research articles Journal of proteome research

Improved Peptide Identification for Proteomic Analysis Based on Comprehensive Characterization of Electron Transfer Dissociation Spectra

Rui-Xiang Sun,*,† Meng-Qiu Dong,*,‡ Chun-Qing Song,‡ Hao Chi,† Bing Yang,‡ Li-Yun Xiu,† Li Tao,‡ Zhi-Yi Jing,‡ Chao Liu,† Le-Heng Wang,† Yan Fu,† and Si-Min He†

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and National Institute of Biological Sciences, Beijing 102206, China

Received June 25, 2010

In recent years, electron transfer dissociation (ETD) has enjoyed widespread applications from sequencing of peptides with or without post-translational modifications to top-down analysis of intact proteins. However, peptide identification rates from ETD spectra compare poorly with those from collision induced dissociation (CID) spectra, especially for doubly charged precursors. This is in part due to an insufficient understanding of the characteristics of ETD and consequently a failure of database search engines to make use of the rich information contained in the ETD spectra. In this study, we statistically characterized ETD fragmentation patterns from a collection of 461 440 spectra and subsequently implemented our findings into *pFind*, a database search engine developed earlier for CID data. From ETD spectra of doubly charged precursors, *pFind* 2.1 identified 63−122% more unique peptides than *Mascot* 2.2 under the same 1% false discovery rate. For higher charged peptides as well as phosphopeptides, *pFind* 2.1 also consistently obtained more identifications. Of the features built into *pFind* 2.1, the following two greatly enhanced its performance: (1) refined automatic detection and removal of high-intensity peaks belonging to the precursor, charge-reduced precursor, or related neutral loss species, whose presence often set spectral matching askew; (2) a thorough consideration of hydrogen-rearranged fragment ions such as $z + \mathrm{H}$ and $c - \mathrm{H}$ for peptide precursors of different charge states. Our study has revealed that different charge states of precursors result in different hydrogen rearrangement patterns. For a fragment ion, its propensity of gaining or losing a hydrogen depends on (1) the ion type (*c* or *z*) and (2) the size of the fragment relative to the precursor, and both dependencies are affected by (3) the charge state of the precursor. In addition, we discovered ETD characteristics that are unique for certain types of amino acids (AAs), such as a prominent neutral loss of SCH_2CONH_2 (90.0014 Da) from *z* ions with a carbamidomethylated cysteine at the N-terminus and a neutral loss of histidine side chain $C_4N_2H_5$ (81.0453 Da) from precursor ions containing histidine. The comprehensive list of ETD characteristics summarized in this paper should be valuable for automated database search, *de novo* peptide sequencing, and manual spectral validation.

Keywords: ETD • hydrogen rearrangement • peptide sequencing • search engine

Introduction

Mass spectrometry-based protein identifications rely on fragmentation spectra of peptides or intact proteins using either the bottom-up or the top-down strategy. As the most widely used fragmentation method, collision induced dissociation (CID) has played a central role in shotgun proteomics,[6−10] but in recent years electron capture dissociation (ECD)[11] and electron transfer dissociation (ETD)[12] have attracted much attention.

The popularity of ETD stems from its technical robustness and commercial availability as offered by various types of ion trap mass spectrometers, such as LTQ,[12] LTQ-Orbitrap,[13,14] Agilent 6340,[15,16] and HCT Ultra.[17,18] On these instruments, it has been demonstrated that ETD is complementary to CID and is favored over CID to analyze large peptides, proteins, and labile post-translational modifications (PTMs).[15−23] The past few years have seen a rapid increase of proteomic applications employing ETD, such as analysis of phosphorylation,[24,25] glycosylation,[26,27] quantification,[28,29] or top-down proteomics.[30,31]

In contrast to the rapid advancement of instrumentation and application, the informatics research on ETD has lagged behind, impeding its full potential.[23] Although there are a few algorithms focused on ETD,[53,54] important spectral characteristics of ETD are ignored or underutilized in popular database

* To whom correspondence should be addressed. Rui-Xiang Sun. Tel.: 86-10-62601016. E-mail: rxsun@ict.ac.cn. Meng-Qiu Dong. Tel: 86-10-80706046. E-mail: dongmengqiu@nibs.ac.cn.

† Chinese Academy of Sciences.

‡ National Institute of Biological Sciences.

10.1021/pr100648r

Published on Web 09/30/2010

Table 1. Ten ETD/ECD Data Sets Analyzed in This Paper

no.	data set	#spectra	MS2 resolution	instrument	species	digestion enzyme	precursor charge states	reference
1	WORM-A	58 424[a]	Normal[c]	LTQ-Orbitrap with ETD	*C. elegans*	Trypsin	+2, +3	This manuscript
2	WORM-B	60 585[a]						
3	WORM-C	56 525[a]						
4	WORM-H	22 258[a]	High[d]					
5	YEAST-B1	52 520[a]	Normal[c]		*S. cerevisiae*	Lys-C	+2, +3, +4, +5	ref 36
6	YEAST-B2R1	59 485[a]						
7	YEAST-B2R2	59 007[a]						
8	YEAST-ETcaD	56 019[a]						
9	SwedECD	11 491[b]	High[e]	LTQ-FT with ECD	*Human* and *E. coli*	Trypsin	+2	ref 38
10	PhosphoETD	36 617[a]	Normal[c]	LTQ-Orbitrap with ETD	*C. elegans*	Trypsin	+2, +3	This manuscript

[a] Number of acquired MS2 ETD spectra. [b] Number of identified MS2 ECD spectra. [c] Fragments measured in LTQ. [d] Fragments measured in Orbitrap. [e] Fragments measured in the ICR cell. Data sets WORM-A, -B -and -C were generated under the same conditions with supplemental activation turned on for ETD. The YEAST-B1, -B2R1, -B2R2, and -ETcaD conditions were the same, except that supplemental activation was turned on only in YEAST-ETcaD. Wherever a cell in this table is left blank, it means *the same as above*.

search algorithms, such as Mascot,[5] SEQUEST,[32] OMSSA,[33] or X!TANDEM.[34] For example, high intensity peaks from the unreacted precursor, charge-reduced precursor(s) and their neutral loss species often dominate an ETD spectrum. They can greatly interfere with the matches between theoretical and experimental fragment ions, particularly for doubly charged peptides. Another feature of ETD is the ample presence of hydrogen-rearranged ions, whose masses are typically one Dalton less or more than regular *c*- or *z*-type ions.[35] These hydrogen-rearranged ions increase the complexity of the ETD spectra, especially for doubly protonated precursors, sometime to the extent that they result in misidentification of peptides.[15,16,18,23,36] In sum, none of the database search algorithms currently available for ETD is sufficiently optimized.[23]

Demonstrating the inadequacy of the current database search algorithms for ETD, a recent study evaluating Mascot, OMSSA, X!TANDEM and Spectrum Mill revealed that only 1/6 of 17 000 identified ETD spectra (false discovery rate, FDR is 5%) were assigned with the same sequences by all four algorithms, while for CID it was reported to be nearly a half.[16] Developing a search algorithm thoroughly optimized for ETD has become a crucial issue in the proteomics field in order to encourage further applications of ETD and explore its full potential.[23]

In this study, we focus on improving the peptide identification algorithm by comprehensively characterizing ETD fragmentation patterns of peptides based on multiple proteomics-scale data sets. Previously, preliminary processing of ETD spectra was based on the assumption that fragmentation patterns of ETD are the same as those of ECD. In reality, it is unknown whether the techniques based on characteristics of ECD can be directly applied to ETD data. For instance, hydrogen rearrangement (HR) of fragment ions from doubly charged peptides was characterized using two model peptides[37] and proteomics-scale ECD spectra.[38,39] However, for higher charged precursors, such as +3, +4, or +5, no HR patterns have been reported for either ECD or ETD. In this work, we analyzed 461 440 ETD spectra collected by us and others, and for comparison, 11 491 high resolution ECD spectra from SwedECD[38] (see Table 1 for these data details). This is the largest and most comprehensive data collection employed so far for comparative studies of ETD spectra. From this analysis, we find that the removal of high-intensity precursor-related peaks in ETD greatly benefits peptide identification. In addition, we find that HR patterns of doubly charged peptides are different from those of higher charged peptides, such as +3, +4, and +5. For example, although $c - \mathrm{H}$ ions are typical in ETD spectra derived from doubly charged peptides, they are seldom observed from triply or quadruply charged peptides.

We then implemented our findings into a search engine, pFind,[1-4] which emphasizes the importance of continuous ion series in spectral matching by a kernel function.[1] Adapting the algorithm to distinct HR patterns of +2, +3, and ≥+4 precursors, we designed in pFind 2.1 a strategy to maximize peptide identifications for all charge states of interest. Furthermore, based on the recognition of charge-reduced precursor ions and associated neutral loss peaks characteristic of ETD, a preprocessing step named pRazor was developed to detect and remove these high-intensity interference peaks. Both the preprocessing step and a thorough consideration of HR improved peptide identification substantially. As such, pFind outperformed Mascot and OMSSA by a large margin. For example, pFind identified 63−122% more unique peptides than Mascot for doubly charged precursors at 1% FDR cutoff. For peptides of higher charge states or phosphopeptides, pFind also outperformed Mascot.

In addition, we discovered ETD characteristics that are unique for certain types of amino acids. In Table 2, we summarize our findings and the characteristics reported previously for ETD and ECD. Our results, based on statistical analysis of hundreds of thousands of ETD spectra, should be useful to others in developing algorithms for database search or *de novo* sequencing, and for manual inspection of ETD spectra.

