is listed before IPI:...). This was to ensure that, for a reverse and a forward hit of identical score, the reverse hit was preferentially retained. Lower scoring identifications were removed, leaving only the top scoring identification for each MS/MS event (remove duplicates for Peptide Scan Title column). The mass error in ppm was calculated for each identification (using the charge-state, theoretical mass and delta mass: delta mass/(theor. mass + chargestate × 1.00728) × 1 000 000).

The OMSSA Browser 2.1.1 was employed to search the DTA files. OMSSA settings were as previously described, with the exception that phosphorylation was not considered as a variable modification.[5] For the ECD search, the 'elimination of charge-reduced precursors in spectrum' option was selected.

Researching ECD Phosphopeptide Data Set. A total of 6080 ECD DTAs were processed as above, that is, removal of noise peak, precursor window and neutral loss peaks. Database searching was as for unmodified (above), but allowing STY phosphorylation as a variable modification. Postsearch filtering was as above.

Results

To test the effect of various search-related parameters, we employed a test data set consisting of 3341 high quality ECD mass spectra obtained from the LC-MS/MS analysis of mouse whole cell lysate. Paired ion trap CID and FT-ICR ECD mass spectra were acquired, as previously described.[4,5] The mouse IPI database was searched; a concatenated forward-reverse version of this database was employed, unless stated otherwise. In all cases, the false-discovery rate (FDR) as estimated by the number of accepted reverse identifications was controlled at less than 1%. Full details of the peptide identifications are supplied as Supplementary Data.

Initial Search. An initial search, without preprocessing of the CID or ECD data, was carried out using both search engines: Mascot and OMSSA. The precursor mass tolerance was set to 0.02 m/z (OMSSA) or 10 ppm (Mascot).

For the initial Mascot search, a forward-only version of the mouse IPI database was employed, in combination with the Mascot 'decoy' option. The 'decoy' option automatically carries out a second search using a randomized database, and thereby gives an estimate of FDR. However, adjusting the FDR to a particular value (1%) was not possible. The search resulted in 633 ECD identifications and 1712 CID identifications. To better control the estimated FDR, we repeated the Mascot search, without "decoy" option, using the concatenated version of the database (as used in all subsequent searches), exported *all* results into Excel, and manually filtered according to Mascot score. That resulted in a doubling of the number of accepted identifications, as shown in Table 1 (ECD Search: row 1 versus row 3). Manually filtered Mascot and OMSSA searches give similar numbers of identifications for both ECD and CID data sets. The identification rates reach 38% for ECD data (1254 identifications) and 69% for CID data (2297 identifications). Clearly, there is a considerable difference, of approximately 30%, in identification success rate between CID and ECD mass spectra.

Postsearch Filtering by Precursor Mass Error. Database searches employing a wide precursor mass tolerance window, with subsequent filtering of results, have previously been shown to improve identification rates.[5,10] While the benefits of post-search filtering are well-established, we were interested in comparing the magnitude of its effect with the other levels of data processing described here and the effectiveness for ECD compared to CID data. We therefore repeated the above searches with a precursor mass tolerance of 1.1 Da and exported all results for subsequent manual filtering of identifications by precursor error (in ppm) and, if the selected ppm range contains more reverse hits than compatible with a 1% FDR, by peptide score. This resulted in an increased number of accepted identifications for all searches (Table 2). We note that the increase in identification efficiency for ECD data is greater than that for CID data, for example, increases of 6.4% and 3.1%, for ECD and CID searches using Mascot. This characteristic may be the result of the lower peptide scores for ECD identifications (Mascot average score of 23 versus 40, for ECD (n = 1468) and CID (n = 2385), respectively), that is, the true identifications are less readily distinguished from reverse hits on the basis of peptide score alone.

Previous work has shown that it is possible for the precursor mass recorded in the DTA file to correspond to the second or third isotopic peak (i.e., one or two ^{13}C more than the monoisotopic peak).[11] This occurrence is particularly common for low resolution ion-trap only experiments. If this occurs, identifications can be rescued by searching with a larger precursor tolerance window (with subsequent narrow mass filtering around the offset precursor mass). We compared searches with 1.1, 2.1, and 3.1 Da tolerances. In none of the cases was there a high-scoring identification resulting from

Table 1. Initial Searches of ECD and CID Data Filtered According to Database Search Algorithm Score[a]

search	postsearch filter	forward hits	reverse hits	ID rate
ECD Search (3341 DTAs)				
Mascot; 10 ppm precursor; Mascot "decoy"		633	2*	18.9
OMSSA; 0.02 Da precursor	Peptide score	1190	11	35.6
Mascot; 10 ppm precursor	Peptide score	1254	12	37.5
CID Search (3341 DTAs)				
Mascot; 10 ppm precursor; Mascot "decoy"		1712	16*	51.2
OMSSA; 0.02 Da precursor	Peptide score	2297	22	68.8
Mascot; 10 ppm precursor	Peptide score	2283	22	68.3

[a] DTA files are unaltered. Asterisks indicate "decoy" hits, from Mascot "Decoy" search option.

Table 2. Searches of ECD and CID Data in Which a Wider Precursor Mass Tolerance Window Was Combined with Postsearch Precursor Mass Error Filtering[a]

search	postsearch filter	forward hits	reverse hits	ID rate
ECD Search (3341 DTAs)				
OMSSA; 1.1 Da precursor	Precursor ppm error and peptide score	1447	14	43.3
Mascot; 1.1 Da precursor	Precursor ppm error	1468	7	43.9
CID Search (3341 DTAs)				
OMSSA; 1.1 Da precursor	Precursor ppm error	2344	9	70.2
Mascot; 1.1 Da precursor	Precursor ppm error	2385	9	71.4

[a] To achieve the estimated FDR of 1%, results were filtered according to database search algorithm scores where necessary (E-value cutoff of 8.01×10^{-1} for OMSSA ECD search).

