DtaRefinery: A Software Tool for Eliminating Systematic Errors from Parent Ion Mass Measurements in Tandem Mass Spectra Datasets

Vladislav A. Petyuk, Anoop M. Mayampurath, Matthew E. Monroe, Ashoka D. Polpitiya, Samuel O. Purvine, Gordon A. Anderson, David G. Camp II, and Richard D. Smith
Pacific Northwest National Laboratory, Richland, WA

Pacific Northwest NATIONAL LABORATORY

Overview

- An algorithm and software tool, DtaRefinery, has been developed that significantly reduces and practically eliminates systematic mass measurement error in parent ions in MS/MS datasets.
- By fitting a regression model the tool can estimate the systematic MME and then correct parent ion mass entries by removing the estimated systematic error components.

Introduction

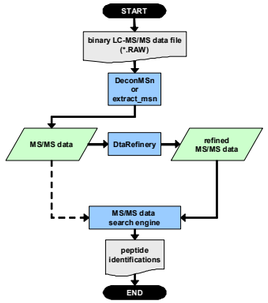
Hybrid two-stage mass spectrometers capable of both highly accurate mass measurement and high throughput MS/MS fragmentation have become widely available in recent years, allowing for significantly better discrimination between true and false MS/MS peptide identifications by the application of a relatively narrow window for maximum allowable deviations of measured parent ion masses. To fully gain the advantage of highly accurate parent ion mass measurements, it is important to limit systematic mass measurement errors.

Based on our previous studies of systematic biases in mass measurement errors, here, we have designed an algorithm and software tool that eliminates the systematic errors from the peptide ion masses in MS/MS data. It is demonstrated that the elimination of the systematic mass measurement errors allows for the use of tighter criteria on the deviation of measured mass from theoretical monoisotopic peptide mass, resulting in a reduction of both false discovery and false negative rates of peptide identification.

A software implementation of this algorithm (called DtaRefinery) reads a set of fragmentation spectra, searches for MS/MS peptide identifications in a FASTA file containing expected protein sequences, fits a regression model that can estimate systematic errors, and then corrects the parent ion mass entries by removing the estimated systematic error components. The output is a new file with fragmentation spectra with updated parent ion masses. The software is freely available at http://omics.pnl.gov/software/.

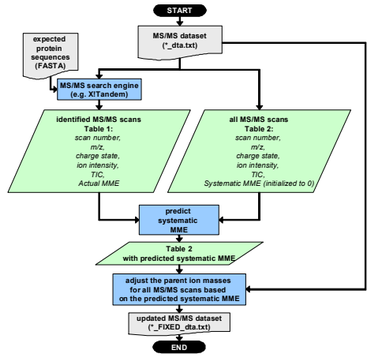
Methods

Updates to the MS/MS processing pipeline



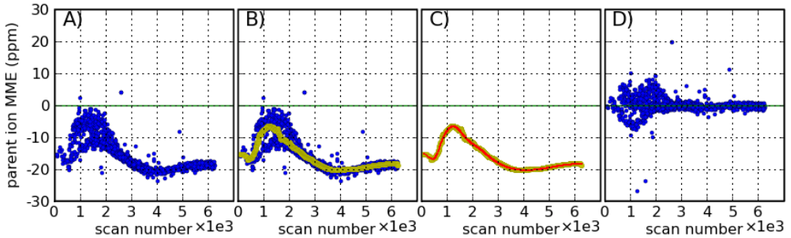
MS/MS data are initially extracted from a binary file with either DeconMSn or extract_msn. The extracted data are processed with DtaRefinery or alternatively, can be directly used for searching peptide identifications. The format for the refined data produced by DtaRefinery is the same as for data originally extracted by DeconMSn or extract_msn. Finally, the refined MS/MS data are submitted for peptide identifications by commonly used search engines.

DtaRefinery process flowchart



The dataset is analyzed by an MS/MS search engine against the expected list of protein sequences. Spectra with identified peptides goes into the "Table 1", which will be further used for training a regression model predicting systematic MME. "Table 2", which is used to store the predicted systematic MME, contains all spectra regardless if they have assigned peptides or not. After the model is trained, the parent ion masses in the original dataset are corrected based on the predicted systematic MME that are stored in "Table 2" and written into an updated MS/MS data file ("*_FIXED_dta.txt").