Experimental Procedures

MudPIT Sample Preparation. Peptide mixtures generated from *Caenorhabditis elegans* lysates: The wide type *C. elegans* strain N2 was cultured and maintained as described previously.[40] Adult worms were collected from NGM plates and washed with M9 buffer three times and then with lysis buffer [20 mM HEPES, pH 7.6, 10 mM KCl, 1.5 mM $\mathrm{MgCl_2}$, 1 mM DTT, and 2× EDTA-free proteinase inhibitors cocktail (Roche)] three times. The worm pellet was resuspended with an equal volume of lysis buffer, frozen in liquid nitrogen and stored away at −80 °C until use. A thawed worm sample was homogenized using 3 volumes of 0.5 mm diameter glass beads (e.g., 300 μL of glass beads for 100 μL worm suspension) in a FastPrep-24 homogenizer (MP Biomedicals) at 6.5 m/s, 20 s/pulse, for a total of 4 pulses with 5 min of sample cooling on ice between pulses. The homogenate was centrifuged at 4 °C at 14 000 rpm for 30 min and proteins in the supernatant were precipitated by methanol/chloroform.[41] The protein pellet was solubilized in

Table 2. Characteristics of Peptide ECD and ETD Spectra

items	ECD	reference	ETD	reference
Backbone fragmentation	Main ion types are the radical $z^{\cdot}$ and c with minor $a^{\cdot}$ and y ions.	11, 12	Main ion types are the radical $z^{\cdot}$ and c with minor $a^{\cdot}$ and y ions.	12
Cleavage N-terminal to Proline	Suppressed due to the cyclic side chain of Proline.	11	Suppressed due to the cyclic side chain of Proline.	12
Charge-reduced precursors and related neutral loss species	Prevalent, often of higher intensities than fragment ions.	46, 47	Prevalent, often of higher intensities than fragment ions. ETcaD or SA can convert some of the charge-reduced species into z and c ions.	Detailed analysis not reported
Side chain loss of Histidine-containing peptide precursors	A prominent side chain loss of 81.0453 Da ($C_4H_5N_2$)	49	A prominent side chain loss of 81.0453 Da ($C_4H_5N_2$)	This manuscript
w and u ions due to side chain loss of amino acid Leu, Ile, Glu, Met, or Gln	Observed in Hot-ECD	47	Rarely observed	This manuscript
Side chain loss of fragment ions containing alkylated Cysteine	Side chain loss of 90.0014 Da (•SCH_2CONH_2) or 91.0092 Da (•SCH_2COOH) from z ions containing a Carbamidomethyl Cysteine or Carboxymethyl Cysteine	48	A prominent $z - 90$ fragment instead of the z fragment was observed for a z ion containing a Carbamidomethylated Cysteine at its very N terminus, due to the loss of •SCH_2CONH_2 (90.0014 Da)	This manuscript
Harmonic peaks	Peaks at 1/2, 1/3, ... to even 1/6 of the precursor m/z	50	Not observed	50
HR of c and z ions from +2 peptides	47% occurrence frequency for HR z ions.	37 (the first report of HR based on two model peptides); 38 (Statistics of HR of z ions based on 11 491 ECD spectra.)	(1) SA can increase the yield of z and c ions with the propensity to produce more HR ions; (2) A higher proportion of $c - \mathrm{H}$ in ETD than in ECD (Figures S11b and S11d). Both $z + \mathrm{H}$ and $c - \mathrm{H}$ ions are readily observed; (3) The larger the z ion, the lower the propensity of the z ion to abstract a hydrogen.	35; This manuscript
HR of c and z ions from +3, +4, or +5 peptides	Not reported	N/A	(1) $z + \mathrm{H}$ ions are still abundant although the occurrence frequency decreases when compared with +2 peptides; (2) The $c - \mathrm{H}$ population has disappeared; (3) The larger the z ion, the higher the propensity of the z ion to abstract a hydrogen. This is opposite to the HR pattern of z ions from +2 peptides.	58; This manuscript

100 mM Tris pH 8.5 containing 8 M urea, reduced with 5 mM TCEP for 20 min, and alkylated with 10 mM iodoacetamide for 30 min. Then the sample was diluted with 3 volumes of 100 mM Tris, pH 8.5, supplemented with $CaCl_2$ to 1 mM and methylamine to 20 mM, and digested with trypsin at 1:50 (enzyme: substrate) ratio at 37 °C for 16 h. After digestion, peptides were acidified with formic acid to a final concentration of 5%.

C. elegans phosphopeptides enriched by IMAC: Phosphopeptides were prepared based on a method described before.[55]

Table 3. Comparison of pFind and Mascot on WORM ETD Spectra

data set	#ETD spectra	identifications	Mascot (charge: +2)	pFind (charge: +2)	Mascot∩pFind[b] (+2)	Mascot (charge: +3)	pFind (charge: +3)	Mascot∩pFind[b] (+3)
WORM-A	58 424	#spectra	3285	9293 (182.9[a])	3111 (94.7)	1726	1869 (8.3[a])	1601 (92.8)
		#peptides	1073	2380 (121.8)	1004 (93.6)	604	644 (6.6)	573 (94.9)
		#proteins	842	1444 (71.5)	801(95.0)	554	606 (9.4)	517 (93.3)
WORM-B	60 585	#spectra	4987	10,608 (112.7)	4707 (94.4)	2565	2805 (9.4)	2418 (94.3)
		#peptides	1750	2853 (63.0)	1616 (92.3)	672	742 (10.4)	646 (96.1)
		#proteins	1092	1595 (46.1)	1009 (92.4)	630	668 (6.0)	590 (93.7)
WORM-C	56 525	#spectra	3173	7693 (142.5)	2984 (94.0)	1779	1954 (9.8)	1663 (93.5)
		#peptides	1364	2544 (86.5)	1266 (92.8)	511	556 (8.8)	477 (93.3)
		#proteins	971	1415 (45.7)	903 (93.0)	456	499 (9.4)	430 (94.3)

[a] Percentage of improvement of pFind over Mascot. [b] Number of overlapping results between Mascot and pFind. The overlap as a percentage of the Mascot result is shown in parentheses.

Table 4. Comparison of pFind and Mascot on WORM CID Data

data set	#CID spectra	identifications	Mascot (charge: +2)	pFind (charge: +2)	Mascot∩pFind[b] (+2)	Mascot (charge: +3)	pFind (charge: +3)	Mascot∩pFind[b] (+3)
WORM-A	58 424	#spectra	16 911	19,241 (13.8[a])	15,829 (93.6)	4975	5289 (6.3[a])	4808 (96.6)
		#peptides	3270	3434 (5.0)	3033 (92.8)	1043	1077 (3.3)	980 (94.0)
		#proteins	1641	1792 (9.2)	1522 (92.7)	834	879 (5.4)	759 (91.0)
WORM-B	60 585	#spectra	17 977	20,075 (11.7)	16,901 (94.0)	6277	6616 (5.4)	6106 (97.3)
		#peptides	3808	4042 (6.1)	3599 (94.5)	1222	1267 (3.7)	1177 (96.3)
		#proteins	1714	1860 (8.5)	1587 (92.6)	978	1017 (4.0)	910 (93.0)
WORM-C	56 525	#spectra	15 117	17,249 (14.1)	14,113 (93.4)	5484	5800 (5.8)	5268 (96.1)
		#peptides	3525	3800 (7.8)	3333 (94.6)	1000	1030 (3.0)	949 (94.9)
		#proteins	1664	1841 (10.6)	1552 (93.3)	802	824 (2.7)	737 (91.9)

[a] Percentage of improvement of pFind over Mascot. [b] Number of overlapping results between Mascot and pFind. The overlap as a percentage of the Mascot result is shown in parentheses.

C. elegans tryptic peptides were generated from lysates as described above. This peptide mixture was desalted by reverse-phase HPLC (Eclipse XDB-C18 column from Agilent) and separated using a strong cation exchange column (PolySULFOETHYL A from PolyLC Inc.) into six fractions. Each fraction was desalted using 100 mg of tC18 SepPak solid-phase extraction cartridges (Waters). Phosphopeptides from each fraction were enriched separately using a Gallium-IMAC spin column (Pierce) according to the manufacturer's instructions.

MudPIT Analysis. For *C. elegans* peptides without IMAC enrichment: MudPIT analysis of digested *C. elegans* proteins (80 μg each time) was performed in triplicates on an LTQ-Orbitrap with ETD mass spectrometer (Thermo Scientific) equipped with an Agilent 1200 quaternary pump. MudPIT conditions were adapted from what had been described before[42] with the following modifications. A 250 μm (ID) × 2 cm (length) desalting column was packed with 5 μm, 125 Å Aqua C18 resin (Phenomenex). The analytical reverse phase column was 100 μm (ID) × 9 cm (length) with a pulled tip, packed with 3 μm, 125 Å Aqua C18 resin (Phenomenex). Between the desalting column and the analytical column is a strong cation exchange column (SCX), 250 μm (ID) by 2 cm (length), containing 5 μm, 120 Å Partisphere SCX material (Whatman). The salt pulses of these 9-step MudPIT experiments were set at 0, 10, 20, 30, 40, 60, 80, 100, and 100, expressed as the percentage of buffer C. MS2 spectra were acquired in data-dependent mode. Full scans (300−2000 m/z) were performed in the Orbitrap (R = 100 000), and each full scan was followed by 5 sets of CID-ETD double play MS2 in LTQ for top 5 peaks. Dynamic exclusion was set to repeat count of 2, repeat duration of 30 s, exclusion list of 300, and exclusion duration of 30 s. Minimal signal threshold for MS2 was 5000. The AGC targets were 2e5 for FTMS full scan, 2e4 for LTQ MS2, and 1e5 for reagent ion. For ETD, supplemental activation (SA) was turned on and the reaction time was 100 ms.

For *C. elegans* phosphopeptides enriched by IMAC: After Enrichment, phosphopeptides from each of the six SCX fractions were pooled together and one-third of it was analyzed using the same MudPIT method as described above except that the very last step was omitted (i.e., the salt pulses were 0, 10, 20, 30, 40, 60, 80, and 100% of buffer C).

Data Analysis. All CID and ETD tandem mass spectra acquired from the whole-cell worm lysate were extracted from Xcalibur 2.0.7 .RAW files. The peak list .ms2 files, including CID and ETD spectra were generated using RawXtractor 1.9.7 (Developed by the research group of Dr. John Yates, III, http://fields.scripps.edu/). Then CID and ETD spectra were separated by an in-house software SeparateCIDETD. 175,534 CID-ETD-pair were derived by triplicate analysis (Data sets WORM-A, WORM-B and WORM-C in Table 1). Tables 3 and 4 list the number of spectra for each analysis and their identifications by pFind 2.1 and Mascot 2.2 at FDR 1%.