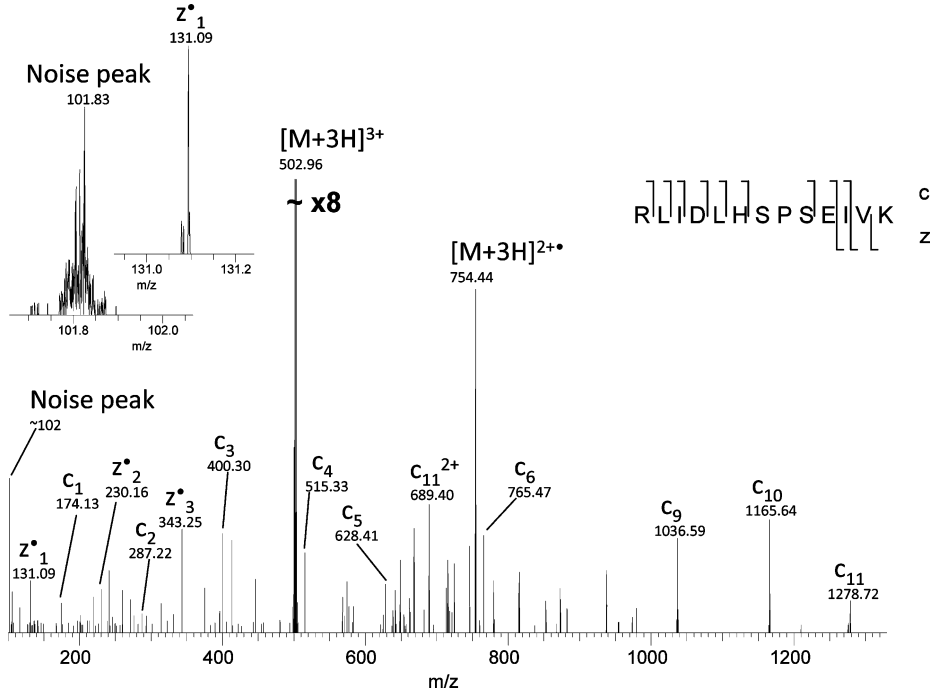

Figure 1. Noise peak at *m*/*z* ~102. Insets show enlargements of the noise peak and a true fragment peak of similar *m*/*z*.

incorrect precursor determination. Precursor mass tolerances of 2.1 and 3.1 Da do not increase the number of identifications, indicating that the Bioworks-generated DTA contains the correct monoisotopic mass. When the second or third isotopic peak was selected for fragmentation, the mass in the DTA file generated by Bioworks was automatically corrected to the monoisotopic value. (As a result of the high precursor threshold for fragmentation, the true monoisotopic ion was always clearly detected.)

Preprocessing of ECD Spectra. Electrical Noise Peak. The results of the initial searches, as reported in Tables 1 and 2, indicate an identification success rate of approximately 70% for CID mass spectra and 45% for ECD mass spectra. We examined the ECD spectra for generic features which may be detrimental to the identification success rate. Electrical noise peaks are unfortunate but commonplace features of mass spectra. In all FT-ICR data from our instrument, we observe a peak of this kind at *m*/*z* 102. No equivalent peaks are observed in the LTQ linear ion trap mass spectra. The peak is broad and lacks isotope peaks, easily distinguishing it from true analyte peaks (Figure 1). Removing the region in which this peak occurs (101.7−102.1 *m*/*z*) from the DTA files is a straightforward way to improve the ratio of informative to uninformative peaks for the database search. This increases the Mascot and OMSSA identification scores and therefore number of identifications (Tables 3 and 4, row 2). The noise peak removal could in theory also remove a z-dot ion from C-terminal valine; however, this will have no effect in practice, since a vast majority of tryptic peptides have lysine or arginine at their C-terminus. (The noise peak would also obscure the valine z-dot ion.)

Coeluting Peaks of Similar *m*/*z* (within the Precursor Isolation Window). To maximize the transfer of large numbers of precursor ions from the linear ion trap into the ICR cell for ECD fragmentation, a precursor isolation window larger than that selected for LTQ CID is usually employed.[12] This window typically ranges from 3 to 10 *m*/*z*.[5,13] Here, we use an isolation width of 6 *m*/*z*. During analysis of complex proteomic samples,

Table 3. Effect of Presearch Trimming of ECD DTA files on the Number of Mascot Identifications[a]

presearch	forward hits (2+)	reverse hits	ID rate
ECD Search (3341 DTAs) Mascot			
No DTA trimming	1468 (870)	7	43.9
Remove noise peak	1509 (898)	12	45.2
Remove precursor window	1556 (934)	9	46.6
Remove precursor window. Remove all nonfragments (MH to MH − 140 *m*/*z*).	1586 (964)	13	47.5
Remove noise peak and precursor window	1609 (972)	10	48.2
Remove noise peak and precursor window. Remove all nonfragments (MH to MH − 140 *m*/*z*).	1643 (1006)	16	49.2

[a] Identifications from doubly-charged precursors shown in parentheses.

coelution of peptides of similar *m*/*z* is frequently observed.[11] In contrast to the LTQ CID process, peaks in the isolation window are not fragmented to completion during ECD making any coelution of peptides more readily apparent. Both OMSSA and Mascot have options to remove the intact precursor peak (these options are employed in all the searches described here); however, this only removes the precursor and its isotopes according to the user-defined fragment ion mass tolerance, and not any coeluting peptide or noise peaks within the precursor isolation window. We investigated whether removing the complete isolation window from the DTA files prior to database searching would improve the resulting identifications (Tables 3 and 4). An example of an additional identification resulting from the removal of the isolation window is shown in Figure 2. It should be noted that *bona fide* fragment peaks may inadvertently be removed during this trimming process. We remove the 6 *m*/*z* region around the precursor. For a typical 1500 Da, 2+ tryptic peptide, the acquired MS/MS ECD mass spectrum ranges from 100 to 1500 *m*/*z*. Therefore, we remove

