Published in final edited form as:

Proteomics. 2010 March ; 10(6): 1190–1195. doi:10.1002/pmic.200900567.

Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation datasets

Eric W. Deutsch¹, David Shteynberg¹, Henry Lam², Zhi Sun¹, Jimmy K. Eng³, Christine Carapito⁴, Priska D. von Haller³, Natalie Tasman¹, Luis Mendoza¹, Terry Farrah¹, and Ruedi Aebersold^{1,5,6,7}

¹ Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA ² Department of Chemical and Biomolecular Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong ³ UW Proteomics Resource, University of Washington, 815 Mercer Street, Seattle, WA USA ⁴ Laboratoire de Spectrométrie de Masse Bio-Organique (LSMBO), IPHC-DSA, University of Strasbourg UDS, CNRS UMR 7178, Strasbourg, France ⁵ Institute of Molecular Systems Biology, ETH Zurich, CH-8049 Zurich, Switzerland ⁶ Faculty of Science, University of Zurich, CH-8049 Zurich, Switzerland ⁷ Center for Systems Physiology and Metabolic Diseases, Zurich Switzerland

Abstract

Electron transfer dissociation (ETD) is an alternative fragmentation technique to collision induced dissociation (CID) that has recently become commercially available. ETD has several advantages over CID. It is less prone to fragmenting amino acid side chains, especially those that are modified, thus yielding fragment ion spectra with more uniform peak intensities. Further, precursor ions of longer peptides and higher charge states can be fragmented and identified. However, analysis of ETD spectra has a few important differences that require the optimization of the software packages used for the analysis of CID data, or the development of specialized tools. We have adapted the Trans-Proteomic Pipeline (TPP) to process ETD data. Specifically, we have added support for fragment ion spectra from high charge precursors, compatibility with charge-state estimation algorithms, provisions for the use of the Lys-C protease, capabilities for ETD spectrum library building, and updates to the data formats to differentiate CID and ETD spectra. We show the results of processing datasets from several different types of ETD instruments and demonstrate that application of the ETD-enhanced TPP can increase the number of spectrum identifications at a fixed false discovery rate by as much as 100% over native output from a single sequence search engine.

Keywords

shotgun proteomics; electron-transfer dissociation; bioinformatics

Introduction

Tandem mass spectrometry (MS/MS) has enabled the identification of large numbers of proteins from complex biological samples. Instrumentation continues to advance, allowing the acquisition of increasing numbers of spectra per run at greater sensitivity. The software tools used to analyze such large datasets have also become increasingly more sophisticated[1]. Simple techniques of applying cutoffs to native search scores have given way to modeling of output scores and other attributes of the peptide identifications to yield improved probabilistic metrics for peptide and protein identifications[2–4].

A recent development in mass spectrometry instrumentation has been the introduction of a new method of fragmenting peptides to generate MS/MS spectra: electron transfer dissociation (ETD)[5–7]. ETD has several advantages over collision induced dissociation (CID), namely that 1) amino acid side chains and modifications are more likely to be left intact, 2) the fragment ion spectra of peptides yield more uniform peak intensities, and 3) precursor ions of longer peptides and higher charge states can be fragmented and identified. It has been shown to be particularly useful in studies of post-translational modifications[8,9], where the ETD fragmentation spectra often are more complex and amenable to identification than their CID counterparts. However, software tools optimized for processing ETD datasets are just beginning to emerge[10], and existing tools for CID data processing are suboptimal for ETD data. In general, the field of ETD-capable software has yet to reach the maturity of the field of CID-capable software.

The Trans-Proteomic Pipeline (TPP)[2] is an open source suite of software tools that supports the analysis of MS/MS data. The TPP includes software tools for MS data representation, MS data visualization, peptide identification and validation, protein inference, quantification, spectral library building and searching, and biological inference. Until now the TPP has only supported the analysis of CID data.

Here we introduce new functionality in the TPP to extend its utility to the analysis of ETD data. We analyzed the performance of the ETD-compatible TPP on search results generated by the search algorithms OMSSA[11] and SEQUEST[12]. The features introduced here are available in version 4.3 of the TPP software suite. In the following sections we describe the features added to the TPP and show the results of processing three different datasets through the tools.

Support for fragment ion spectra from higher charge state precursors

For CID LC-MS/MS analysis of tryptic digests, there are relatively few interpretable spectra from precursor ions with charge greater than 3+. However, for ETD analysis where alternate enzymes are used, peptides with higher precursor ion charges are more likely to be identified, often with 50% of interpretable spectra having charge greater than 3+. To accommodate the analysis of such spectra PeptideProphet has been enhanced to build models for an arbitrary number of charge states and now by default attempts to create models for charge states up to 7+ for ETD datasets.

