Published in final edited form as:

J Proteome Res. 2012 February 3; 11(2): 668–677. doi:10.1021/pr200597j.

Improving CID, HCD, and ETD FT MS/MS degradome-peptidome identifications using high accuracy mass information

Yufeng Shen¹, Nikola Tolić², Samuel O. Purvine², and Richard D. Smith¹

¹Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354

²Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99354

Abstract

MS dissociation methods, including CID, HCD, and ETD, can each contribute distinct peptidome identifications using conventional peptide identification methods (Shen et al. *J. Proteome Res.* 2011), but such samples still pose significant informatics challenges. In this work, we explored utilization of high accuracy fragment ion mass measurements, in this case provided by FT MS/MS, to improve peptidome peptide dataset size and consistency relative to conventional descriptive and probabilistic scoring methods. For example, we identified 20–40% more peptides than SEQUEST, Mascot, and MS-GF scoring methods using high accuracy fragment ion information and the same FDR (e.g., <10 mass errors) from CID, HCD, and ETD spectra. Identified species covered >90% of the collective identifications obtained using various conventional peptide identification methods, which resolves the issue of different data analysis methods generating different peptide datasets. Choice of peptide dissociation and high-precision measurement-based identification methods presently available for degradomic-peptidomic analyses needs to be based on the coverage and confidence (or specificity) afforded by the method, as well as practical issues (e.g., throughput). By using accurate fragment information, >1000 peptidome peptides can be identified from a single human blood plasma sample with low peptide-level FDRs (e.g., 0.6%), providing an improved basis for investigating potential disease-related peptidome components.

Keywords

FT MS/MS; CID; HCD; ETD; peptides; non-tryptic peptides; peptidome; degradome; merging of spectra; scoring of spectra

INTRODUCTION

Endogenous peptides (the peptidome) resulting from protein proteolytic processing and degradation (the degradome) can be important in physiological and pathological processes (1). For example, endogenous protease activities are well-established components of many diseases, e.g., cancer (2), and our recent MS-based comparative analysis of pooled plasma samples from early-stage breast cancer patients and matched healthy controls revealed >1000 endogenous peptides and small proteins that suggested distinctive patterns of substrate proteolysis for cancer patients (3).

Requests for materials should be addressed to Yufeng Shen (Yufeng.shen@pnnl.gov) or Richard D. Smith (rds@pnnl.gov).

Supplementary figures and data files are included with this manuscript.

Characterization of the degradome-peptidome is challenging (4), as endogenous peptides generated from protein intracellular and intercellular degradation can have molecular weights of ~0.5–15 kDa (or more) and multiple terminal cleavage specificities (5,6). Identification of peptidome components benefits significantly from high-resolution/accuracy tandem MS (e.g., FT MS/MS) (3,5,6). Several dissociation methods, (e.g. CID, HCD, and ETD) can be used to generate tandem mass spectra (MS/MS) to extend identification coverage (7). The extent of useful and complementary information from each dissociation method depends significantly on the software tools and methods used. Additionally, an evaluation of CID, HCD, and ETD effectiveness for peptide identification using conventional software tools (e.g., SEQUEST, Mascot) and methods has been observed to produce inconsistent peptidome datasets with respect to size and content (7), and new methods are necessitated to improve both peptidome coverage and dataset consistency through better utilization of high accuracy mass information (e.g., provided by FT MS/MS).

A variety of software tools for spectral database searching and peptide validation have been developed (8). However, most were developed for low-resolution MS/MS data of smaller peptides with known termini (e.g. tryptic peptides), and their effectiveness generally declines with peptide size and terminal cleavage complexity due largely to the exponential growth of possible species. More confident peptide identifications can be obtained using high accuracy fragment ion data, especially for peptides with unknown modifications or amino acid substitutions (9). Software tools developed for top-down proteomics [e.g., the Thrash algorithm (10), Xtract (11), Decon_MS_n (12), MS-Decconv (13), YADA (14)] can be used to 'de-convolute' high-resolution peptide mass spectra, while Mascot (15), ProLuCID (16), and ProSight (17) have some applicability for searching databases of high-resolution fragment spectra. Mascot has been extensively applied for tryptic peptides, while ProLuCID can use ~20 ppm error data (15), but its performance with higher accuracy data (e.g., with <10 ppm errors) has not been reported. ProSightPC and modified Mascot [e.g., BIG Mascot, (18)] are presently the only commercial tools specifically designed for identifying large polypeptides (e.g., intact proteins) from high-resolution spectra, but their effectiveness for searching an extremely large number of spectra against large protein databases in an automated fashion remains uncertain.

Peptide fragments measured with high-precision FT MS/MS allow certain identification of peptides, which has been demonstrated using the UStags strategy (9). Long sequences (e.g., >7 residues) measured with high precision can unambiguously assign proteins and peptides from a large protein database (e.g., the IPI human protein database) provided the sequences uniquely belong to specific proteins (6). As a result of this strict requirement, some peptide assignments may be missed from spectra that would otherwise be assigned using more relaxed peptide identification criteria (or a certain false discovery rate, FDR). Reducing the sequence length required for the use of inconsecutive fragments can be applied for this purpose.