Table 4. Effect of Presearch Trimming of ECD DTA Files on OMSSA Identifications[a]

presearch	forward hits (2+)	reverse hits	ID rate
ECD Search (3341 DTAs) OMSSA			
No DTA trimming	1390 (733)	12	41.6
Remove noise peak	1409 (745)	14	42.2
Remove precursor window	1453 (786)	13	43.5
Remove precursor window. Remove all nonfragments (MH to MH − 140 m/z).	1470 (803)	13	44.0
Remove noise peak and precursor window	1469 (801)	13	44.0
Remove noise peak and precursor window. Remove all nonfragments (MH to MH − 140 m/z).	1480 (813)	14	44.3

[a] Identifications from doubly-charged precursors shown in parentheses. As processing also affects the scores of the reverse hits, the optimum *E*-value cutoff varies between 2.36×10^{-1} and 8.01×10^{-1}. For direct comparison, the more conservative value of 2.36×10^{-1} is used for all searches.

0.04% of this range. The fraction of useful fragment peaks removed is expected to be less than 0.04%, as fragment peaks coinciding with the precursor will, in any case, not be distinguishable from the high intensity precursor. The increase in identifications shown in Tables 3 and 4 indicates that the net effect is positive. By plotting every identification score, with and without removal of the precursor isolation window, we see that a vast majority of identifications result in an increased score (Figure 3). (In a small number of cases, a previously accepted identification was replaced with an unacceptable identification (reverse hit or unacceptable mass error), due to the score of the unacceptable identification increasing by more than the score of the previously accepted identification. These cases are plotted as a score of zero.)

To control for any unanticipated effects of this trimming process, a window of the same size (6 m/z), but shifted away from the precursor isolation region (+25 m/z) was removed. This had no net effect on the number of ECD identifications (data not shown). Removing the precursor isolation window from CID DTAs also has no net effect, positive or negative (data not shown). This observation agrees with the lack of high intensity intact precursor surviving in this region during the ion-trap excitation event, leaving no intense peaks to potentially confound the database search.

Neutral Losses from Charge-Reduced Precursor. A salient feature of ECD is the potential for neutral losses from the charge-reduced precursor.[6,7] These losses are particularly evident in ECD mass spectra of doubly charged precursors. The tryptic peptides are predominantly doubly charged: 67% of the ECD mass spectra collected were from doubly charged precursors. However, the success rate for the ECD identification of doubly charged precursors was lower than that for triply charged species (32% versus 54%, for the Mascot unaltered search). Figure 4 shows the proportion of identifications by charge-state for both CID (4b) and ECD (4c), alongside the DTA input proportion (4a). The observed lower success rate for ECD identifications from 2+ precursors agrees with the previously reported data for ETD of 2+ and 3+ precursors.[14] We hypothesized that neutral loss peaks from the charge-reduced precursor which are not anticipated by the database search engine might be detrimental to the identification of doubly charged peptides. Rather than remove specific neutral losses from the 2+ DTAs, we chose to retain all potential true fragment ions within 140 m/z of the charge-reduced precursor, specifically all possible c, z, z-prime and y ions from tryptic peptides were listed for retention. (The m/z values of potential true fragment ions were calculated on the basis of the following known parameters: m/z of the charge-reduced precursor, mass of amino acid residues, and structures of c, z, z-prime and y ions.) That resulted in a net increase in 2+ identification efficiency of 3.5% and 1.5% for Mascot and OMSSA, respectively. The cumulative effect of the three precursor trimming operations on the number of 2+ identifications is an increase of 16% for Mascot (from 870 to 1006) and 11% for OMSSA (733 to 813). Figure 5 plots the Mascot scores of all 1006 2+ identifications after removing the neutral loss peaks, alongside the scores prior to the additional processing. For 98.5% of the previously identified 2+ peptides, the same peptide was identified. In the remaining 15 cases, the previously accepted identification was relegated to second place, behind either a decoy hit or a hit with an unacceptably large mass error (presumed false-positive). In some cases, it was not clear why the Mascot score of the previously accepted hit failed to increase as much as the alternative identification. This is likely related to the overall number of peaks present and the windows into which Mascot divides the spectrum (D. Creasy, personal communication). Large fragment ions from nontryptic peptides (e.g., C-terminal peptides) could potentially be lost; however, in none of these 15 cases did we observe the loss of a real fragment ion from the Mascot identification.

We compared the strategy described above (retaining all true fragment ions) with the simpler approach of removing the entire 140 m/z region below the $[M + 2H]^{\bullet}$ peak, followed by a Mascot search and postsearch filtering as before. The identification rate was reduced, with 7% fewer identifications from doubly charged precursors (936 vs 1006). The removal of any of the 33 true c/z/y fragment ions which fall in this region will be detrimental to identification.

Validation of Identifications. A large number of the accepted ECD identifications have low Mascot peptide scores. On the basis of the Mascot scores alone, these identifications would usually be rejected; however, the FDR estimate suggests that the majority are correct. We employ two strategies in order to validate the ECD identifications: (1) check for agreement between identifications from paired CID and ECD events; (2) manually assess a small number of low-scoring ECD identifications, where no paired CID identification was made.

Of the 1643 ECD identifications, there are only 83 identifications without a paired CID identification, that is, 1560 (95%) of the ECD identifications are from paired events which also led to a CID identification. This high degree of overlap is due to the high CID success rate (>70%), which is in turn attributable to the fragmentation of high intensity (>40 000 counts) tryptic peptides. The estimated FDR of less than 1% predicts fewer than 15.6 false-positives in this ECD data set and the same number in the CID data set. The FDR therefore predicts fewer than 31.2 identification conflicts (as a false-positive for either the ECD or CID identification will result in a conflict). In fact we find 21 conflicts, well within the limit of 31 expected for a 1% FDR (Supplementary Table 2). Of these conflicts, 18 out of 21 are isobaric peptides, often with similar sequences (e.g., VAPDEHPILLTEAPLNPK and VAPEEHPVLLTEAPLNPK).