In cases where there are several reasonable interpretations of the same spectrum for different possible charge states, PeptideProphet now is able to apportion the probabilities of the various interpretations so that the sum of probabilities for the one spectrum is always less than or equal to 1.0.

Charge State Prediction

For CID data from a low resolution instrument such as the frequently used linear ion trap, it is usually not possible to know the charge state of the precursor ions in advance (except for 1+ spectra, which are relatively easy to distinguish). This is not too troublesome with CID spectra since spectra of peptide ions with charge state greater than 3+ are rarely identifiable and thus searches are commonly performed on both 2+ and 3+ charge states. However, since ETD data commonly have charge states greater than 3+, it becomes impractical to search all the possibilities.

To address this, several charge estimation algorithms have emerged, such as Charger[13] and the Charge Prediction Machine (CPM)[14]. Unfortunately, the licenses of the Charger and CPM programs do not allow bundling with the TPP. We have therefore implemented

tools in the TPP that allow users to use the results generated by Charger and CPM with the TPP. The workflow is to convert the mzXML[15] files to ms2 format files with the MzXML2Search tool, run the CPM command-line program, transform the results into simple text format with the TPP createChargeFile tool, and then finally update the mzXML files with the TPP mergeCharges tool. This workflow results in mzXML files with updated charge state information, either with the precursorCharge attribute when a single charge is estimated with reasonable certainty, or with the possibleCharges attribute in cases where several charge states are implicated. The new mzML format[16] already supports both of these concepts.

Spectral Library Searching

As with CID spectra, we expect that a spectral searching approach to peptide identification will also improve the identification of ETD spectra compared to sequence database searching alone[17]. To support spectral library searching of ETD data, we adapted SpectraST[17,18], a spectral library building and searching tool packaged in the TPP, to accommodate ETD spectra. The peak annotation algorithm now expects c- and z[•]-type ions in addition to b- and y-type ions, and the activation method is recorded so that CID and ETD spectra can co-exist in the same spectral library.

The dominant peaks in ETD MS/MS spectra are often the charge-reduced precursor (CRP) peaks, which are not useful for peptide identification (although they are useful for charge-state determination). Therefore, it is advantageous for algorithms to remove these peaks before searching each spectrum[19]. OMSSA includes this functionality as does the SEQUEST version 27 build UW20090611 used here. We adapted the SpectraST tool to remove CRP peaks from both the library and query spectra prior to spectral matching.

Other Changes

A few other minor changes were made to the TPP to improve its handling of ETD data. Specifically, the spectrum display components now annotate spectra with c- and z[•]-type ions and display the corresponding fragmentation tables by default when viewing ETD spectra (Figure 1); the mzXML and pepXML file formats now include attributes for the activation method, so that software can properly handle both CID and ETD spectra appropriately (handling for this is already present in mzML[16]); and peptides generated by the Lys-C protease that is popular with ETD experiments due to the increased peptide length compared to trypsin are now handled properly. Finally, the TPP graphical user interface Petunia was updated to enable users to run createChargeFile and mergeCharges.

Application to Sample Datasets

To demonstrate the significant improvement in search results achieved by the ETD-adapted TPP, we applied the tools to three datasets and compared the TPP results with the native search output. All three datasets were derived from yeast whole-cell lysate samples digested with Lys-C. Dataset 1 was acquired on a custom LTQ Orbitrap (Thermo-Fisher) outfitted with ETD capability[20,21] by the Coon Lab and has been previously described[22]. Datasets 2 and 3 were acquired on ordinary LTQ ETD (Thermo-Fisher) instruments in different labs. All three datasets are described in detail in the Supplementary Materials section.

The native vendor format RAW files were converted to mzXML[15] using the TPP tool ReAdW. Since dataset 1 was acquired on an LTQ Orbitrap, the precursor m/z values and charge states were known with quite high accuracy. In contrast, datasets 2 and 3 were acquired on the currently more common LTQ ETD instruments and we used the CPM tool

with the TPP tools to store estimated precursor charge values in the mzXML files using the workflow described above.

The protein list (database) used for searching was a non-redundant merge of several sources, including SGD, Ensembl, and NCI, plus the cRAP list of contaminants from GPM (<http://www.thegpm.org/cRAP/index.html>). In order to assess the performance of the software, decoy sequences were appended to the target sequences. The decoys were generated by taking all the target sequences, fixing the positions of all the arginine and lysine amino acids and any prolines that immediately follow them, and scrambling the amino acids between these anchor residues. The identifiers of the decoy sequences were the original protein names prepended with DECOY1_ and DECOY2_ interleaved. Thus there were as many decoy sequences as target sequences, but the decoys could be easily split into two populations for modeling and testing. This file is available with the data package as described in the Supplementary Materials section.