In this work, we explore identification of the peptidome using high-accuracy CID, HCD, and ETD FT MS/MS data categorized by counts of peptide backbone cleavages (CBC) for total fragment ion species, total cleavage sites from observed fragments, total residues from observed consecutive fragments, and sequence length from consecutive residues. We show that these approaches provide improvements in both peptide identification coverage and confidence without tradeoff of peptide identification FDR compared to representative spectral descriptive and probabilistic scoring methods currently applied for peptide identification. Further, we show that different CBC approaches in conjunction with a targeted FDR level work better for some dissociation methods than others.

METHODS

Descriptions of datasets

CID, HCD, and ETD FT MS/MS datasets (<http://www.ebi.ac.uk/pride/>) were obtained previously (7) using reversed-phase LC in conjunction with a LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Briefly, peptidome components were isolated from a human blood plasma using affinity and size exclusion chromatography (7). Reversed-phase LC was performed using a long packed capillary column (100 cm $\times$ 100 μ m i.d. $\times$ 3 μ m C4). The LC separation was performed at a flow rate of $\sim$ 0.8 μ L/min and with a gradient from mobile phase A acetonitrile/H₂O/acetic acid, 10:90:0.2, v/v/v) to B (acetonitrile/isopropyl alcohol/H₂O/acetic acid/trifluoroacetic acid, 60:30:10:0.2:0.1, v/v/v/v/v). FT MS and FT MS/MS measurements were obtained at 60K resolution with two micro scans and AGC targets of 1×10^6 and 3×10^5 for the first and second scans, respectively. A survey scan at $400 \leq m/z \leq 2000$ was followed by FT MS/MS of the three most intense ions from the survey scan (monoisotopic precursor selection not enabled). Each precursor was fragmented using CID, ETD, and HCD in this order prior to analysis of the next precursor. A normalized collision energy of 35% was applied for HCD and CID, and a reaction time of 300 ms as the default CS 2 was employed for ETD with supplementary activation enabled. Fragmentation of the most intense precursor was completed with an isolation window of 6 m/z units and a minimal signal of 2000. Dynamic exclusion was enabled with no repeat counts, using a 3 m/z tolerance and a duration cycle of 5 min. Mass calibration was performed according to the method provided by the instrument manufacturer.

CID, HCD, and ETD spectra were merged using VBA macros developed in house, and resulting datasets included CID/HCD, CID/ETD, HCD/ETD, and CID/HCD/ETD spectra.

Creation of input files for database searching

Both monoisotopic (non-de-convoluted) and deconvoluted precursor input files were created for spectral database searching. The monoisotopic precursor input files were created using Extract_MS_n (version 5.0, Thermo Fisher Scientific) as previously described (7). De-convoluted precursor input files were created using Decon_MS_n (12) wherein the precursor mass was corrected by -4.5 ppm based upon statistical evaluation of a confidently identified peptide subset. The scan header information was used to apply the appropriate charge state (CS) and parent mass values for each MS/MS spectrum, and an upper mass tolerance of 25 KDa was set for input file creation.

The protein database for this work was constructed by combining the IPI human protein database with a decoy database of reversed IPI database protein sequences (7). The combined database contained 139,462 protein entries in total. SEQUEST (version 27, revision 12, Thermo Fisher Scientific) and Mascot (version 2.3.01, Matrix Science Inc., Boston, MA) were employed for spectral database searches. For SEQUEST, the monoisotopic files were searched with 5-Da precursor tolerance and 0.05-Da fragment ion tolerance (7), while de-convoluted files were searched with 2.5- and 10-ppm precursor mass tolerances and 0.05 Da monoisotopic fragment ion tolerance for this work. CID and HCD employed *b*- and *y*-type ions, while ETD spectra employed *c*- and *z*-type ions for all files searched. With Mascot, CID and HCD spectra were searched using Instrument option ESI-FTICR, and ETD spectra using Instrument option ETD-Trap, as described previously (7) with peptide mass tolerances of 5 Da, 10 ppm, and 2.5 ppm and fragment mass tolerances of 0.05, 0.1 and 0.005 Da for monotopic masses and isotopic correction. Neither amino acid modifications nor enzymes were specified for both SEQUEST and Mascot database searches. The same database search methods were used for merged spectra; CID/HCD

spectra employed *b*- and *y*-type ions, while CID/ETD, HCD/ETD, and CID/HCD/ETD spectra employed *b*-, *y*-, *c*-, and *z*-type ions.

Peptide validation using descriptive and probabilistic scoring methods

Previously described scoring methods (7) were used to validate candidates output from SEQUEST and Mascot database searches. Briefly, validation using SEQUEST was accomplished as follows: top-1 candidates were filtered with $\Delta C_n > 0.1$ and were accepted if their Xcor values were higher than the thresholds used to generate the desired spectrum-level FDR for each charge state. For Mascot, top-1 candidates were accepted when scores were higher than the targeted FDR threshold for each charge state. The spectrum-level FDR was calculated as $2N_{decoys} / (N_{targets} + N_{decoys})$, where N represents the number of spectra assigned for targets or decoys from the combined protein database (19). The same calculation was applied for peptide-level FDR, but in this case N is the number of different peptides assigned for targets or decoys.