To examine the distribution of conflicts, the 1560 ECD identifications with paired CID identifications were divided

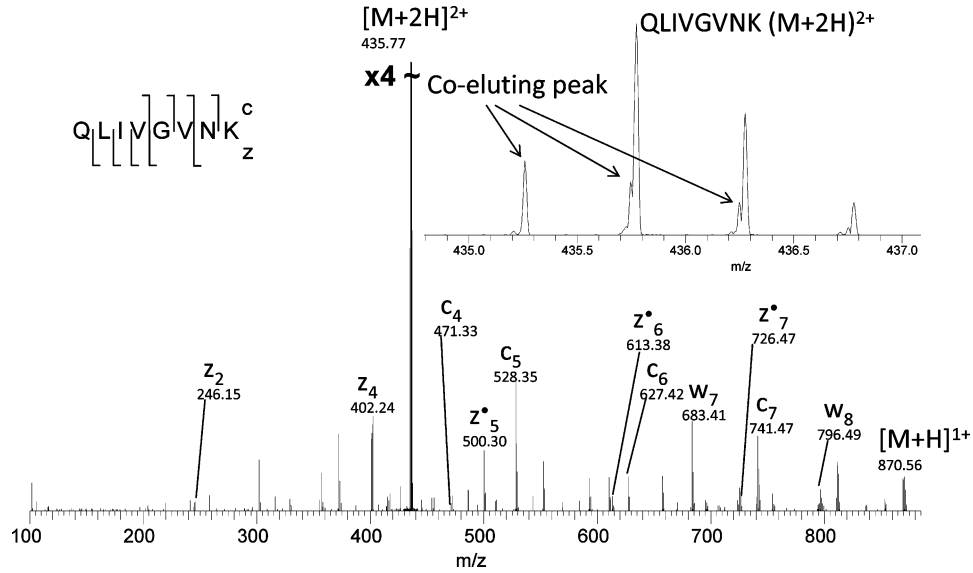

Figure 2. Example of multiple precursor ions in isolation window. The identification was made by both Mascot and OMSSA after removal of the entire precursor window. Unassigned peaks may be due to fragmentation of the unidentified coeluting precursor. (Note that the noise peak has not been removed from this mass spectrum.)

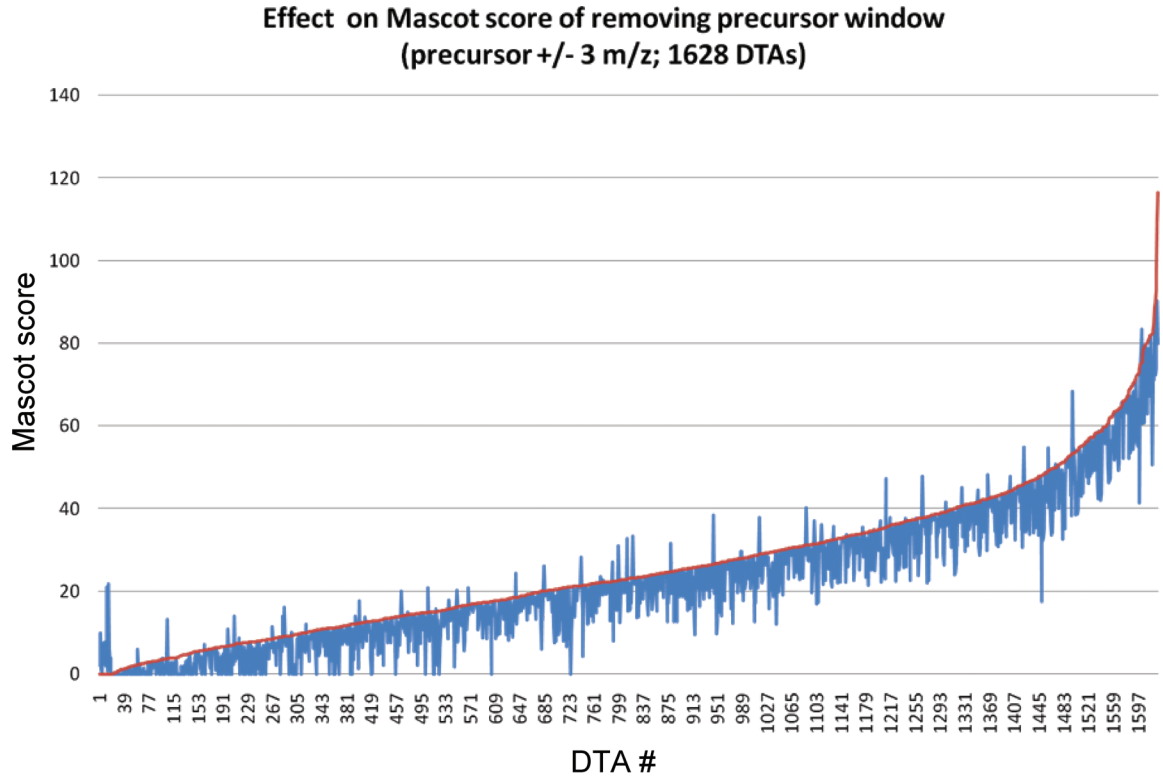

Figure 3. Effect on Mascot score of removing precursor window. A total of 3341 ECD DTAs (noise peak removed) were searched without precursor window removal and with precursor window removal resulting in 1509 and 1609 identifications, respectively. A total of 1628 DTA were identified in one or both searches. The identifications are plotted, by ascending Mascot score of trimmed version (precursor window removed) (red). The identification score for each DTA prior to trimming is shown alongside (blue). Reverse hits or rejected hits (unacceptably large ppm error) were assigned a score of zero.

evenly into 10 bins, according to descending ECD Mascot peptide score. The 21 conflicts and 16 reverse hits from the initial search are shown separated across these bins, according to the ECD Mascot score (Figure 6). Both conflicts and reverse hits cluster around the lower Mascot scores, as would be expected; however, even the lowest-scoring 10% of hits (Mascot scores from 7.27 to 0.04) show high agreement with CID identifications (96%) (6 conflicts) and an estimated FDR of only 6% (10 reverse hits).