These CID and ETD spectra were searched in parallel against the concatenated forward and reversed worm database, wormpep201 (March 26, 2009, 23 993 proteins, downloaded from http://www.sanger.ac.uk/Projects/C_elegans/). We searched all spectra by both pFind 2.1 and Mascot 2.2 and then integrated their results together to get a highly confident ETD data set. The precursor mass tolerance was set to ±1.1 Da. The fragment mass tolerance was ±0.5 Da for normal resolution spectra and ±0.02 Da for high resolution spectra. Two missed cleavage sites were allowed for the enzyme trypsin. All cysteines were considered carbamidomethylated (fixed modification). For phosphopeptide data, phosphorylation on S, T or Y was additionally set to variable modifications. The filtering of the search results was performed by pBuild, an in-house software

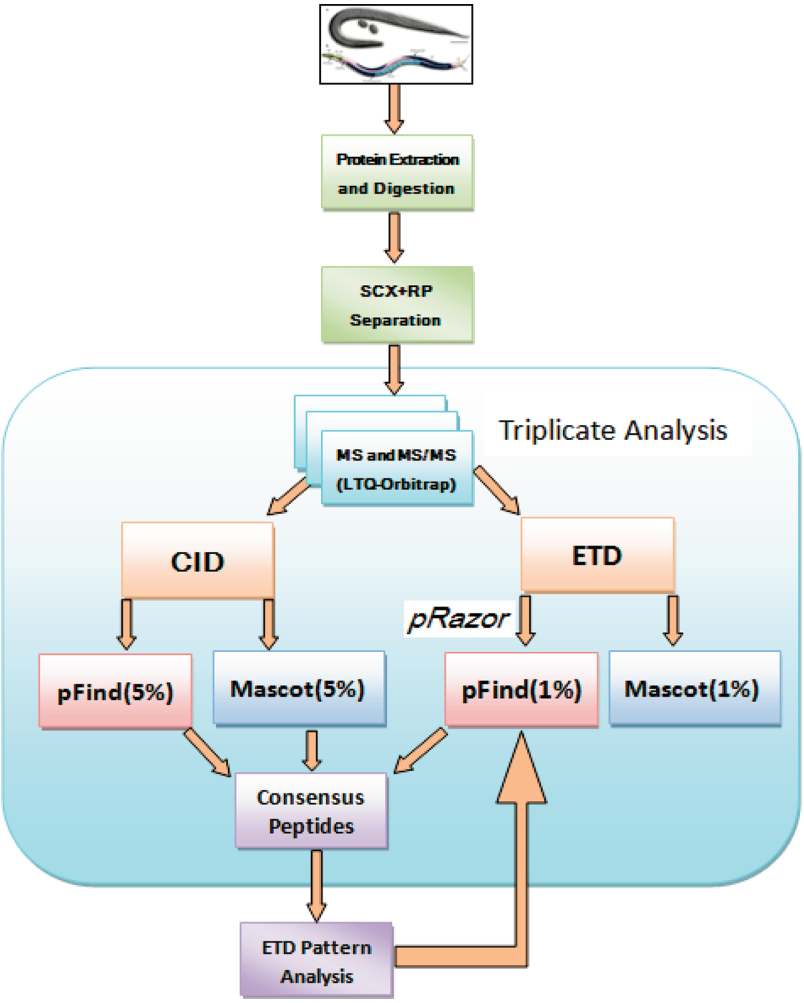


Figure 1. Workflow for worm CID and ETD spectra acquisition and analysis. A trypsin digested *Caenorhabditis elegans* lysate was analyzed by LC−MS/MS on an LTQ-Orbitrap XL with ETD. Three replicate experiments generated 58 424, 60 585, and 56 525 ETD spectra, each accompanied by a cognate CID. CID spectra were analyzed by pFind 2.1 and Mascot 2.2 (FDR = 5%). ETD spectra were searched using pFind 2.1 (after preprocessed using pRazor) and Mascot (FDR = 1%). The ETD results were further filtered by requiring that, for each ETD spectrum, its cognate CID spectrum must be assigned the same sequence by both Mascot and pFind.

for analyzing and combining the results of search engines, such as pFind, Mascot or SEQUEST (http://pfind.ict.ac.cn). Any peptide containing less than 7 amino acids or with a mass of less than 700 Da was excluded from further analysis. FDR was computed according to the target-decoy strategy.[43−45] Considering isotopic peaks, mass deviation within the interval of [−4 ppm, 1.02 Da] was allowed under the desired FDR cutoff, 1% or 5%.

We compared pFind and Mascot on CID and ETD under the same level of FDR, such as 1%, at the spectrum and peptide levels. Our comparison was also based on the receiver operating characteristic (ROC) curves, which can quantitatively evaluate the performance of these two search engines.

In addition to the normal resolution worm ETD data stated above, we also generated 22 258 high resolution MS2 ETD spectra (Data set WORM-H in Table 1), that is, fragment m/z was measured in Orbitrap (resolution 7500 at $m/z = 400$). The high accuracy afforded by Orbitrap can help probe ETD's behaviors more precisely, such as the mass deviations of hydrogen-rearranged ions. For comparative analysis of the characteristics of high resolution ETD spectra, we also analyzed 11 491 ECD spectra from the database of SwedECD.[38]

Results

Optimizing Instrument Parameters of LTQ-Orbitrap to Improve Data Collection. CID-ETD double play data were generated on LTQ-Orbitrap XL with ETD (Thermo Scientific). A digested *C. elegans* lysate sample was analyzed three times using a nine-step MudPIT method. To make sure the data are of high quality, we optimized experimental parameters such as automatic gain control (AGC) targets for LTQ MSn, FT MSn and reagent ion using a standard sample-a mixture of eight proteins digested with trypsin (details in Supplementary Text and Supplementary Table S1, Supporting Information). We chose the best setting obtained from the test experiments (Set2, AGC targets: 2e4 for LTQ MS2, 2e5 for FT MS1, and 1e5 for reagent ions) to perform MudPIT analysis of worm lysates by CID and ETD. Additional MudPIT experiments were carried out to collect high resolution CID and ETD MS2 spectra using the same worm lysate sample.

Obtaining a High-Quality ETD Data Set to Characterize Peptide Fragmentation. In order to select peptide ETD spectra with highly confident sequence identifications from the triplicate analyses of the *C. elegans* lysate sample, we require that each ETD spectrum in this collection be assigned a peptide sequence by pFind 2.1 at 1% FDR using the reversed database as a decoy. In addition, for each ETD spectrum in this data set, its cognate CID spectrum must be assigned the same sequence by both Mascot (FDR = 5%) and pFind (FDR = 5%) as shown in Figure 1. Mascot was also used to analyze the ETD spectra as it was previously reported to be one of the best search engines for ETD.[16] However, for our ETD data the number of spectra identified by Mascot was far less than that by pFind (details in Table 3). So, we did not use the results from the Mascot ETD search. Out of a total of 175 534 spectra from the triplicate analyses (data sets WORM-A, WORM-B, and WORM-C in Table 1), we obtained 23 649 high-confidence ETD spectra (mass accuracy with ±6 mDa and *E*-values <0.001, see Supplementary Figure S1, Supporting Information) for doubly charged peptides. If a peptide had more than one copy of spectrum, the one with the best score was kept to represent the sequence. As such, we obtained 3471 high-quality ETD spectra corresponding to unique peptides for further study.

Our ETD spectra were generated with supplemental activation (electron transfer and collisionally activated dissociation, or ETcaD) from a *C. elegans* lysate. To guard against possible sample- or lab-originated bias, we also analyzed four additional sets of published ETD data of the yeast proteome from the Coon lab,[36] one of which was also generated with ETcaD and the other three were collected without supplemental activation (data sets Nos. 5−8 in Table 1). For comparison, the SwedECD data containing 11 491 high resolution ECD spectra were examined in parallel.[38] We also generated a set of high resolution ETD spectra on LTQ-Orbitrap to confirm some of the features detected in normal resolution ETD spectra. Lastly, we produced a phosphopeptide data set with sequential acquisition of CID and ETD spectra for each precursor in LTQ. In total, we analyzed 10 proteome-scale data sets, details of which are shown in Table 1.

Statistical Characterization of ETD Peptide Fragmentation. High-intensity precursors, charge-reduced (CR) precursors, and their neutral-loss peaks dominate peptide ETD spectra. We analyzed the peaks near the precursors and CR precursors and found that precursor neutral loss is rare (Figure 2a), whereas CR precursors have an abundance of neutral loss peaks spreading to as far as −35 Da relative to the CR precursor masses. Most of the high-intensity neutral loss peaks are concentrated at −17 ± 3 Da and a minor fraction at −28 Da (Figure 2b). The neutral loss peaks may be accounted for by loss of NH_3, H_2O, or CO with or without hydrogen rearrangement.[46,47]

For histidine-containing peptides, we noticed a prominent peak about 81 Da lower than the charge-reduced precursor (An example is shown in Figure S2, Supporting Information). To further probe the phenomenon, we divided the 3471 unique peptides from WORM-A, -B, and -C ETD data sets into two classes, one containing histidine, and the other not. We then plotted the relative intensities of the peaks near the 81 Da neutral loss region of CR precursors (Figure 3a). The result clearly shows that this neutral loss is specific for peptides containing histidine. Moreover, the intensity of this neutral loss is higher for peptides containing two or more histidines than those with only one (Figure 3b). This neutral loss is probably due to a loss of $C_4N_2H_5$ (81.0453 Da) from the histidine side chain in conjunction with the gain or loss of a hydrogen. The high-resolution high-mass accuracy SwedECD data confirmed our result (Figure 3c). This highly specific side chain loss of histidine-containing peptides can be used to validate sequence identifications.