MS_GF (<http://proteomics.ucsd.edu/Software/MSGeneratingFunction.html>), an improved spectral probabilistic scoring method (20), was also tested for database searching. The version tested was not effective for peptidomic data since it did not support “no-enzyme” data analysis and MS_GF method is now being improved by its developers. Therefore, in this work, MS_GF probabilities were used to verify candidates from SEQUEST and Mascot searches. The top hit output from SEQUEST and Mascot searches was used for each spectrum and the probability value was calculated. Validation was based on either probability values of individual spectra or the lowest probability value of all spectra assigned to the same candidate to achieve the targeted peptide-level FDR.

Peptide verification using high accuracy fragment measurements and relevant backbone cleavage information

The top hits from SEQUEST and Mascot searches were used as candidates for peptide validation. The unique sequence tags (USTags) methodology (7) was used to determine the peptide fragments within a mass error tolerance of 10 ppm. For merged spectra, the candidate *c* or *z* fragments assigned from ETD spectra were mapped to the corresponding *b* or *y* fragments, respectively, of the same candidate, using ICR2LS (<http://ncrr.pnl.gov/software/>) and then added to the lists of candidates derived from CID and/or HCD spectra. The fragments were used to determine cleavages, amino acid residues, and sequences. In other cases candidates were first ordered according to search scores (i.e., from large to small) and the observed total fragments (TF), total cleavage sites (TS), total residues (TR), or the sequence length (SL) (from high to low values, depending on the variation of CBC used) assigned for each candidate. Then, the TF, TS, TR, and SL counts were selected to achieve the targeted spectrum-level FDR (as described above) and peptide-level FDRs (the spectrum providing the best CBC value was used for assignment of each candidate regardless of CS).

RESULTS

Improving identification scoring using high accuracy measurements

We first examined SEQUEST and Mascot scoring with high-accuracy mass measurements (Table 1). Without de-convolution of high-resolution mass spectra (e.g., using Extract_MS_n to create SEQUEST input files), 1012 peptidome components were identified using the combination of CID, HCD, and ETD in the previous work (7). With control of the same spectrum-level FDR (i.e., 2%), the combination of identifications from individual dissociation spectra was observed to increase the peptide-level FDR. Reducing precursor mass tolerance from 5 Da to 2.5 ppm (the input files were created with Decon_MS_n de-

convolution function) resulted in a 7% increase in the number of identified peptides. The peptides identified with large (5 Da) mass tolerance were mostly (91%) covered by those identified with small (2.5-ppm) mass errors. For those with 5-Da mass errors, most peptides excluded (96%) had precursor mass errors $\gg 10$ ppm with allowance of ± 1.000 Da shift for possible de-convolution errors. Based on variations in mass accuracy for precursor species, we believe that these peptide identifications are incorrect, i.e., 9% false positives exist although the identifications were achieved with 2% spectrum-level FDR.

Mascot increased the number of peptides collectively identified from CID, HCD, and ETD spectra by $>60\%$ by reducing precursor mass tolerance from 5 Da to 2.5 ppm. Reducing fragment mass tolerance from 0.05 Da to 0.01 Da further increased the number of identified peptides by 20%. Such improvements stopped at fragment mass tolerance of 0.005 Da. Combinations of CID, HCD, and ETD consistently improved the number of identified peptides, and also increased peptide-level FDRs regardless of the mass tolerances applied for peptide identification. Mascot still provided fewer peptide identifications than SEQUEST even at small mass tolerances (e.g., 2.5 ppm).

Identification of peptidome components from merged spectra using SEQUEST and Mascot

We next investigated use of merged spectra for potential improvements in peptide identifications; results are listed in Table 2. Merging different dissociation spectra led to identification of fewer peptides than CID spectra alone with either SEQUEST or Mascot at a specific spectrum-level FDR (e.g., 2%), regardless of precursor or fragment mass accuracies applied. This was observed even for merging spectra that contained the same types of fragment ions (e.g., CID and HCD spectra). We examined the database scores of merged spectra and found increased SEQUEST scores for both targets and decoys (Supplementary Figures 1 and 2). The increased scores for decoys were necessary to prevent increasing SEQUEST identifications from the merged spectra at a targeted FDR. Merging different dissociation spectra had minor effects on Mascot scores for both targets and decoys (Supplementary Figure 3), and no improvement in the number of peptide identifications were observed with Mascot.

Extending peptidome identifications using high accuracy fragment information

The ability of CBC to distinguish fragment ion matches from targets and decoys was examined. As shown in Figure 1, the difference between targets and decoys was evident for <10 ppm mass error (few decoys had ≥ 7 matched fragments). The ability to distinguish is reduced for mass errors of 0.05 Da (i.e., 25–50 ppm errors for 1000–2000 Da fragments) where large numbers of decoys had significant matched fragments. The Orbitrap FT MS/MS provided the mass measurement accuracy needed for effective CBC use (Supplementary Figures 4 and 5).