After conversion to mzXML and updating the precursor ion charge states, the data were searched with both the SEQUEST and OMSSA sequence database search engines as described in the Supplementary Materials section. The results of these searches were post-processed and validated with the tools in the TPP as summarized in Figure 2. The PeptideProphet[23] tool was used to develop a mixture model of correct and incorrect identifications in the search results and to assign a probability of being correct for each peptide spectrum match (PSM).

The PeptideProphet analysis was performed with the accurate mass and retention time models enabled. Validation of the SEQUEST results was performed with the standard parametric model. However, the OMSSA search results were processed with the semi-parametric model[24] using the DECOY1_ matches.

The iProphet[25] analysis followed the PeptideProphet analysis to improve the discrimination between correct and incorrect assignments. This was accomplished by developing models of how to adjust the probabilities based on the corroborating evidence of the number of sibling search results, the number of replicate spectra, the number of sibling ions, and the number of sibling modifications. Probabilities for distinct peptide sequences were calculated and the individual PSM probabilities corrected according to the models.

Finally, ProteinProphet[26] was used to derive protein-level probabilities by combining the multiple observations of peptides that map to each protein. Peptides that map to multiple proteins were apportioned to result in the simplest list of proteins that can explain the observations. ProteinProphet was run in the mode designed to follow iProphet.

The processed search results of the three datasets were then used to build a consensus spectral library using the updated version of SpectraST. During library building, whenever available, replicate spectra from the same peptide ion are merged to form a consensus spectrum, hence improving the quality of the reference spectrum. We then demonstrated the feasibility of spectral searching of ETD data by re-searching the same three datasets using SpectraST. A substantial number of largely lower-quality spectra that failed to be identified by sequence searching were identified by this strategy (see Supplementary Materials section).

The results of the analysis of these three datasets are shown in the three panels of Figure 3. Each panel shows the number of PSMs as a function of the decoy-estimated false discovery rate (FDR). In each case, we show 6 curves, one each for: SEQUEST search alone, SEQUEST search with PeptideProphet+iProphet results; OMSSA search alone, OMSSA search with PeptideProphet+iProphet results; both searches combined with iProphet;

SpectraST search with PeptideProphet+iProphet validation based on the ETD spectrum libraries created from the three datasets.

For each of the datasets, application of the TPP modeling increased the number of identified spectra at 1% FDR by 35 – 70% for SEQUEST native scores, and by 10 – 25% for OMSSA native expect scores. Further, the combining of search results from both engines typically yielded an additional 10 – 25% gain in identified PSMs at an FDR of 1% over a single search engine validated with the TPP. Finally, a spectrum library reanalysis with SpectraST added a 5 – 10% gain in identified PSMs at an FDR of 1% over the combined iProphet results. For these three datasets derived from different yeast whole-cell lysate samples, we found the relative performance of the search engines and the improvement in the results when applying the TPP tools quite similar, even between the samples run on the LTQ Orbitrap ETD (Dataset 1) and the LTQ ETD instruments (Datasets 2 and 3).

Conclusion

We described the optimization of several modules of the TPP for the analysis of fragment ion spectra generated by ETD. We demonstrated that the implemented incremental improvements contributed to a significantly enhanced performance of the system for ETD spectrum-to-peptide matching compared to the performance of commonly used database search systems. Further, we demonstrated that combining the results of both OMSSA and SEQUEST searches with the iProphet tool yielded a marked improvement in identification rates since each algorithm excels at identifying somewhat different subsets of spectra and the iProphet tool can extract the best identifications from each. The combination of two search engines via TPP yielded a 10 – 25% increase in PSMs at a constant 1% FDR over one search engine alone. Further, library searching of these datasets identified an additional 5 – 10% PSMs. These results are consistent with the results achieved by searching CID data with the conventional form of the TPP since the TPP tools take advantage of information that is not apparent from the search scores themselves.

The TPP tools are all free and open source, and have become easier to install and use in the past few years. For the Windows platform, there is an install link available at the Seattle Proteome Center web site (<http://www.proteomecenter.org>). The TPP is also compatible with Linux and MacOS X, although the installation for those platforms is somewhat more complex. However, the entire process is sufficiently simple that all users can easily take advantage of the advanced tools in the TPP to analyze their data[27].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank D.L. Swaney, J.J. Coon, G.C. McAlister and collaborators for making their datasets public via PeptideAtlas, thereby enabling works such as this. This work has been funded in part with Federal funds from the National Heart, Lung and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179, from the University of Washington's Proteomics Resource (UWPR95794), and from PM50 GMO76547/Center for Systems Biology.