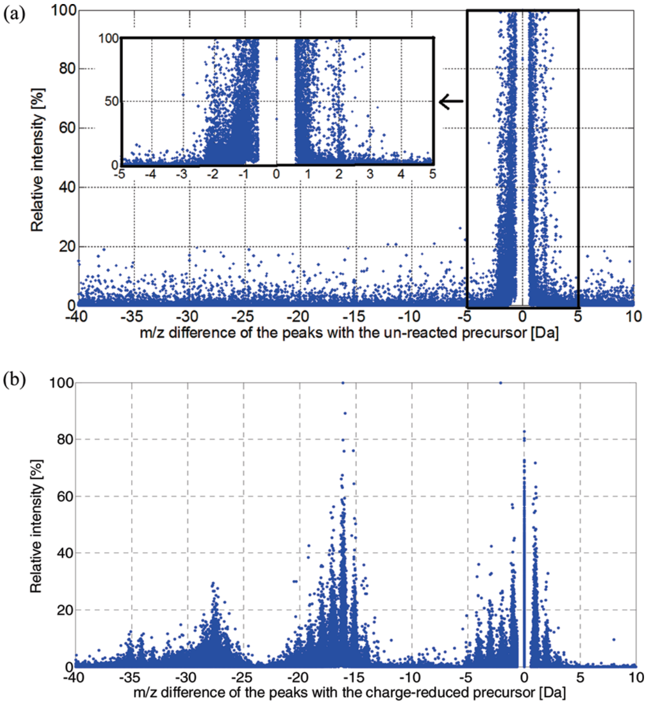


Figure 2. Characteristic distribution of peaks around unreacted precursors and CR precursors in ETD spectra. (a) Neutral loss peaks are seldom observed around the unreacted precursors. (b) Neutral loss peaks associated the CR precursors display a characteristic distribution. Each point corresponds to a peak in one of the 3471 high-quality ETD spectra (see Results).

Hydrogen rearrangement (HR) or hydrogen transfer is a common phenomenon in ECD.[37,38] Unlike ECD, ETD is often performed with supplemental activation, which is a very low energy CID aimed at CR precursors in order to dissociate fragment ions held together by noncovalent interactions. It was pointed out that SA could increase the proportion of hydrogen-rearranged ions.[35] Hydrogen-rearranged *z* ions usually have a mass addition of about one Dalton to regular *z* ions, whereas masses of hydrogen-rearranged *c* ions are usually one Dalton less than regular *c* ions (Supplementary Figure S3, Supporting Information). However, systematic analysis of HR based on large-scale ETD data has not been reported. Here we examined the HR patterns of +2, +3, +4, and +5 ETD spectra separately and found that fragment ions of +2 peptides exhibit a different HR pattern from those of +3, +4 or +5. For example, analysis of both WORM-ETcaD and YEAST-ETcaD data revealed that in +2 ETD spectra HR manifests itself through *z* ion gaining a hydrogen and *c* ion losing a hydrogen. As shown in Figure 4a, $z + 1$ are more abundant than *z* ions whereas $z - 1$ are close to nonexistent. This is accompanied by a large number of *c* and $c - 1$ ions, but not $c + 1$ ions (Figure 4b). Although in our normal resolution ETD data, *z* ions gaining a hydrogen cannot be distinguished from isotopic peaks containing one ^{13}C or ^{15}N atom, we can estimate the interference of isotopic peaks by examining *b* and *y* ions in CID spectra because the isotope composition of *b* and *y* ions should be nearly identical to that of matching *c* and *z* ions. Our analysis indicated that the interference from isotopic peaks is less than 5% (Figure 4c and d). In sharp contrast, triply charged peptides exhibit a substantial decrease in the occurrence of $z + 1$ ions in the ETD spectra whereas $c - 1$ ions diminish almost completely (Figure 4e and f). A recent publication by Chalkley *et al.* comparing

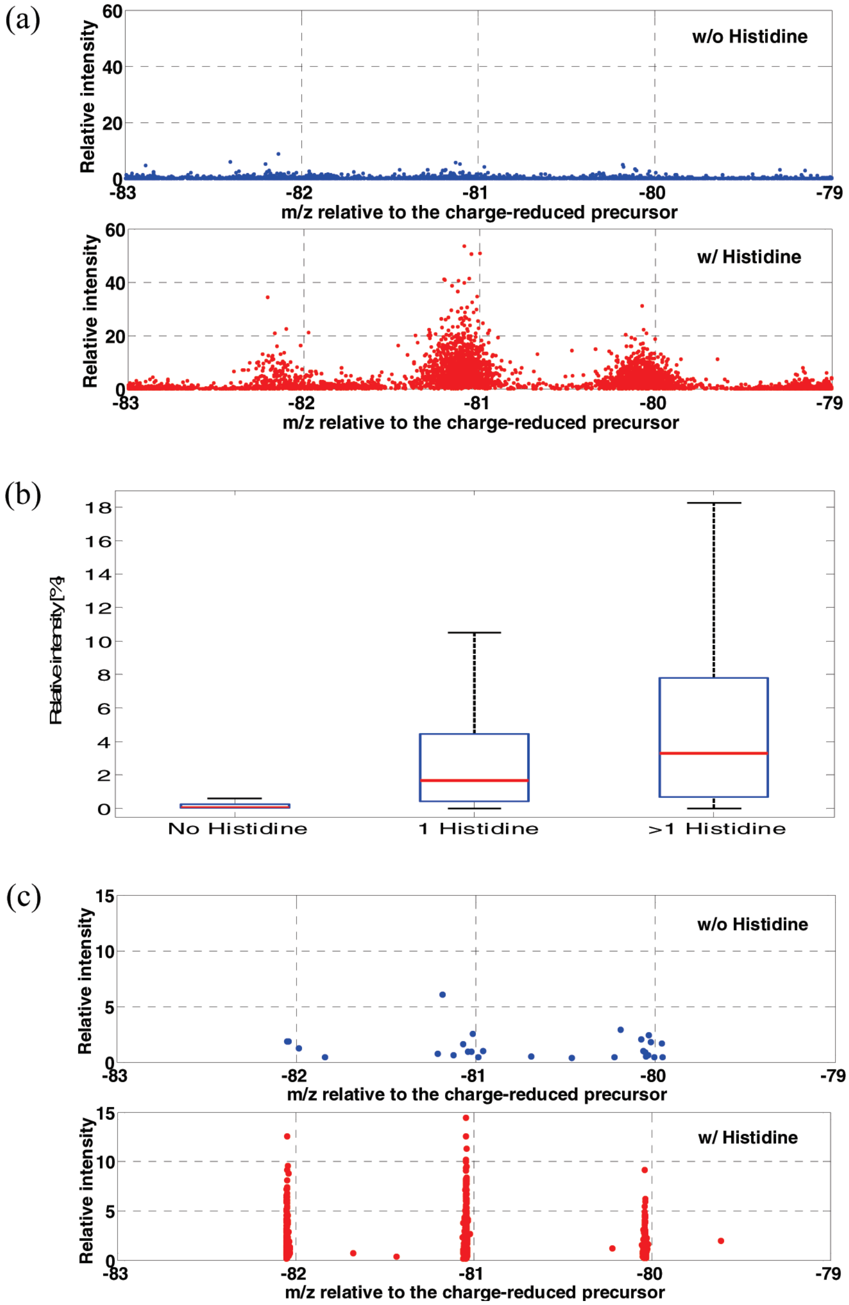
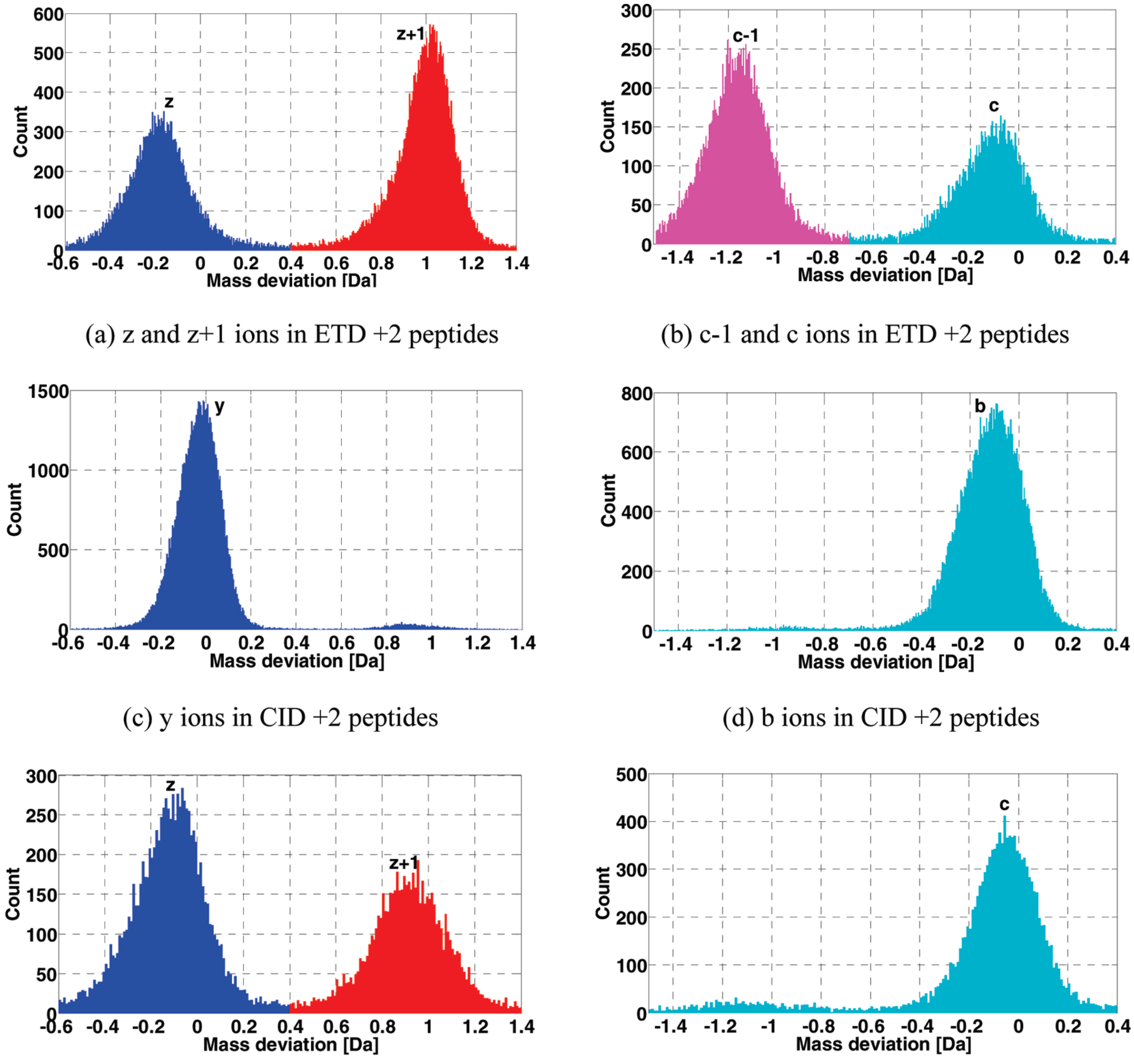