Performance for CS 2–5 peptides is shown in Figures 2 to 4 for two CBC versions. For the assignment of CID spectra (Figure 2), CBC_TS and CBC_SL behaved similarly over a 0–10% FDR range. CBC distinguished targets and decoys than SEQUEST and Mascot scores, which is evidenced by the number of assigned spectra increasing slowly with increasing FDR using CBC compared with conventional SEQUEST and Mascot. Note, CBC was particularly effective for assignment of spectra at low FDR, e.g., $<3\%$. CBC and SEQUEST approached the same performance for assignment of CS 2 peptides, and both methods assigned more spectra than Mascot for low CS peptides. CBC assigned ~ 2 -fold more spectra for CS 3–4 peptides and >5 -fold more spectra for CS 5 peptides than SEQUEST at low FDR (e.g., $<1\%$). For assignment of HCD spectra (Figure 3), two versions of CBC exhibited similar performance for various CS peptides. CBC assigned a similar number of HCD spectra as SEQUEST, but slightly more than Mascot for CS 2 peptides. For higher CS

peptides CBC assigned ~2-fold more spectra than the two scoring methods at low FDR (e.g. <2%). For assignment of ETD spectra (Figure 4), two versions of the CBC method behaved differently, i.e., CBC_TS assigned more ETD spectra than CBC_SL for all CS examined. CBC_TS was more effective than the two scoring methods for ETD with CS 2–4 peptides, while few ETD CS 5 peptides were identified regardless of the method applied. We note that different CBC values can be needed for different dissociation spectra to achieve the same FDR level. For example, the TF values needed to achieve a 2% spectrum-level FDR were 7 for CID, 11 for HCD, and 3 for ETD; and the TS, TR, and SL values required were generally one less than the corresponding TF values at the same FDR level.

CBC allows straight-forward selection of peptide-level FDR (Method section), a capability of more practical value for peptide dataset quality control than spectrum-level FDR levels as applied in SEQUEST and Mascot scoring methods. Figure 5 shows CBC identifications from CID, HCD, and ETD spectra for peptide-level FDRs. Four CBC versions had similar performance for CID spectra at moderate (e.g., $\geq 2\%$) FDRs. However, these same four versions behaved differently for HCD and ETD spectra, i.e., CBC_SL and CBC_TS were the best and worst, respectively, for HCD, but CBC_TF and CBC_TR were the best and worst respectively for ETD. These observations reflect HCD and ETD differences in generating consecutive fragments (needed for CBC_TR and CBC_SL) and total fragments (used for CBC_TF and CBC_TS, which does not require consecutive fragments). CBC significantly improved the number of identifications over a large FDR range compared with other peptide-level FDR-controlled identification methods, such as MS_GF (20), SEQUEST-output fragment ions, and SP values.

We examined the coverage and peptide-level FDR quality of peptide datasets identified by CBC assignment of CID, HCD, ETD, and their merged spectra at a specific spectrum-level FDR; results are listed in Table 3. CBC identified 20–120% more peptides from CID spectra than HCD and ETD without sacrificing the identification peptide-level FDR. Combining the HCD and ETD contributions can add 5–10% more peptides to that obtained from CID spectra; however, such additional identifications were at the cost of nearly doubling the peptide-level FDR. (Note, these additional peptides were identified at the same spectrum-level FDR. Merging CID, HCD, and ETD spectra did not significantly increase the number of identified peptides; the best performance was observed for CID/ETD spectra, which provided ~2% more CBC_TF peptides than CID spectra with a slight reduction of peptide-level FDR. HCD merged spectra, including HCD/CID, HCD/ETD, and HCD/CID/ETD provided less identifications than CID alone when FDR was used to control spectral assignment due to HCD spectra introducing significant random hits of decoy fragments, even though measured with high precisions (e.g., <10 ppm mass errors). This occurrence was further evidenced by higher FDR values produced from HCD spectra-assigned UStags (see results below). Compared to the maximal identifications obtainable from all CID, HCD, and ETD spectra from conventional scoring methods (1089 peptides at ~4.5% peptide-level FDR; results shown in Tables 1 and 2), CBC identified more peptides (1170–1180 peptides) at lower peptide-level FDRs (e.g., ~2.6%) from a single type of spectra (CID spectra).

In addition to improving peptide identification coverage, CBC also provided datasets that covered most peptides identified by other different methods (7), i.e., improving peptide dataset consistency. As shown in Figure 7, 94–99% of SEQUEST or Mascot identifications from CID, HCD, or ETD spectra were covered by CBC, and 92% of all peptides identified with both scoring methods were covered by CBC. MS_GF provided ~15% of the peptides that differed from SEQUEST or Mascot scoring methods; however, CBC covered >95% of all peptides identified by either SEQUEST plus MS_GF or Mascot plus MS_GF without sacrificing confidence levels (Figure 8).

Modifying UStags for improving peptidome identifications at low FDR

The UStags approach is a variation of CBC_SL that excludes sequence homologues (9) to provide identifications with extremely low (near zero) FDR (7, 9). In this work, we modified the UStags to utilize different types of accurate fragment ions (or merged fragment ions of *b* and *y* or *c* and *z* ions) for extended coverage of peptidome peptide identifications.