References

1. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007;4(10):787–97. [PubMed: 17901868]
2. Keller A, et al. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 2005;1:2005 0017. [PubMed: 16729052]

3. Kall L, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 2007;4(11):923–5. [PubMed: 17952086]
4. Ma ZQ, et al. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J Proteome Res.* 2009
5. Coon JJ, et al. Electron transfer dissociation of peptide anions. *J Am Soc Mass Spectrom* 2005;16(6):880–2. [PubMed: 15907703]
6. Syka JE, et al. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 2004;101(26):9528–33. [PubMed: 15210983]
7. Good DM, et al. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* 2007;6(11):1942–51. [PubMed: 17673454]
8. Domon B, et al. Electron transfer dissociation in conjunction with collision activation to investigate the *Drosophila melanogaster* phosphoproteome. *J Proteome Res* 2009;8(6):2633–9. [PubMed: 19435317]
9. Swaney DL, et al. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* 2009;106(4):995–1000. [PubMed: 19144917]
10. Sadygov RG, et al. A new probabilistic database search algorithm for ETD spectra. *J Proteome Res* 2009;8(6):3198–205. [PubMed: 19354237]
11. Geer LY, et al. Open mass spectrometry search algorithm. *J Proteome Res* 2004;3(5):958–64. [PubMed: 15473683]
12. Eng J, McCormack AL, Yates JR. 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.
13. Sadygov RG, Hao Z, Huhmer AF. Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Anal Chem* 2008;80(2):376–86. [PubMed: 18081262]
14. Carvalho PC, et al. Charge Prediction Machine: Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra. *Anal Chem.* 2009
15. Pedrioli PG, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004;22(11):1459–66. [PubMed: 15529173]
16. Deutsch E. mzML: a single, unifying data format for mass spectrometer output. *Proteomics* 2008;8(14):2776–7. [PubMed: 18655045]
17. Lam H, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007;7(5):655–67. [PubMed: 17295354]
18. Lam H, et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods* 2008;5(10):873–5. [PubMed: 18806791]
19. Good DM, et al. Post-Acquisition ETD Spectral Processing for Increased Peptide Identifications. *J Am Soc Mass Spectrom* 2009;20(8):1435–40. [PubMed: 19362853]
20. McAlister GC, et al. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal Chem* 2007;79(10):3525–34. [PubMed: 17441688]
21. McAlister GC, et al. A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *J Proteome Res* 2008;7(8):3127–36. [PubMed: 18613715]
22. Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 2008;5(11):959–64. [PubMed: 18931669]
23. Keller A, et al. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74:5383–5392. [PubMed: 12403597]
24. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J Proteome Res* 2008;7(1):286–92. [PubMed: 18078310]
25. Shteynberg D, et al. Postprocessing and validation of tandem mass spectrometry datasets improved by iProphet. in preparation.
26. Nesvizhskii AI, et al. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75:4646–4658. [PubMed: 14632076]

27. Deutsch EW, et al. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics*. (this issue) accepted.
28. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 2008;9(5):429–34. [PubMed: 18451766]
29. Falkner JA, Andrews PC. Tranche: Secure Decentralized Data Storage for the Proteomics Community. *Journal of Biomolecular Techniques* 2007;18(1):3.

Figure 1.

Example screenshot of a spectrum from Dataset 1 in the TPP interface for 4+ ALPIRRDDEVLVVRGSK. Each identified peak is labeled with its ion series, series number, and charge identification. Charge reduced precursor peaks are labeled with the prefix M++++. On the left are some user-settable parameters for the spectrum display. On the right, the m/z values for expected c, z*, and y series ions are listed, and shaded if detected in the spectrum.

Figure 2.

Schematic overview of the workflow of the post-processing and validation of the results of the SEQUEST and OMSSA sequence database searches using the TPP tools. PeptideProphet was used to model and validate the peptide spectrum matches (PSMs). The iProphet tool was used to coalesce the results to the distinct peptide sequence level. ProteinProphet was used to infer the identified proteins and assign probabilities to each protein based on the upstream analysis.

Figure 3.

Number of PSMs as a function of FDR for the SEQUEST search before (red dotted) and after TPP processing (red solid), OMSSA search before (blue dotted) and after TPP processing (blue solid), combined (orange), and SpectraST spectral library search (black) for the three datasets. All three datasets are from different yeast whole-cell lysate samples; Dataset 1 was obtained with an LTQ Orbitrap ETD, while Datasets 2 and 3 were obtained with LTQ ETD instruments in different labs. In all three cases, applying two search engines and the TPP tools nearly doubled the number of identified spectra at a 1% FDR over using SEQUEST raw scores.