To further validate these extremely low-scoring ECD identifications, the 83 identifications without a paired CID identification were examined manually. Characteristics of true positive FT-ICR ECD identifications are the following: fragment ion errors smaller than 12 ppm; y fragment ions, if present, of

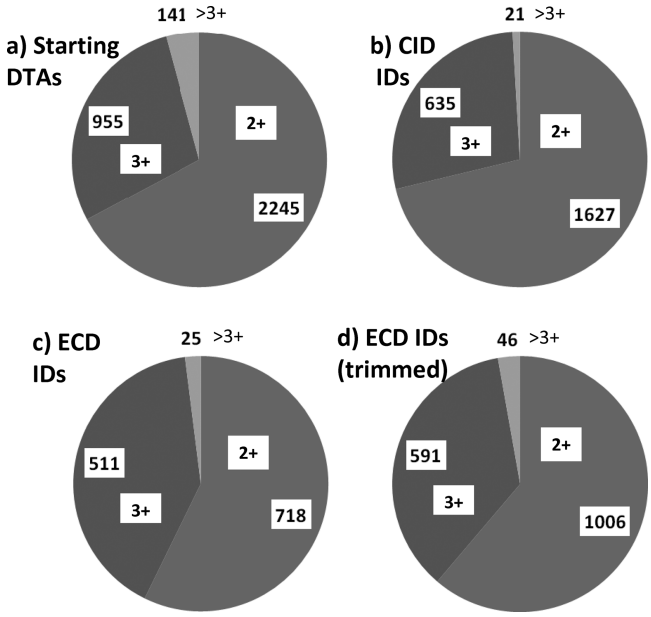

Figure 4. DTAs and identifications by charge-state. (a) Data set of 3341 DTAs (CID and ECD pairs); (b) CID identifications (2283) using Mascot with subsequent filtering; (c) ECD identifications (1254) using Mascot with subsequent score filtering; (d) ECD identifications (1643) using trimmed DTAs, Mascot and subsequent score and mass accuracy filtering.

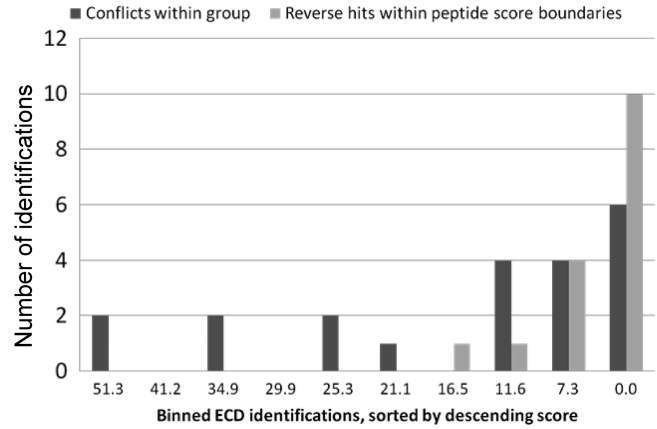

Figure 6. Distribution of identification conflicts and reverse hits by ECD identification Mascot peptide score. The abscissa shows 1560 ECD identifications, with paired CID identifications, binned according to descending Mascot score: ten bins, each containing 156 ECD IDs, labeled with lowest score in bin.

lower intensity than z fragments (with the exception of y fragments N-terminal to proline); runs of consecutive c or z fragments. The manual validation suggests that the number of false-positives within this subset of 83 identifications is greater than for the data set overall, with four putative false positives giving a FDR of 5%. Nonetheless, 95% of these ECD identifications pass manual scrutiny. Two examples of low-scoring ECD identifications are shown in Figure 7.

Comparison with Existing ETD Spectral Processing Algorithm. In recent work by Good et al., an algorithm designed for ETD fragmentation mass spectra preprocessing was re-