CID contributed the most UStags identifications of the three dissociation methods by way of different types of fragment ions-constructed UStags (Table 3), while HCD was the major contributor for peptide identifications from single type fragment ions-constructed UStags (7). The use of different types of fragment ions from CID to construct UStags provided ~60% more identifications than identified with single-type fragment ions-constructed UStags (7). Merging different dissociation spectra resulted in additional 15–20% increases in UStags identified peptides, and CID/HCD was slightly better based upon peptide coverage.

We note that improved UStags coverage by merging different types of fragment ions or dissociation spectra modestly reduced identification confidence. Merging different types of fragment ions can result in 0.5% peptide-level FDR for UStags-identified peptide datasets (typically, 1–2 decoy peptides were observed from CID and HCD spectra), compared with 0% FDR produced from the use of single type fragment ions. Merging different dissociation spectra introduced some new false UStags (e.g., CID/HCD spectra produced one new decoy peptide) and increased peptide-level FDR. In spite of this, the modified UStags method was effective for peptide identification at relatively low FDR levels (e.g., <1%). For example, 10% more total peptides were identified with UStags from CID, HCD, and ETD at six-fold lower peptide-level FDR than with Mascot, and an equivalent number of peptides as SEQUEST with one-sixth the FDR (Tables 1 and 3). When CBC peptide identification FDR was the same as that produced by UStags, the UStags provided 19–70% more peptides than CBC_TS and CBC_TR, but ~6% fewer than CBC_TF and CBC_SL from CID spectra; and 10–400% more peptides than CBC_TF, CBC_TS, and CBC_TR, but an equivalent number of peptides as CBC_SL from HCD spectra. Similar to residue and sequence-based CBC_TR and CBC_SL, UStags were less effective than individual fragment-CBC_TF and CBC_TS for ETD. Regardless, the modified UStags method enabled identification of >1000 components at 0.8% peptide-level FDR from a single plasma sample LC-MS/MS analysis, in contrast to conventional methods where fewer plasma peptidome components were identified from more than two dozen samples at an unknown peptide-level FDR (21).

DISCUSSION

Broad, confident, and consistent identification of peptidome components, including cleavages and terminal modifications will benefit elucidation of degradome activities and discovery of new disease markers. Many approaches have been developed for small tryptic peptides, and in this work we examined two database search tools, i.e., descriptive SEQUEST and probabilistic Mascot, for identification of peptidome components. The scoring systems benefit from use of the high accuracy mass measurements (e.g. <10 ppm), and SEQUEST was found to provide more identifications than Mascot at the same FDR level, likely due to its better effectiveness with multiple terminal cleavages and with multiple CS species. Both scoring methods, however, displayed coverage limitations regardless of mass tolerances for precursors (5 Da to 2.5 ppm) and fragments (0.05 Da to 0.005 Da).

High accuracy fragment data were found effective for distinguishing targets from decoys, and direct use of high-accuracy fragment information (i.e., CBC) provided a significant improvement in the peptidome identification coverage compared with conventional

methods. Different variations of CBC method were explored, and all were similarly effective using CID spectra. However, HCD and ETD benefitted from different CBC versions; the residue and sequence-based CBC variations (e.g., CBC_TR and CBC_SL) were most suitable for HCD, while the individual fragment-based CBC methods (CBC_TF and CBC_TS) were best for ETD. A similar approach (referred to as “H-score”, ref. 22) as CBC_TS has been recently applied for improving identifications of tryptic peptides from HCD FT MS/MS spectra; however, other CBC versions should be better for this purpose based upon this work. In contrast to SEQUEST and Mascot, CBC identifications (regardless of peptide CS) allowed direct control of peptide-level FDR. Based upon the same peptide-level FDR, CBC identified more (e.g., by 30–40%) peptides than SEQUEST or Mascot from CID, HCD, and ETD spectra. In addition to improved coverage, CBC provided datasets that covered most (e.g., >90%) peptides identified by SEQUEST, Mascot, and MS_GF, and without sacrifice of confidence level. Further improvement is needed for spectra that contain a limited number (e.g., 3–6) of high accurately matched fragments. In this regard, we have initially explored combinational use of sequenced cleavage sites, residues, and sequence length information, but no gains were obtained.

UStags (9) is a sequence-based variation of CBC and the modified UStags method presented in this work identified 60% more peptides than the previous approach; equivalent to or better than that obtainable from SEQUEST and Mascot, but with 1/6 the peptide-level FDR. In comparison with other variations of CBC, the UStags identified more peptides than CBC_TS, CBC_TR, and CBC_SL from CID and HCD spectra, but slightly fewer (6%) than CBC_TF using CID spectra, at the same peptide-level FDR. ETD was less effective for UStags identification. Modification of UStags for merging of different types of dissociation spectra reduced their specificity, but the resultant peptide-level FDR can be limited to <1%.