Figure 3. Peptides containing histidine residues have a prominent side chain losses of ~81 Da from the charge-reduced precursors. (a) Relative intensities of peaks found in the region −83 to −79 Da away from the CR precursors for peptides with or without histidine. The results are based on the WORM-A, -B, and -C data sets. (b) Box plot of the relative intensities of peaks around −81 Da in (a). (c) Analysis of the SwedECD data shows in high resolution the same histidine side chain loss from the CR precursors as observed in (a).

only +2 and +3 ETD spectra reported a similar observation.[58,59] Lys-C peptides from the YEAST-ETcaD data display a similar HR pattern as trypsin peptides from the WORM data (Figure S4, Supporting Information), in despite of a higher proportion of +3, +4 and +5 peptides. In Figure S4 (Supporting Information), the percentages of $z + 1$ ions over all z ions in +2, +3, +4 and +5 ETD spectra are 58.5, 22.2, 48.0, and 51.6%, respectively. Only in +2 ETD spectra did we observe a significant population of $c - 1$ ions, which accounted for 45.8% of all c ions. Figure S5 (Supporting Information) showed this analysis of y and b ions of +2, +3, +4, and +5 CID spectra.

In addition to analyzing HR of z ions as a whole, we also examined the relationship between the propensity of HR and the z ion size. We normalized the length of each z ion against the length of its precursor peptide. The normalized length (NL) of a z ion is its amino acid length divided by that of its precursor peptide. We found that in +2 ETD spectra, little HR is observed for z ions when their NL values are close to 1 (Figure 5a and b). In contrast, in +3 and +4 ETD spectra, singly charged larger z ions (NL > 0.8) have a higher propensity of abstracting a hydrogen than smaller ones (NL below 0.4) (Figure 5c and d). This is the first report that HR of a z-ion is differentially affected by its relative size and the charge state of its precursor peptide.

Neutral loss off fragment ions is common in CID spectra, mostly in the form of an ammonia or water loss off b or y



(a) z and z+1 ions in ETD +2 peptides

(b) c-1 and c ions in ETD +2 peptides

(c) y ions in CID +2 peptides

(d) b ions in CID +2 peptides

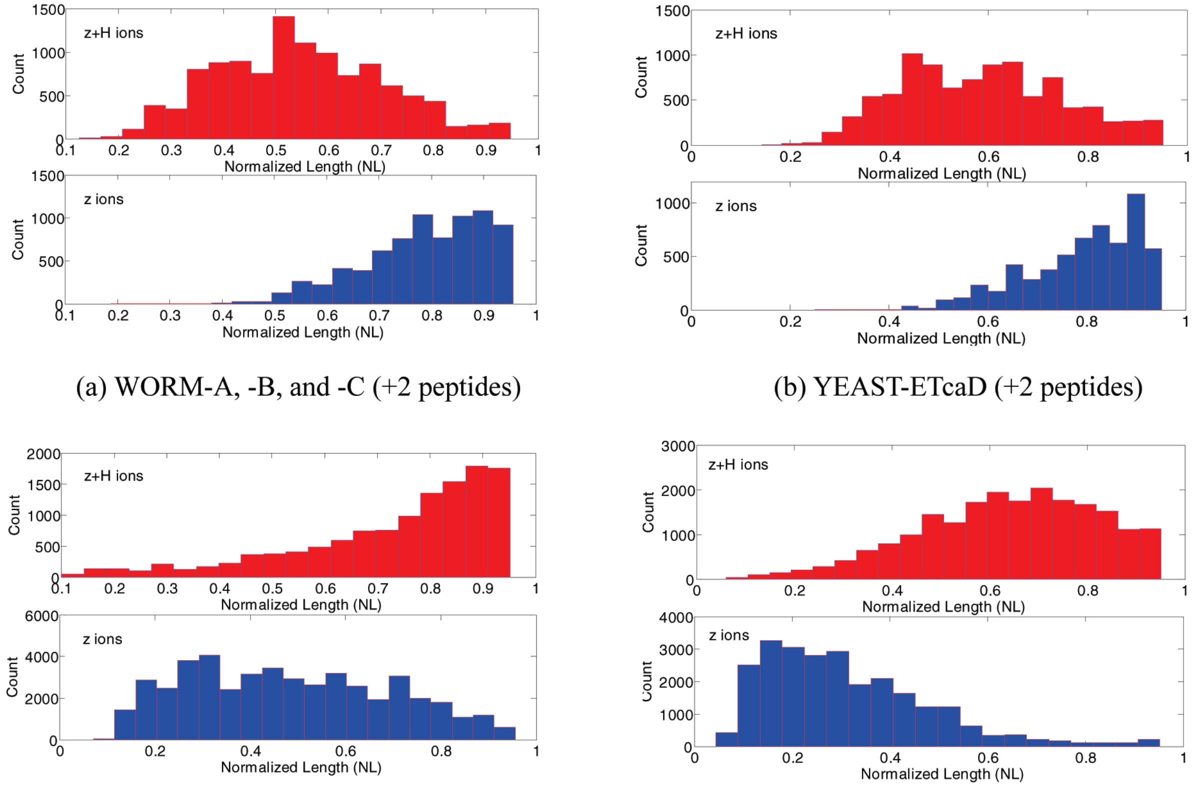
(e) z and z+1 ions in ETD +3 peptides

(f) c ions in ETD +3 peptides

Figure 4. Doubly charged and triply charged peptide ETD spectra show different HR patterns. In +2 spectra, HR ions appear at a higher frequency than regular *c* and *z* ions such that in the histograms there are more *z* + 1 ions than *z* ion (a) and more *c* − 1 ions than *c* (b). The abundance of *z* + 1 ions cannot be explained by isotopic peaks only, because in the matching CID spectra *y* + 1 population is insignificant compared to the *y* ion population (c). Likewise, few *b* − 1 ions are found in the +2 CID spectra (d). HR is less pronounced in +3 ETD spectra—*z* + 1 ions are outnumbered by *z* ions (e) while *c* − 1 ions are hard to observe (f).

ions.[6−8] In contrast, ECD or ETD spectra exhibit no obvious ammonia or water loss off *c* or *z* ions, although ammonia or water loss of CR precursor are clearly seen (see above and Figure 2). However, we did find a very specific neutral loss of ∼90 Da from *z* fragments starting with a carbamidomethylated cysteine; for example, in Figure S6a (Supporting Information), z7 was missing while z4, z5, z6, and z8 were readily observed for the peptide THCFEWTAK. This started with our observation that a *z* ion starting with a carbamidomethylated cysteine is always missing in an otherwise contiguous ion series. Closer inspection found that the missing peak was almost always accompanied by the appearance of another peak ∼90 Da smaller than the expected *z* ion. This corresponds to the loss of -SCH_2CONH_2 from the side chain of carbamidomethylated cysteine when it is the most N-terminal residue of a *z* ion. A previous study on ECD spectra of standard proteins containing carboxymethylated or carbamidomethylated cysteine confirms this interpretation.[48] This should provide a useful clue to *de novo* sequencing of cysteine-containing peptides using ETD.

Collectively, our statistical analysis of multiple large-scale ETD data sets revealed that ETD spectra are information-rich. Many features of peptide ETD spectra are different from those of CID. For example, HR is common in ECD and ETD. Abstraction of a hydrogen from *c* ions occurs frequently in +2 ETD spectra, but not in +3 or +4 spectra. We also find that HR of a *z* ion is differentially affected by its relative size and the charge state of its precursor peptide. Furthermore, neutral loss of CR precursors is prevalent while that of fragment ions is rare. Lastly, we have observed a highly specific neutral loss of ∼81 Da (theoretical mass 81.0453 Da) from histidine-containing peptides. This can be utilized to validate sequence identifications. A detailed summary of our statistical results of peptide ETD spectra is listed in Table 2, along with the characteristics of ETD or ECD spectra of peptides learnt from



(a) WORM-A, -B, and -C (+2 peptides)

(b) YEAST-ETcaD (+2 peptides)

(c) YEAST-ETcaD (+3 peptides)

(d) YEAST-ETcaD (+4 peptides)

Figure 5. Larger *z* ions are less likely to undergo HR in +2 spectra whereas in +3 and +4 spectra they have a high propensity of gaining a hydrogen. *z* and *z* + 1 ions are counted separately according to their normalized length (NL), which is the AA length of a *z* or *z* + 1 ion normalized against the length of its precursor peptide. (a) for +2 peptides in data sets WORM-A, -B, and -C; (b) for +2 peptides in the data set YEAST-ETcaD; (c) for +3 peptides in YEAST-ETcaD; (d) for +4 peptides in YEAST-ETcaD.

previously studies.[46−50] In this paper, we have employed many of these characteristics in our search engine pFind 2.1 to improve peptide identification from ETD data.