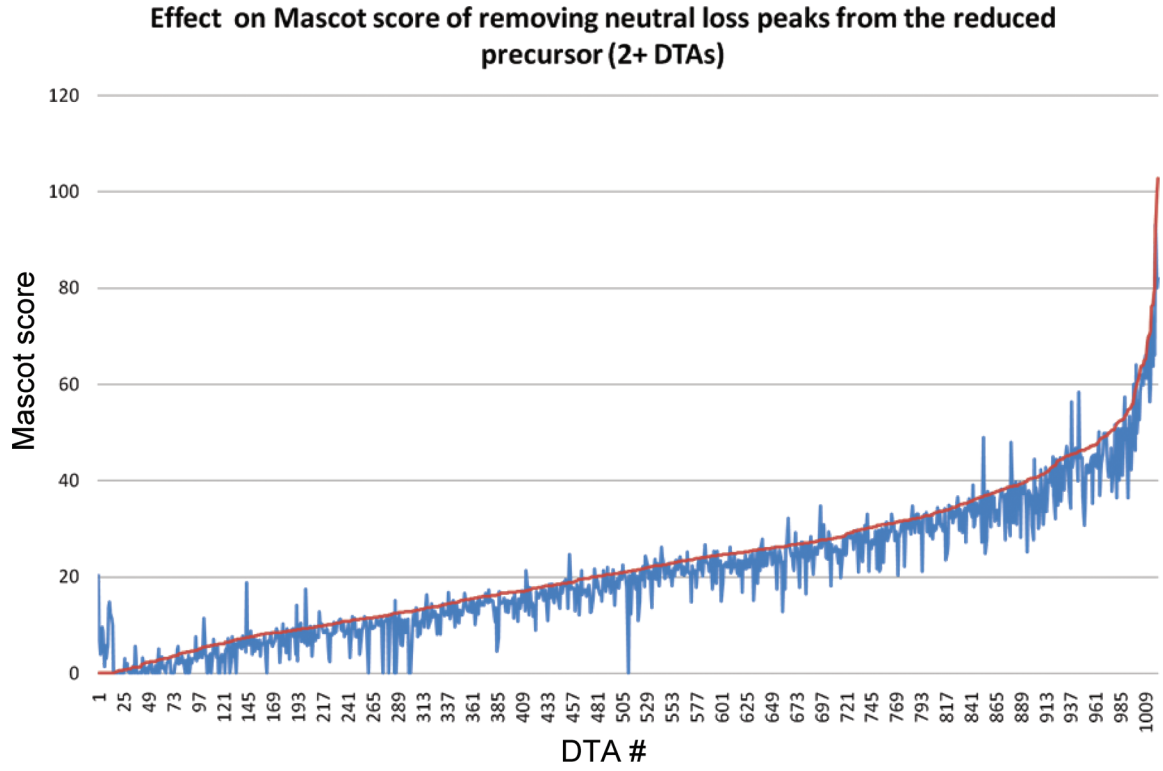

Figure 5. Effect on Mascot score of removing non-c, z, y fragment peaks within 140 *m*/*z* of the charge-reduced precursor (RP-140) of doubly charged peptides. A total of 3341 ECD DTAs were searched, without neutral loss peak removal and with neutral loss peak removal resulting in 972 and 1006 2+ identifications, respectively. A total of 1021 DTAs resulted in an identification in one or both searches. The identifications are plotted, by ascending Mascot score of RP-140 trimmed version (red). The identification score for each DTA prior to RP-140 trimming is shown alongside (blue). In both cases, the precursor window and noise peak at *m*/*z* 102 were removed. Reverse hits or rejected hits (unacceptably large ppm error) were assigned a score of zero.

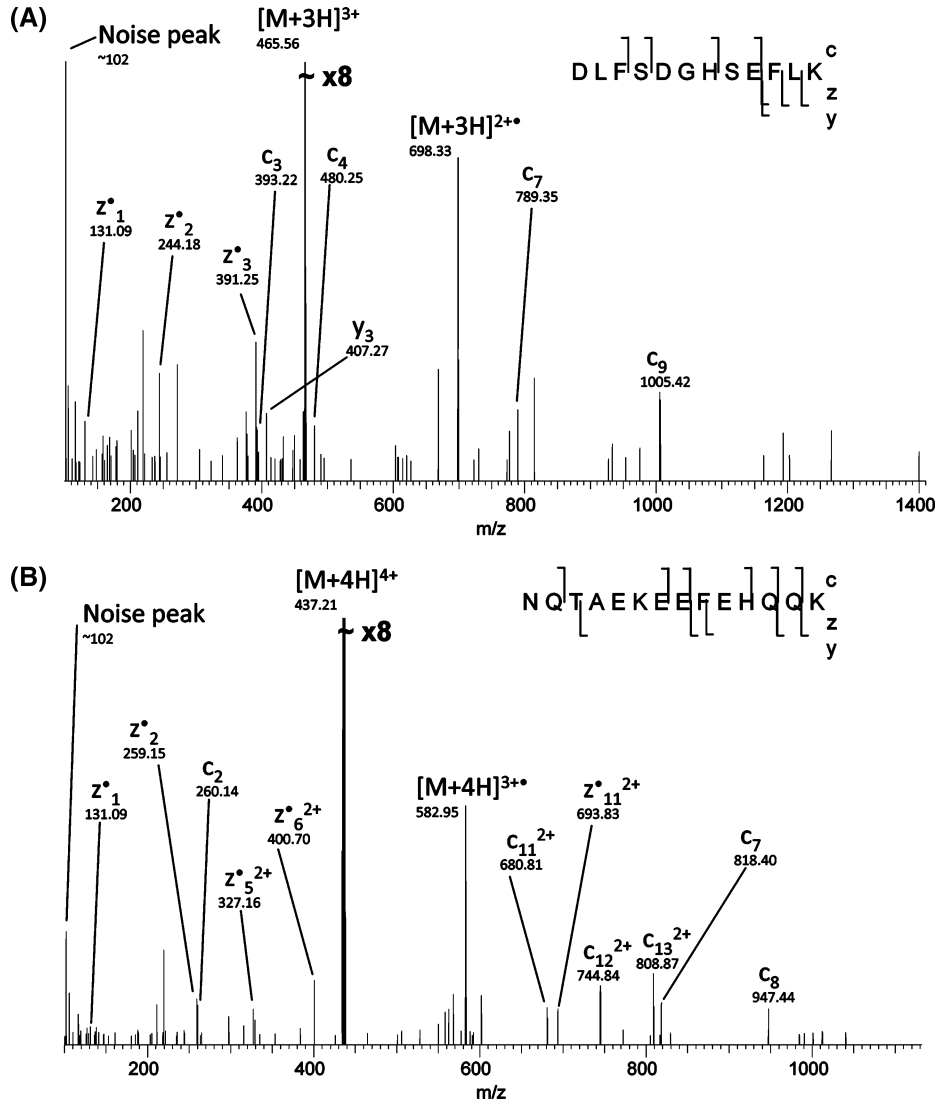

Figure 7. (A) ECD identification with a Mascot peptide score of 0.36. (B) ECD identification with a Mascot peptide score of 1.96. In both cases, all fragments are identified with mass accuracy better than 10 ppm. Mass spectra are shown as acquired, i.e., prior to trimming.