The use of different dissociation spectra to improve identifications is of interest (23–25) but needs to be carefully considered, e.g., in the context of achievable FDR. Merging CID, ETD, and/or HCD spectra tended to match more fragments from both target and decoy peptides, although these fragments were measured with high precision (e.g., <10 ppm mass error). For FDR-controlled methods such as SEQUEST, Mascot, and CBC that were applied in this work, increased random decoy hits make it difficult for the use of merged spectral fragments to improve identification coverage. The fragment intensity information may have potential to further improve discrimination of targets and decoys, and then merged spectral identification coverage. For non-FDR-controlled methods such as UStags, merging different dissociation spectra improved identification coverage, but resulted in increased FDR values.

A range of peptide dissociation and identification approaches is now available for degradomic-peptidomic analyses using high-accuracy tandem mass spectrometry, and deciding which is the best for a particular application should be based upon their coverage, confidence (or specificity), and practical issues (e.g., throughput). UStags constructed with single type fragment ions are a first choice when the goal is to achieve specific identification of individual peptides. Then, different types of ions-constructed UStags and CBC can be used to search for peptides that cannot be specifically identified, but may be assignable with certain FDRs (e.g., 0.5–5%). When higher peptide-level FDRs (e.g., 5–10%) are allowed, different peptide dissociations can be used as an alternative to maximize identification coverage. In addition, combinational use of CID, HCD, and ETD spectra may provide information useful for locating modified amino acid residues not investigated in this work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

CID	collision induced dissociation
HCD	high energy collision dissociation
ETD	electron transfer dissociation
FT MS/MS	Fourier transform tandem mass spectrometry
FDR	false discovery rate
UStags	unique sequence tags
CS	charge state(s)

Acknowledgments

This research was partially supported by the NIH National Center for Research Resources (RR18522). Work was performed in the Environmental Molecular Science Laboratory, a U.S. Department of Energy (DOE) Office of Biological and Environmental Research (DOE/BER) national scientific user facility located on the campus of Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle for the DOE under contract DE-AC05-76RLO-1830.

REFERENCES

- Walsh, CT. Posttranslational modification of proteins. Expanding nature's inventory. Robert and Company Publishers; Colorado, USA: 2006. Chapter 8.
- Edwards, D.; Høyer-Hansen, G.; Blasi, F.; Sloane, BF., editors. The cancer degradome: proteases and cancer biology. Springer-verlag; New York: 2008.
- Shen Y, Tolić N, Liu T, Zhao R, Petritis BO, Gritsenko MA, Camp DG II, Moore RJ, Purvine SO, Esteva FJ, Smith RD. Blood peptidome-degradome profile of breast cancer. PLoS ONE. 2010; 5(10):e13133. [PubMed: 20976186]
- Lópezín-Otín C, Overall CM. Protease degradomics: a new challenge for proteomics. Nat. Rev. Mol. Cell Biol. 2002; 4:544–558.
- Shen Y, Hixson KK, Tolić N, Camp DG, Purvine SO, Moore RJ, Smith RD. Mass spectrometry analysis of proteome-wide proteolytic post-translational degradation of proteins. Anal. Chem. 2008; 80:5819–5828. [PubMed: 18578501]
- Shen Y, Liu T, Tolić N, Petritis BO, Zhao R, Moore RJ, Purvine SO, Camp DG II, Smith RD. Strategy for degradomic-peptidomic analysis of the human blood plasma. J. Proteome Res. 2010; 9:2339–2345. [PubMed: 20377236]
- Shen Y, Tolić N, Xie F, Zhao R, Purvine SO, Schepmoes AA, Moore RJ, Anderson GA, Smith RD. Effectiveness of CID, HCD, and ETD in FT MS/MS analysis of peptide: evaluations from different peptide identification methods. J. Proteome Res. 2011
- Nesvizhskii A, Vitek O, Aebersold RA. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods. 2007; 4:787–797. [PubMed: 17901868]
- Shen Y, Tolić N, Hixson KK, Purvine SO, Pasă-Tolić L, Qian WJ, Adkins JN, Moore RJ, Smith RD. Proteome-wide identification of proteomics and their modifications with decreased ambiguities and improved false discovery rates using unique sequence tags. Anal. Chem. 2008; 80:1871–1882. [PubMed: 18271604]
- Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large proteins. J. Am. Soc. Mass Spectrom. 2000; 11:330–332.
- Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL. New and automated ms approaches for top-down identification of modified proteins. J. Am. Soc. Mass Spectrom. 2005; 16:2027–2038. [PubMed: 16253516]
- Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JK, Smith RD. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics. 2008; 24:1021–1023. [PubMed: 18304935]