Taking into Consideration Characteristics of ETD Spectra Greatly Enhances the Performance of pFind. We implemented our findings about peptide ETD fragmentation patterns into pFind, a database search engine we developed six years ago.[1] The result is striking. The new version, pFind 2.1, identified 63−122% more unique peptides than Mascot 2.2[5] from three ETD data sets (121.8, 63.0, and 86.5% in Table 3). For phosphopeptide ETD spectra, pFind identified 74.4% more phosphopeptides than Mascot (351−612 in Table 6). Because most of the increase comes from doubly charged peptides, and 60−75% tryptic peptides are doubly charged, pFind 2.1 would make a big difference for ETD data analysis of tryptic samples. Trypsin is the most commonly used protease in mass spectrometric sample preparation owing to its high specificity and low cost. We believe that pFind 2.1 would be helpful for many proteomic studies. Of the features built into pFind 2.1, data preprocessing and a thorough consideration of HR fragments are the most important. We discuss the details of these two features in the following sections.

pRazor, a Highly Effective Preprocessing Step for ETD Data Analysis. As described above, ETD spectra are typically dominated by intense peaks of leftover precursors, CR precursors and related neutral loss species (Figure 2). A representative spectrum is shown in Figure S7. On one hand, such peaks can be used to determine the charge states of peptide precursors;[51,52] on the other, they increase the chance of random match with wrong sequences. To eliminate this undesirable effect, we designed an efficient preprocessing algorithm, pRazor, to find such peaks and remove them. For +2 peptides, before the spectra are searched against a protein database, pRazor removes the peaks within a ± 3 Da region surrounding the precursor and from another region (−20 Da to +5 Da) flanking the +1 CR precursor. For peptides of higher charge states such as +3 or +4, pRazor removes all CR precursors (+2 or +2 and +3, and so on) and their neutral loss peaks using the same mass windows. Compared to previous algorithms,[56,57] pRazor uses a narrower window in order to prevent or reduce accidental removal of fragment ions.

To verify the effectiveness of the pRazor algorithm, we tested it with the WORM-A data set which contains 58 424 spectra (Table 1). As shown in Figure 6, data preprocessing using pRazor dramatically increases the number of pFind identifications on the spectral and peptide levels (FDR fixed at 1%). This improvement is well pronounced regardless of how HR fragment ions are handled by the algorithm (to be discussed later). For example, pRazor boosts peptide identifications from a mere 55 to 971 when HR is ignored. Similarly, when *c* and *z* ion HRs are both taken into consideration, pRazor again increases peptide identifications by a thousand, from 1376 to 2380 (Figure 6b). The ROC curves comparing the data analysis results with or without pRazor are shown in Figure S8 (Supporting Information).

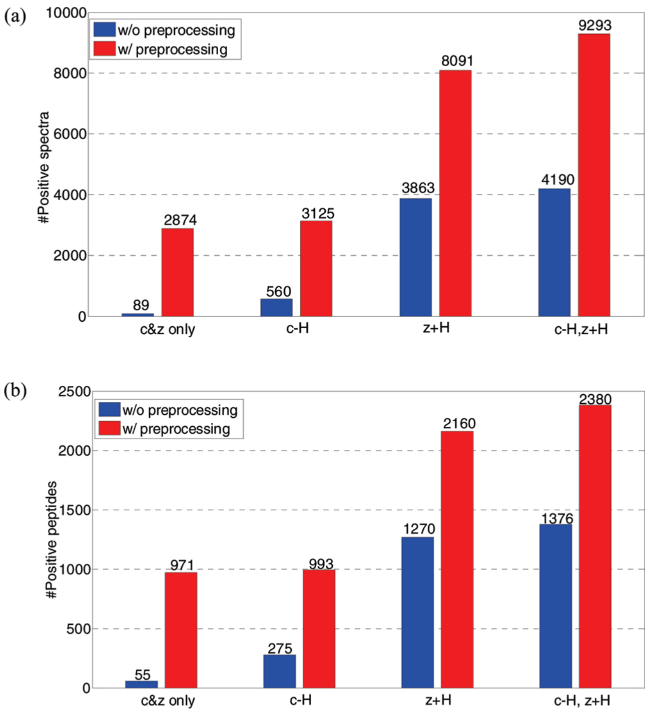


Figure 6. Both preprocessing and consideration of HR ions greatly enhance the performance of pFind 2.1 on ETD data analysis. (a) Improvement of spectral identifications. (b) Improvement of peptide identifications. ETD data were preprocessed (w/preprocessing) or not (w/o preprocessing) using pRazor to remove high-intensity unreacted precursors, CR precursors and related neutral loss species. "*c*&*z* only" means that only regular *c* and *z* ions were considered in spectral matching. "$c - H$", "$z + H$" and "$c - H$, $z + H$" indicate that in addition to regular *c* and *z* ions, $c - H$, $z + H$, and both $c - H$ and $z + H$ ions were considered in spectral matching, respectively.

To further test the effectiveness of pRazor, we submitted ETD spectra preprocessed by pRazor to Mascot and compared the identification results to those from unpreprocessed ones. Similar to the effect on pFind, pRazor increased Mascot identifications by 32% on the spectral level and 28% on the peptide level for doubly charge peptides (Figure S9a, Supporting Information). Mascot scores for most ETD spectra were also improved after they were preprocessed by pRazor (Figure S9b, Supporting Information).

Utilization of HR Patterns Greatly Enhances pFind 2.1. Fragment ions that have undergone HR possess distorted isotopic peak clusters, making it difficult to determine the monoisotopic masses of the fragments.[37] This may mislead database search engines into assigning a wrong or low-confidence sequence to the ETD spectrum. Currently, most database search algorithms do not, or not thoroughly, consider HR in ETD or ECD data analysis. This might compromise the confidence level of search results.[15,16,18,23,36] In this study we show that indeed a full consideration of HR in a precursor charge state-specific manner is crucial for successful peptide identification.

As described above, in +2 ETD spectra, there are many $c - H$ ions and $z + H$ ions besides regular *c* and *z* ions, and no $z - H$ or $c + H$ ions (Figure 4a and b). In contrast, in +3, +4, and +5 ETD spectra most of the fragment ions are regular *c* and *z* type ones, although $z + H$ ions are still visible (Figure S4, Supporting Information). Consistent with these statistical results, we found that for +2 peptides inclusion of both $z + H$ and $c - H$ peaks in the pFind scoring algorithm gave the best identification result (Figure 6, Figure S8 and Figure S10a, Supporting Information). Specifically, twice as many peptides were identified if $z + H$ ions were considered (from 993 to 2160, comparing "$z + H$" to "$c - H$" with preprocessing in Figure 6b). On top of this, adding $c - H$ ions ("$c - H$, $z + H$") increased identification by another 15% for spectra and 10% for peptides (comparing "$c - H$, $z + H$" and "$z + H$" with preprocessing in Figures 6a and 6b). In contrast, for peptides carrying three or more positive charges, the best performance was achieved with only $z + H$ ions added to regular *c* and *z* ions (Figure S10b−d, Supporting Information). Similar results were obtained from analysis of high resolution ETD or ECD data. We also tested other features such as *y* ions, but their effect on peptide identification was not significant due to their low intensity in ETD spectra.

pFind 2.1 Outperforms Mascot and OMSSA in ETD Data Analysis by a Large Margin. We systematically compared the performance of pFind with Mascot and OMSSA on ETD and CID data. pFind 2.1 outperformed Mascot in both ETD and CID data analysis (Table 3, Table 4, Table 6 and Figure 7). At 1% FDR, pFind 2.1 identified from worm ETD data (doubly

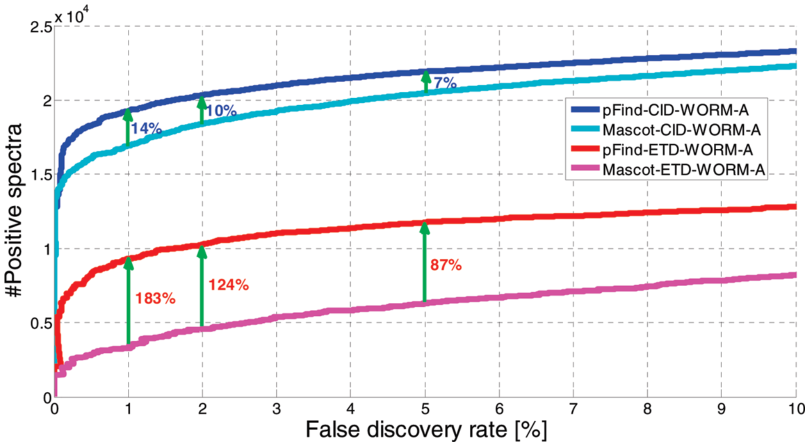
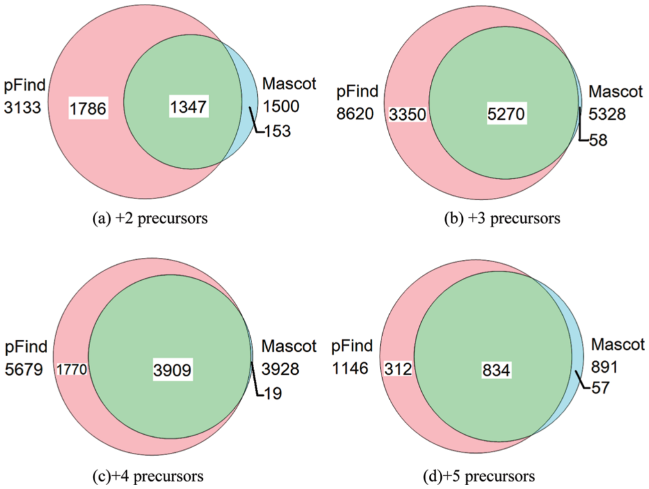


Figure 7. ROC curves of pFind and Mascot on WORM-A CID and ETD data sets. pFind identified 14% more CID spectra and 183% more ETD spectra than Mascot under at FDR 1%.