ported.[9] ETD and ECD result in similar fragmentation patterns and it is therefore likely that the described ETD algorithm would be useful for ECD spectral processing. We compared the two data processing methods directly, using the same set of DTAs (generated by the Good algorithm from raw files). The results are shown in Table 5. The ETD algorithm allows removal of the precursor window, neutral losses up to 60 Da from the charge-reduced precursor and the charge-reduced precursor ions themselves. Both processing options result in an increase in identifications compared to the unprocessed DTAs, with the ECD processing algorithm resulting in the most identifications. These searches resulted in fewer identifications than the equivalent searches using DTAs generated by Bioworks. Bioworks automatically corrects the precursor mass to the monoisotopic value even if the second isotopic peak was selected for fragmentation. The DTAs generated using the ETD algorithm may be out by 1 Da if the second isotopic peak was selected for fragmentation. While these identifications are recovered by filtering two windows, one around 0 Da error and the second around 1 Da error, the likelihood of a false-positive is now increased.

Table 5. Comparison with Good[9] Algorithm[a]

presearch	forward hits (2+)	reverse hits	ID rate
ECD search (3341 DTAs); Good Algorithm; Mascot			
DTAs generated using Good algorithm.	1423 (848)	14	42.6
DTAs generated and processed using Good algorithm.	1470 (873)	14	44.0
DTAs generated using Good algorithm; processed using ECD algorithm (as in Table *4*).	1507 (915)	15	45.1

[a] Identifications from doubly-charged precursors are shown in parentheses. To achieve the estimated FDR of 1%, results were filtered according to Mascot scores (scores of 4.09, 0.79 and 8.4, respectively).

Applicability to Phosphoproteomic Data Set. In earlier work, we identified over 900 phosphopeptides from a similar mouse whole-cell lysate sample after TiO_2-based phosphopeptide enrichment.[5] From a total of 6080 ECD DTAs, 1087 phosphopeptides identifications were made (with redundancy). These identifications were made using the same search (OMSSA) and postsearch strategies (filtering by precursor mass error and

score) described here; however, no presearch processing was employed. We reanalyzed this data set, keeping the search and postsearch steps the same, but employing presearch processing of DTAs, as described above. This increased the number of phosphopeptide identifications from 1087 to 1155 (a 6% increase) with no change in the number of reverse identifications. This increase was largely due to an increase in the number of identifications from doubly charged precursor ions: from 285 to 338 (a 19% increase). Employing Mascot for the search step resulted in slightly fewer phosphopeptide identifications.

Discussion

We have considered three stages of data processing: presearch, search and postsearch.

Postsearch. We confirm that searching using a wide precursor mass tolerance window, with subsequent filtering by ppm, substantially improves the identification rate for ECD data as well as CID data. This improvement is due to the tendency for false-positive hits to scatter across a wider mass range of the search space than true hits, which cluster within a window of approximately 10 ppm.[10] We show that postsearch filtering has a greater effect on ECD identification efficiency than CID data, which may be due to the lower scores assigned to ECD spectra by the database search engines employed. We also find that postsearch processing has the largest effect of any individual data processing step (6.4% increase in ECD identifications, compared to 5.3% increase for all preprocessing steps).

Search. Ideally, both precursor *and* fragment mass errors should be specified in ppm, as ppm errors are fairly constant across the measured m/z range. By contrast, an error of 0.02 m/z corresponds to a ppm error of 12.5 for a peak at m/z 1600, but an error of 50 ppm at m/z 400. Observed fragment ion errors range from 0 to 12 ppm. Precursor errors can be converted to ppm for postsearch filtering (as we have shown). However, fragment errors are less easy to convert and filter. Both the Mascot and OMSSA ECD searches employed a fragment tolerance of 0.02 m/z, which is not ideal. Nevertheless, fragment errors greater than 12 ppm are indicative of an incorrect match and are useful information for manual validation of identifications.

We note that the "Mascot decoy" search results in a significantly lower number of identifications than the other searches. This observation draws attention to the conservative nature of the Mascot scoring system, as previously discussed.[15] The scoring scheme was developed prior to the use of decoy searches to estimate FDR. This type of conservative scoring system is particularly valuable when the number of identifications is too low to allow a meaningful estimate of FDR.

ECD fragmentation of peptides results in both z-dot and z-prime fragment ions.[5,8,16] Mascot allows identification of both types. That might be expected to provide an advantage over OMSSA, which cannot identify z-prime fragments. We do observe a slightly higher performance for Mascot for the main test data set studied here; however, OMSSA slightly outperforms Mascot for the phosphopeptide data set (data not shown). Both algorithms perform reasonably well for ECD analysis; however, there is a clear need for a purpose-built ECD/ETD algorithm.

Presearch. The starting premise for all presearch DTA processing is that intense peaks corresponding to anticipated c, z or y fragment ions result in high-scoring identifications; conversely, intense peaks unattributable to anticipated fragment ions detract from an identification score. We removed, or reduced, contributions from three types of uninformative peaks: electrical noise peaks, coeluting precursor peaks, and neutral loss peaks from the charge-reduced precursors. Each of these steps improved both the number of identifications and the average search engine scores of those identifications. Note that neutral loss peaks are not necessarily uninformative;[6,7] however, the search engines we employed are incapable of interpreting these peaks. We remove neutral loss peaks only from ECD mass spectra from doubly charged precursors. This enables us to retain every possible true c/z/y fragment ion, while removing all other peaks within a 140 Da region of the $[M + 2H]^{\bullet}$ reduced precursor. An alternative strategy described by Good et al. removes a smaller region around all reduced precursors, with the side-effect that some true c/z/y fragments are also removed.[9] Our strategy removes fewer true c/z/y fragments and results in a greater number of identifications (Table 5). Retention of true fragments is facilitated by the high resolution ECD MS/MS data: in the region from $[M + 2H]^{\bullet} - 57$ to -140 m/z (73 m/z region), there are 33 potential true fragment masses which are retained (Supplemental Table 1). However, even using a relatively wide retention window of ± 12 ppm, we retain less than 0.8 m/z of the 73 m/z region. Eighteen neutral losses have been described which fall into this 73 m/z region, and which we remove.[6,17]