13. Carvalho PC, Xu T, Han X, Cociorva D, Barbosa VC, Yates JR III. YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics*. 2009; 25:2734–2736. [PubMed: 19684088]
14. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edward N, Souda P, Whitelegge JP, Bafna V, Pavzner PA. Deconvolution and database search of complex tandem mass spectra of intact proteins. *Mol. Cell. Proteomics*. 2010; 9:2772–2782. [PubMed: 20855543]
15. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification searching sequence database using mass spectrometry data. *Electrophoresis*. 1999; 20:3551–3567. [PubMed: 10612281]
16. casb.ucsd.edu/proteomics/posters/Prolucid_by_taoxu.pdf
17. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. ProSight PTM 2.0: Improved protein identification and characterization for top down mass spectrometry. *Nucleic. Acids Res*. 2007; 35:W701–W706. [PubMed: 17586823]
18. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong P, Easterling ML, Agar JN. Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Mol. Cell. Proteomics*. 2009; 8:846–856. [PubMed: 19074999]
19. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*. 2007; 4:207–214. [PubMed: 17327847]
20. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy database. *J. Proteome Res*. 2008; 7:3354–3363. [PubMed: 18597511]
21. Bakun M, Karczmarski J, Poznanski J, Rubel T, Rozga M, Malinowska A, Sands D, Hennig E, Oledzki J, Ostrowski J, Dadlez M. An integrated LC-ESI-MS platform for quantitation of serum peptide ladders. Application for colon carcinoma study. *Proteomics Clin. Appl*. 2009; 3:932–946. [PubMed: 21136997]
22. Savitski M, Mathieson T, Becher I, Bantscheff M. H-Score, a mass accuracy derived rescoring approach for improved peptide identification in modification rich samples. *J. Proteome Res*. 2010; 9:5511–5516. [PubMed: 20836569]
23. Nielsen ML, Savitski MM, Zubarev RA. Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell. Proteomics*. 2005; 4:835–845. [PubMed: 15772112]
24. Dayon L, Pasquarello C, Hoogland C, Sanchez J-C, Scherl A. Combining low- and high-energy tandem mass spectra for optimized peptide quantification with isobaric tags. *J. Proteomics*. 2010; 73:769–777. [PubMed: 19903544]
25. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJ, Pevzner PA. The generating function, of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics*. 2010; 9:2840–2852. [PubMed: 20829449]

Figure 1. Array of peptide fragment counts matched for targets (blue solid dots) and decoys (red circles). All top peptide candidates output from SEQUEST for individual FT MS/MS CID spectra are displayed.

Figure 2.

Assignments of CID FT MS/MS spectra with CBC and scoring methods for CS 2–5 peptides. CBC_TS and CBC_SL are used for the CBC method; assignments of spectra with the CBC, SEQUEST, and Mascot at a specific FDR as described in the test. (A): CBC vs. SEQUEST; (B): CBC vs. Mascot.

Figure 3. Assignments of HCD FT MS/MS spectra with CBC and scoring methods for CS 2–5 peptides. (A): CBC vs. SEQUEST; (B): CBC vs. Mascot. Other conditions same as Figure 2.

Figure 4. Assignments of ETD FT MS/MS spectra with CBC and scoring methods for CS 2–5 peptides. (A): CBC vs. SEQUEST; (B): CBC vs. Mascot. Other conditions same as Figure 2.

Figure 5. Peptide identification with four versions of CBC method from CID, HCD, and ETD spectra through control of identification peptide-level FDR.

Figure 6. Comparison of CBC, MS_GF probability score, and SEQUEST fragments and SP value for peptide identification from CID spectra with control of peptide-level FDR.

Figure 7.

Overlaps of peptides identified with CBC, SEQUEST, and Mascot from CID, HCD, and ETD spectra. (A): CBC_TF, CBC_TS, and CBC_SL; (B): CBC_TS, CBC_SL, and SEQUEST scoring; (C): CBC_TS, CBC_SL, and Mascot scoring; (D) CBC, SEQUEST scoring, and Mascot scoring. All peptides were identified at a 2% spectrum-level FDR.

Figure 8.

CBC covers most peptides identified from SEQUEST, Mascot, and MS_GF scoring methods. CBC_TF for the CBC method. SEQUEST and Mascot datasets obtained with 10 ppm mass tolerance for precursors and 0.05 Da for fragments; the MS_GF peptide dataset was identified by filtration of SEQUEST or Mascot top-1 hits with MS_GF probability scores; all peptides were identified from CID spectra at 2% spectrum-level FDR.

Table 1
Numbers of peptides identified from conventional scoring of CID, HCD, and ETD FT MS/MS spectra.*