Table 5. Comparison of pFind, Mascot, and OMSSA on Four YEAST ETD Data Sets

data set (#spectra)	identifications	OMSSA (%[a])	Mascot (%[a])	pFind (%[a])	Mascot∩pFind[b] (%)	pFind-Mascot (%[c])	pFind-OMSSA (%[d])
YEAST-B1 (52 520)	#spectra	5603 (10.7)	3613 (6.9)	6388 (12.2)	3537 (97.9)	76.8	14.0
	#peptides	3951	2602	4130	2538 (97.5)	58.7	4.5
	#proteins	1472	1205	1592	1166 (96.8)	32.1	8.2
YEAST-B2R1 (59 485)	#spectra	12 193 (20.5)	10 124 (17.0)	17 040 (28.6)	9944 (98.2)	68.3	39.8
	#peptides	4962	3765	5723	3666 (97.4)	52.0	15.3
	#proteins	1648	1518	1983	1465 (96.5)	30.6	20.3
YEAST-B2R2 (59 007)	#spectra	11 906(20.2)	10 160 (17.2)	16 638 (28.2)	9984 (98.3)	63.8	39.7
	#peptides	4895	3822	5612	3712 (97.1)	46.8	14.7
	#proteins	1604	1554	1939	1482 (95.4)	20.9	20.9
YEAST-EtcaD (56 019)	#spectra	11 470 (20.5)	11 647 (20.8)	18 578 (33.2)	11 360 (97.5)	59.5	62.0
	#peptides	4585	4376	6325	4230 (96.7)	44.5	38.0
	#proteins	1582	1598	1999	1525 (95.4)	25.1	26.4

[a] Percentage of identified ETD spectra. [b] Number of overlapping results between Mascot and pFind. The overlap as a percentage of the Mascot result is shown in parentheses. [c] Difference between the number of pFind IDs and Mascot IDs, divided by the number of Mascot IDs. [d] Difference between the number of pFind IDs and OMSSA IDs, divided by the number of OMSSA IDs.



(a) +2 precursors

(b) +3 precursors

(c)+4 precursors

(d)+5 precursors

Figure 8. Venn diagrams showing the overlap of identified ETD spectra by pFind and Mascot. The number of identified ETD spectra of +2, +3, +4 or +5 peptides from the YEAST-ETcaD data are plotted separately.

Table 6. Comparison of pFind and Mascot on Phosphopeptide ETD data

	identifications (FDR=1%)	Mascot	pFind	Mascot∩ pFind[a]	Mascot∩ EpFind[b]
Phosphopeptides	#spectra	596	1581	560	1622
	#peptides	351	612	329	637
	#proteins	516	792	481	827

[a] Number of overlapping results between Mascot and pFind. [b] Number of combined results of Mascot with pFind.

charged tryptic peptides) twice to thrice as many spectra as Mascot, corresponding to an increase of 63−122% more peptides. Between the pFind and Mascot results, the overlap is 93−95% on either the spectrum, peptide or protein level (Table 3). When tested on the yeast ETD data (Lys-C peptides), pFind again outperformed Mascot with 45−59% more unique peptides as detailed in Table 5. The Venn diagrams in Figure 8 show that independent of the precursor charge state, the pFind results nearly contain all the Mascot results (overlap is ∼98% in the column "pFind∩Mascot" of Table 5 on the spectral level), and pFind identified many more additional ETD spectra (Table 5). With respect to CID data analysis, pFind 2.1 also performed better than Mascot (Table 4 and Figure 7).

We further compared the performance of pFind with OMSSA on ETD data analysis. pFind 2.1 showed a markedly better performance than OMSSA on the YEAST-ETcaD data by identifying 6325 peptides compared to 4585 by OMSSA, corresponding to a 38% increase (Table 5). On other YEAST ETD data sets, pFind 2.1 identified 5−15% more peptides over OMSSA (Table 5).

ETD is often used in analysis of peptides with "labile" PTMs such as phosphorylation at Ser or Thr residues, so we tested the performance of pFind 2.1 on phosphopeptides. For the phosphopeptide data set (Table 6), pFind identified 1581 phosphopeptide ETD spectra for 612 phosphopeptides while Mascot identified 596 phosphopeptide ETD spectra, corresponding to 351 phosphopeptides. Of the 612 phosphopeptides identified by pFind from ETD spectra, 527 (86.1%) were also confidently identified by pFind from their cognate CID spectra. This result argues for a high degree of reliability of phosphopeptide identification by pFind.

All these comparisons have shown that pFind 2.1 is a highly effective search engine for both ETD and CID data analysis, especially with ETcaD spectra of tryptic peptides. Much of the performance enhancement of pFind 2.1 comes from data preprocessing and taking into consideration the effects of precursor charge states on HR. This work shows for the first time that the HR patterns of ETD spectra can be used to effectively improve peptide identification.

In a previous study, ETcaD and ETD data from the same sample were analyzed using OMSSA and no improvement of ETcaD was found over ETD alone (without supplemental activation).[36] OMSSA identified very few doubly charged peptides although +2 ETD spectra accounted for 39.2% of the data.[36] Analyzing the same data using pFind and Mascot, we find that ETcaD resulted in many more peptide identifications than ETD alone, including 2,393 doubly charged peptides by pFind and 1,209 by Mascot. We think that this discrepancy has to do with how HR ions are taken into consideration in these search engines. OMSSA ignores HR ions while Mascot considers $z + \mathrm{H}$ but not $c - \mathrm{H}$ ions. pFind considers both $z + \mathrm{H}$ and $c - \mathrm{H}$ ions in a precursor charge state-dependent manner, resulting in a significant improvement over both OMSSA and Mascot (Table 5 and Figure 8). Coon predicted that "newer search engines built around ETD fragmentation patterns.. .will further improve ETD performance".[23] Our work has shown that this prediction is true and the newer search engine envisioned previously has been realized through pFind.

Discussion

Although ETD is now in widespread use in proteomics, the lack of an effective search engine has limited its application and further technology development. Let us take, as an example, the worm data acquired by CID/ETD double play (Table 1, WORM-A, -B, and -C). From a total of 175 534 pairs of CID/ETD spectra, Mascot identified on average 38.0% of the CID spectra and a mere 9.7% of the ETD spectra (calculated from Table 4 and Table 3, including both +2 and +3 peptides). With an identification rate almost one-fourth of that for CID, ETD makes itself a difficult choice for shotgun proteomics. Even for applications capitalizing on the unique advantages of ETD such as analysis of labile post-translational modifications, it is a pity to lose valuable information that is already acquired in the spectra. With pFind 2.1, we have doubled the identification rate of ETD spectra over that by Mascot. From the same 175 534 pairs of CID/ETD spectra, pFind2.1 identified on average 42.3% of the CID spectra and 19.4% of the ETD spectra (calculated from Table 4 and Table 3, including both +2 and +3 peptides). The identification rate of ETD spectra is still lower than that of CID. This difference might be due to a lower fragmentation efficiency of ETD than CID. In a CID spectrum, the precursor is almost completely fragmented. In an ETD spectrum, only a fraction of the precursor signal is converted to informative fragment ion signal while much is left in the form of an intact precursor or charge-reduced precursors, both of which make little contribution to peptide identification. So, we anticipate that additional improvement may come from increasing peptide fragmentation efficiency of ETD, better ionization, better ion transmission, or all of the above. From the informatics side, the region of CR precursors might be used to validate sequence identifications or flag the presence of a spectrum containing two or more precursors and trigger subsequent analysis.

Searching ETD Data with a Large Fragment Mass Tolerance is Not an Optimal Way to Utilize HR Information. A recent study recommended that a larger fragment tolerance window should be used to search ETD data compared to that used for CID, for example, ±1.1 Da with Mascot[18] and ±1.2 Da with Protein Prospector.[26] Such large mass tolerance windows increase the identification rate of ETD spectra, compared with the common setting such as ±0.5 Da. We think that the reason why a large fragment mass tolerance window can improve identification rate is that it allows HR ions (e.g., $z + \mathrm{H}$ and $c - \mathrm{H}$) to be included for spectral matching. Mascot considers $z + \mathrm{H}$ ions but it ignores $c - \mathrm{H}$ ions. We manually examined a subset of ETD spectra that were either missed or identified with a low confidence score by Mascot but whose cognate CID spectra had positive identifications by both Mascot and pFind. If we let pFind consider $c - \mathrm{H}$ ions besides c, z, and $z + \mathrm{H}$ ions, these spectra were then identified with high confidence (data not shown).

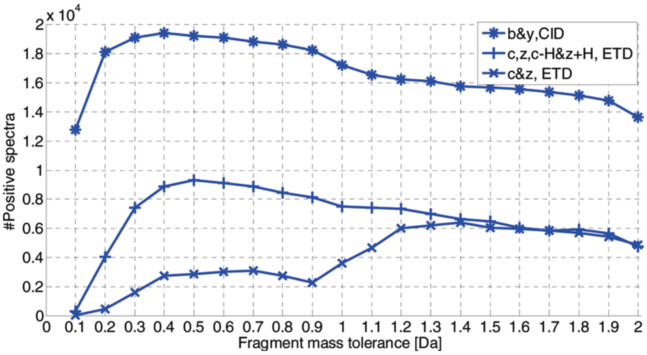


Figure 9. Effect of fragment mass tolerance on ETD or CID peptide identification. Doubly charged ETD or CID spectra from the WORM-A data set were searched repeatedly using pFind 2.1 while the fragment mass tolerance window was varied from ±0.1 to ±2.0 Da. The false discovery rates were no more than 1%.

Our results indicate that a larger fragment tolerance window is not the optimal solution to searching ETD spectra. If we do not let pFind consider $z + \mathrm{H}$ and $c - \mathrm{H}$ ions (only z and c ions matched), a fragment mass tolerance of ±1.4 Da gives a better result than other smaller or larger tolerance windows (Figure 9). However, a fragment mass tolerance of ±0.5 Da along with the inclusion of $z + \mathrm{H}$ and $c - \mathrm{H}$ ions increased the number of identified spectra by 45.6% (9293 to 6382) over a ±1.4 Da mass tolerance window (Figure 9). We think that although a larger mass tolerance can include HR ions in spectral matching, it is a compromised solution because it simultaneously increases random matches. Thus, we recommend searching ETD data generated by ion trap instruments with a mass tolerance window of ±0.5 Da along with the inclusion of $z + \mathrm{H}$ and $c - \mathrm{H}$ ions in a precursor charge state-dependent manner.