In a recent comparison of different search engines for identification of ETD mass spectra, Kandasamy et al. found that OMSSA identified far fewer doubly charged peptides than Mascot.[18] We do not find a similarly dramatic difference for ECD mass spectra: identifications from doubly charged peptides make up 55% and 61% of the total for OMSSA and Mascot, respectively. The shortfall observed by Kandasamy et al. may be related to the fact that y ions were not utilized in their OMSSA search. The proportion of identifications from doubly charged peptides approaches the proportion selected for fragmentation (67%).

In conclusion, we show that search results for ECD data are highly dependent on the search strategy employed, varying from an identification rate of 19% (Mascot decoy search) up to 49% (Mascot search with pre- and postsearch processing). We also demonstrate that the absolute database search engine peptide score is unimportant; rather the relative scores of forward and reverse hits are more useful in determining correct identifications.

Acknowledgment. The authors gratefully acknowledge EU Endotrack (S.M.M.S.), EPSRC (A.W.J.), Cancer Research UK (D.L.C., J.K.H.) and the Wellcome Trust (074131) (H.J.C.) for funding.

Supporting Information Available: Supplementary Table 1: Masses retained in the m/z region ($[M + 2H]^{+} - 57$) > m/z > ($[M + 2H]^{+} - 140$). Supplementary Table 2: CID and ECD pairs giving conflicting IDs. Supplementary File: Perl script for removal of non-c,z,y peaks from DTA files, Trim_DTAs_May2009.pl. Supplementary Data: Peptide identifications; Supplementary Tables Identifications.xlsx. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **1998**, *120* (13), 3265–3266.
(2) Cooper, H. J.; Håkansson, K.; Marshall, A. G. The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom. Rev.* **2005**, *24* (2), 201–222.

(3) Sweet, S. M. M.; Cooper, H. J. Electron capture dissociation in the analysis of protein phosphorylation. *Expert Rev. Proteomics* **2007**, *4* (2), 149–159.
(4) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (6), 835–845.
(5) Sweet, S. M. M.; Bailey, C. M.; Cunningham, D. L.; Heath, J. K.; Cooper, H. J. Large-scale localization of protein phosphorylation by use of electron capture dissociation mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (5), 904–912.
(6) Falth, M.; Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Andren, P. E.; Zubarev, R. A. Analytical utility of small neutral losses from reduced species in electron capture dissociation studied using SwedECD database. *Anal. Chem.* **2008**, *80* (21), 8089–8094.
(7) Cooper, H. J.; Hudgins, R. R.; Hakansson, K.; Marshall, A. G. Characterization of amino acid side chain losses in electron capture dissociation. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (3), 241–249.
(8) Savitski, M. M.; Kjeldsen, F.; Nielsen, M. L.; Zubarev, R. A. Hydrogen rearrangement to and from radical z fragments in electron capture dissociation of peptides. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (1), 113–120.
(9) Good, D. M.; Wenger, C. D.; McAlister, G. C.; Bai, D. L.; Hunt, D. F.; Coon, J. J. Post-acquisition ETD spectral processing for increased peptide identifications. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (8), 1435–1440.
(10) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285–1292.
(11) Scherl, A.; Tsai, Y. S.; Shaffer, S. A.; Goodlett, D. R. Increasing information from shotgun proteomic data by accounting for misassigned precursor ion masses. *Proteomics* **2008**, *8* (14), 2791–2797.
(12) Sweet, S. M. M.; Cooper, H. J., On-line liquid chromatography electron capture dissociation for the characterisation of phosphorylation sites in proteins. In *Methods in Molecular Biology*; de Graauw, M., Ed.; Humana Press Inc.: Totowa, NJ: 2009; Vol. 527, pp 191−199.
(13) Sweet, S. M. M.; Creese, A. J.; Cooper, H. J. Strategy for the identification of sites of phosphorylation in proteins: Neutral loss triggered electron capture dissociation. *Anal. Chem.* **2006**, *78* (21), 7563–7569.
(14) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E. P.; Coon, J. J. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal. Chem.* **2007**, *79* (2), 477–485.
(15) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell. Proteomics* **2008**, *7* (5), 962–970.
(16) Tsybin, Y. O.; He, H.; Emmett, M. R.; Hendrickson, C. L.; Marshall, A. G. Ion activation in electron capture dissociation to distinguish between N-terminal and C-terminal product ions. *Anal. Chem.* **2007**, *79* (20), 7596–7602.
(17) Cooper, H. J.; Hakansson, K.; Marshall, A. G.; Hudgins, R. R.; Haselmann, K. F.; Kjeldsen, F.; Budnik, B. A.; Polfer, N. C.; Zubarev, R. A. Letter: The diagnostic value of amino acid side-chain losses in electron capture dissociation of polypeptides. Comment on: "Can the (Mdot-X) region in electron capture dissociation provide reliable information on amino acid composition of polypeptides?", Eur. J. Mass Spectrom. 8, 461 (2002). *Eur. J. of Mass Spectrom.* **2003**, *9* (3), 221–222.
(18) Kandasamy, K.; Pandey, A.; Molina, H. Evaluation of several MS/MS search algorithms for analysis of spectra derived from electron transfer dissociation experiments. *Anal. Chem.* **2009**, *81* (17), 7170–7180.

PR9008282