Methods	SEQUEST	Mascot	Methods	SEQUEST	Mascot
<i>Extract_MSn [5Da, 0.05Da]</i>					
CID	747 (3.0%)	324 (5.0%)	CID	899 (2.6%)	704 (2.2%)
HCD	665 (2.8%)	329 (5.4%)	HCD	705 (2.2%)	569 (2.0%)
ETD	439 (2.2%)	276 (3.6%)	ETD	591 (2.0%)	448 (2.2%)
CID+HCD	916 (4.4%)	450 (7.6%)	CID+HCD	987 (2.0%)	786 (3.0%)
CID+ETD	894 (3.6%)	449 (5.8%)	CID+ETD	1030 (3.0%)	799 (3.2%)
HCD+ETD	809 (3.4%)	441 (6.4%)	HCD+ETD	898 (2.6%)	690 (2.6%)
CID+HCD+ETD	1012 (4.8%)	536 (7.8%)	CID+HCD+ETD	1086 (4.4%)	859 (4.0%)
<i>Decon_MSn [10 ppm, 0.05Da]</i>					
CID	837 (2.4%)	646 (2.2%)	CID	---	737 (2.8%)
HCD	669 (2.0%)	516 (2.4%)	HCD	---	664 (2.4%)
ETD	567 (2.0%)	363 (2.0%)	ETD	---	544 (2.0%)
CID+HCD	928 (3.4%)	727 (3.6%)	CID+HCD	---	840 (4.2%)
CID+ETD	986 (3.2%)	727 (2.8%)	CID+ETD	---	854 (3.0%)
HCD+ETD	860 (2.8%)	631 (2.8%)	HCD+ETD	---	787 (2.8%)
CID+HCD+ETD	1039 (4.2%)	791 (3.6%)	CID+HCD+ETD	---	926 (4.6%)
<i>Decon_MSn [2.5 ppm, 0.01Da]</i>					
CID	837 (2.4%)	646 (2.2%)	CID	---	737 (2.8%)
HCD	669 (2.0%)	516 (2.4%)	HCD	---	664 (2.4%)
ETD	567 (2.0%)	363 (2.0%)	ETD	---	544 (2.0%)
CID+HCD	928 (3.4%)	727 (3.6%)	CID+HCD	---	840 (4.2%)
CID+ETD	986 (3.2%)	727 (2.8%)	CID+ETD	---	854 (3.0%)
HCD+ETD	860 (2.8%)	631 (2.8%)	HCD+ETD	---	787 (2.8%)
CID+HCD+ETD	1039 (4.2%)	791 (3.6%)	CID+HCD+ETD	---	926 (4.6%)

* All peptides were identified at a 2% spectrum-level FDR; the numbers labeled in the parentheses represent the resultant peptide-level FDR values from control of the spectrum-level FDR; [xxx, yyy] represents the precursor mass tolerances of xxx and fragment mass tolerance of yyy applied for database search and peptide identification.

--- not available from automatic database search.

Table 2
Number of peptides identified from scoring of CID, HCD, and ETD merged FT MS/MS spectra.*

Methods	SEQUEST	Mascot	Methods	SEQUEST	Mascot
<i>Extract_MSn [5Da, 0.05Da]</i>					
CID/HCD	771 (3.2%)	425 (3.2%)	CID/HCD	835 (2.2%)	623 (2.0%)
CID/ETD	657 (3.0%)	368 (2.0%)	CID/ETD	718 (2.2%)	568 (2.0%)
HCD/ETD	460 (2.2%)	418 (2.4%)	HCD/ETD	592 (2.0%)	578 (2.8%)
CID/HCD/ETD	632 (2.2%)	376 (2.2%)	CID/HCD/ETD	598 (2.0%)	639 (2.2%)
<i>Decon_MSn [10 ppm, 0.05Da]</i>					
CID/HCD	721 (3.6%)	545 (2.2%)	CID/HCD	---	679 (2.0%)
CID/ETD	643 (2.2%)	422 (2.4%)	CID/ETD	---	669 (3.0%)
HCD/ETD	507 (2.0%)	454 (2.0%)	HCD/ETD	---	636 (2.0%)
CID/HCD/ETD	466 (2.6%)	527 (2.0%)	CID/HCD/ETD	---	693 (2.0%)

* All the conditions and symbols are the same as for Table 1.

Table 3

Number of peptides identified from CBC measurements of CID, HCD, ETD, and their merged FT MS/MS spectra.*

Methods	CBC_TF	CBC_TS	CBC_TR	CBC_SL	UStags
CID	1173 (2.6%)	1174 (2.8%)	1146 (2.6%)	1180(2.6%)	884 (0.2%)
HCD	856 (2.8%)	837 (2.6%)	946 (2.6%)	913 (2.8%)	808 (0.5%)
ETD	689 (2.4%)	689 (2.4%)	518 (2.2%)	556 (2.6%)	372 (0.0%)
CID+HCD	1229 (4.8%)	1216 (4.4%)	1248 (4.2%)	1251 (4.4%)	994 (0.6%)
CID+ETD	1236 (3.6%)	1227 (4.0%)	1247 (3.4%)	1225 (3.6%)	910 (0.2%)
HCD+ETD	1009 (4.0%)	1002 (3.8%)	1046 (3.4%)	995 (4.0%)	834 (0.5%)
CID+HCD+ETD	1268 (5.6%)	1254 (5.6%)	1307 (5.0%)	1273 (5.4%)	1020 (0.6%)
CID/HCD	1131 (3.2%)	1033 (3.4%)	1017 (3.0%)	1091 (3.2%)	1031 (0.8%)
CID/ETD	1199 (2.2%)	1184 (2.0%)	1021 (2.4%)	1124 (2.4%)	958 (0.2%)
HCD/ETD	969 (2.4%)	931 (2.4%)	884 (2.0%)	934 (2.0%)	883 (0.5%)
CID/HCD/ETD	1160 (2.0%)	1136 (2.2%)	986 (2.0%)	1146 (2.0%)	1063 (0.8%)

* All CBC peptides were identified at 2% spectrum-level FDR; the numbers labeled in the parentheses represent the resultant peptide-level FDR values.