Abbreviations: CID, Collision induced dissociation; ECD, Electron capture dissociation; ETD, Electron transfer dissociation; ETcaD, Electron transfer and collisionally activated dissociation; SA, Supplemental activation; HR, Hydrogen rearrangement; LTQ, Linear ion trap; AGC, Automatic gain control; MS1, Parent scan mass spectra; MS2, Tandem mass spectra; AA, Amino acid; FDR, False discovery rate; ROC, Receiver operating characteristic.

Acknowledgment. We thank the reviewers for constructive suggestions that have helped improve this paper. This work was supported by National High Technology Research and Development Program of China (863) under Grant Nos. 2008AA02Z309, 2007AA02Z315, and 2007AA02Z1A7, National Key Basic Research and Development Program of China (973) under Grant No. 2010CB912701, CAS Knowledge Innovation Program under Grant No. KGGX1-YW-13, and National Natural Science Foundation of China under Grant No. 30900262.

Supporting Information Available: All WORM raw data produced in this manuscript can be downloaded from the ProteomeCommons.org Tranche network using the following three hashes: “Q7K9CAFZP6Af1Ls2H+D6xM6FhbNobUYxf0750MXbg/c1DaOohALKe6va6yPhVvUFY1N4oGgOSfAkrvxSPyDi9vG5FTsAAAAAAAAQgg==” “wCxYMJ105NfMCeME630L/7SqIFSwa5k94reX8I9V/k/m6Dkrcs8LBFYafgEjZrQee/K0tUGaP5Y67DphtvgejTmsMoQAAAAAAAAG9g==”“V+NCsJiGw40L1yeGaizW63jxmQhjcUkuEAHNmUa6KKgo+IqSH/Rh2sL5CGFlK6Flc8LUKLQipH8mAGTa4KMhOixoVooAAAAAAAAE6Q==”. pFind is freely available for academic users and can be obtained from the pFind website: http://pfind.ict.ac.cn. Supplementary text, figures, and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Fu, Y.; Yang, Q.; Sun, R.; Li, D.; Zeng, R.; Ling, X.; Gao, W. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **2004**, *20*, 1948–1954.

(2) Li, D.; Fu, Y.; Sun, R.; Ling, X.; Wei, Y.; Zhou, H.; Zeng, R.; Yang, Q.; He, S.; Gao, W. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **2005**, *21*, 3049–3050.

(3) Wang, L.; Li, D.; Fu, Y.; Wang, H.; Zhang, J.; Yuan, Z.; Sun, R.; Zeng, R.; He, S.; Gao, W. pFind 2.0: a software package for peptide and

protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2985–2991.

(4) pFind Web site: http://pfind.ict.ac.cn.

(5) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(6) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.

(7) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207.

(8) Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science* **2006**, *312*, 212–217.

(9) Han, X.; Aslanian, A.; Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12*, 483–490.

(10) Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.* **2009**, *11*, 49–79.

(11) 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*, 3265–3266.

(12) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.

(13) Williams, D. K.; McAlister, G. C.; Good, D. M.; Coon, J. J.; Muddiman, D. C. Dual electrospray ion source for electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal. Chem.* **2007**, *79*, 7916–7919.

(14) McAlister, G. C.; Berggren, W. T.; Raming, J. G.; Horning, S.; Makarov, A.; Phanstie, D.; Stafford, G.; Swaney, D. L.; Syka, J. E. P.; Zabrouskov, V.; Coon, J. J. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *J. Proteome Res.* **2008**, *7*, 3127–3136.

(15) Molina, H.; Matthiesen, R.; Kandasamy, K.; Pandey, A. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal. Chem.* **2008**, *80*, 4825–4835.

(16) 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*, 7170–7180.

(17) Sobotta, F.; Wattb, S. J.; Smithc, J.; Edelmannd, M. J.; Kramerd, H. B.; Kesslerd, B. M. Comparison of CID Versus ETD based MS/MS fragmentation for the analysis of protein ubiquitination. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1652–1659.

(18) Leinenbach, A.; Hartmer, R.; Lubeck, M.; Kneissl, B.; Elnakady, Y. A.; Baessmann, C.; Muller, R.; Huber, C. G. Proteome analysis of sorangium cellulosum employing 2D-HPLC-MS/MS and improved database searching strategies for CID and ETD fragment spectra. *J. Proteome Res.* **2009**, *8*, 4350–4361.

(19) Coon, J. J.; Ueberheide, B.; Syka, J. E. P.; Dryhurst, D. D.; Ausio, J.; Shabanowitz, J.; Hunt, D. F. Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 9463–9468.

(20) Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques* **2005**, *38*, 519–523.

(21) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F. The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta* **2006**, *1764*, 1811–1822.

(22) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **2007**, *6*, 1942–1951.

(23) Coon, J. J. Collisions or electrons? protein sequence analysis in the 21st century. *Anal. Chem.* **2009**, *81*, 3208–3215.

(24) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2199–2204.

(25) Swaney, D. L.; Wenger, C. D.; Thomson, J. A.; Coon, J. J. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 995–1000.

(26) Chalkley, R. J.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 8894–8899.

(27) Alley, W. R.; Mechref1, Y.; Novotny, M. V. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161–170.

(28) Han, H.; Pappin, D. J.; Ross, P. L.; McLuckey, S. A. Electron transfer dissociation of iTRAQ labeled peptide ions. *J. Proteome Res.* **2008**, *7*, 3643–3648.

(29) Phanstiel, D.; Unwin, R.; McAlister, G. C.; Coon, J. J. Peptide quantification using 8-plex isobaric tags and electron transfer dissociation tandem mass spectrometry. *Anal. Chem.* **2009**, *81*, 1693–1698.

(30) Chi, A.; Bai, D. L.; Geer, L. Y.; Shabanowitz, J.; Hunt, D. F. Analysis of intact proteins on a chromatographic time scale by electron transfer dissociation tandem mass spectrometry. *Int. J. Mass Spectrom.* **2007**, *259*, 197–203.

(31) Bunger, M. K.; Cargile, B. J.; Ngunjiri, A.; Bundy, J. L.; Stephenson, J. L. Automated proteomics of E. coli via top-down electron-transfer dissociation mass spectrometry. *Anal. Chem.* **2008**, *89*, 1459–1467.

(32) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(33) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.

(34) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(35) 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*, 477–485.

(36) Swaney, D. L.; McAlister, G. C.; Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nature Methods* **2008**, *5*, 959–964.

(37) O'Connor, P. B.; Lin, C.; Cournoyer, J. J.; Pittman, J. L.; Belyayev, M.; Budnik, B. A. Long-lived electron capture dissociation product ions experience radical migration via hydrogen abstraction. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 576–585.

(38) Savitski, 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*, 113–120.

(39) Savitski, M. New proteomics methods and fundamental aspects of peptide fragmentation; PhD theses, Uppsala University, 2007.

(40) Brenner, S. The genetics of Caenorhabditis elegans. *Genetics* **1974**, *77*, 71–94.

(41) Wessel, D.; Flugge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **1983**, *138*, 141–143.

(42) McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R. Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int. J. Mass Spectrom.* **2002**, *219*, 245–251.

(43) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2*, 667–675.

(44) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207–214.

(45) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787–797.

(46) Cooper, H. J.; Hudgins, R. R.; Håkansson, K.; Marshall, A. G. Characterization of amino acid side chain losses in electron capture dissociation. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 241–249.

(47) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. Side-chain losses in electron capture dissociation to improve peptide identification. *Anal. Chem.* **2007**, *79*, 2296–2302.

(48) Chalkley, R. J.; Brinkworth, C. S.; Burlingame, A. L. Side-chain fragmentation of alkylated cysteine residues in electron capture dissociation mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1271–1274.

(49) 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*, 8089–8094.

(50) Zeller, M.; Ueckert, T.; Delanghe, B. High confident protein identification of ETD and ECD spectra with a new mass list

preprocessor. In *Proceedings of the 56th ASMS Conference on Mass Spectrometry and Allied Topics*; Denver, CO, USA, June 1–5, 2008, M635.

(51) Sadygov, R. G.; Hao, Z.; Huhmer, F. R. Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Anal. Chem.* **2008**, *80*, 376–386.

(52) Carvalho, P. C.; Cociorva, D.; Wong, C. L.; Carvalho, D. C.; Barbosa, V. C.; Yates, J. R. Charge prediction machine: tool for inferring precursor charge states of electron transfer dissociation tandem mass spectra. *Anal. Chem.* **2009**, *81*, 1996–2003.

(53) Sadygov, R. G.; Good, D. M.; Swaney, D. L.; Coon, J. J. A new probabilistic database search algorithm for ETD spectra. *J. Proteome Res.* **2009**, *8*, 3198–3205.

(54) Liu, X.; Shan, B.; Xin, L.; Ma, B. Better score function for peptide identification with ETD MS/MS spectra. *BMC Bioinform.* **2010**, *11* (Suppl 1), S4.

(55) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villén, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12130–12135.

(56) 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*, 1435–1444.

(57) Sweet, S. M.; Jones, A. W.; Cunningham, D. L.; Heath, J. K.; Creese, A. J.; Cooper, H. J. Database search strategies for proteomic data sets generated by electron capture dissociation mass spectrometry. *J. Proteome Res.* **2009**, *8*, 5475–5484.

(58) Chalkley, R. J.; Medzihradszky, K. F.; Lynn, A. J.; Baker, P. R.; Burlingame, A. L. Statistical analysis of peptide electron transfer dissociation fragmentation mass spectrometry. *Anal. Chem.* **2010**, *82*, 579–584.

(59) Baker, P. R.; Medzihradszky, K. F.; Chalkley, R. J. Improving software performance for peptide ETD data analysis by implementation of charge-state and sequence-dependent scoring. *Mol. Cell. Proteomics* **2010**, *9*, 1795–1803.

PR100648R