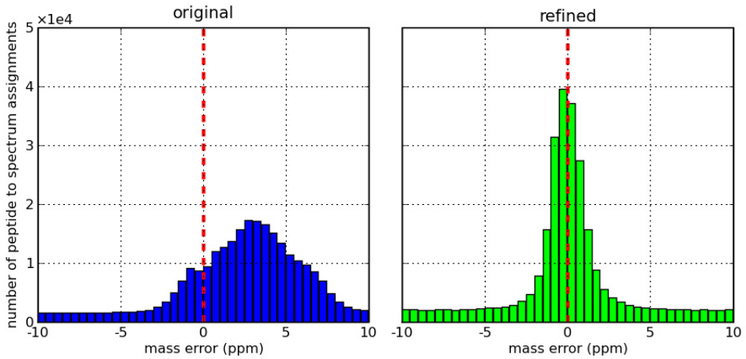
Example of correcting highly pronounced systematic parent ion MME



The example is an actual LC-MS/MS analysis on a not up-to-date calibrated LTQ Orbitrap instrument with significant sample overloading and outdated calibration. Because of sample overloading, the automatic gain control system was not capable of properly modulating the ion population within the Orbitrap cell resulting in space charge effects that cause noticeable systematic MME. However, after applying DtaRefinery and subtracting the systematic MME components predicted by the regression models trained in the space of all four parameters (scan number, *m/z*, log10 of ion intensity and TIC), the mean of the MME distribution shifts from -16 ppm to practically 0 ppm and the standard deviation contracts from 4.3 ppm to 0.8 ppm (data not shown). **A)** The original parent ion MME plotted as a function of scan number (blue circles). **B)** Smoothing the MME residuals with Tukey's running median (**yellow** circles). **C)** Fitting a spline function into smoothed data to have a continuous function for prediction of systematic MME (**red** line). **D)** Corrected parent ion MME by subtracting the systematic MME predicted by the model trained using only the scan number parameter.

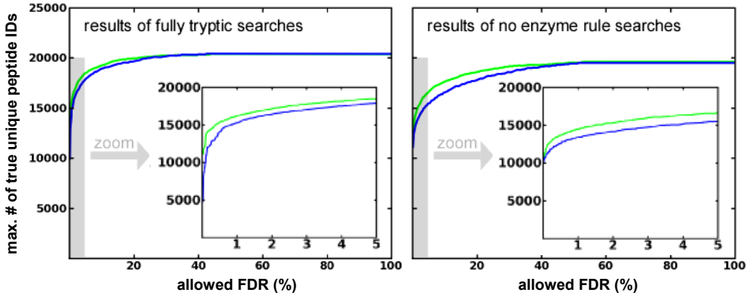
Results

Mass measurement error improvements



Mass error distribution histograms for all peptide to spectrum assignments with 2+ charge produced by SEQUEST searches for fully tryptic peptides from 72 LC-MS/MS datasets. No XCorr or ΔCn filtering criteria have been applied. The bin width is 0.5 ppm. The DtaRefinery tool had noticeably decreased the width of the MME distribution histogram and thus maximum allowable deviation of the parent ion mass for true peptide identifications from about 10 (left, blue) to about 3 (right, green) ppm.

Improvements in FDR



Estimates of maximum number of true positive unique peptides with 2+ charge that can be identified by SEQUEST within the allowed FDR values. The green curve represents the results of SEQUEST searches of 72 LC-MS/MS datasets preprocessed by DtaRefinery. The results of non-preprocessed datasets are shown in blue. The maximum number of peptide identification was obtained by searching multiple combinations of ΔM, XCorr, and ΔCn parameters within the following ranges: ΔM from 1 to 10, XCorr from 0 to 8, and ΔCn from 0 to 0.4 (288,000 combinations in total). Note that for any given FDR value, the results from refined datasets searches provide more true peptide identifications. It is also true that for a given number of true peptide identifications it is always possible to achieve noticeably lower FDR by preprocessing LC-MS/MS datasets with DtaRefinery.

Conclusions

- DtaRefinery aids in reducing the parent ion MME tolerance up to 10-fold (typically down to +/-2 ppm for hybrid instruments), and thus reduces the number of both false positive and false negative peptide identifications.
- With the increasing use of hybrid MS instrumentation, we anticipate that the DtaRefinery software tool will be widely used in tandem with DeconMSn, in particular for proteomics applications in which peptide identification confidence remains challenging due to a significantly increased search space.
- Applications DtaRefinery can benefit most include identification of peptides with post-translational modifications, identification of peptides resulting from non-specific proteolysis, and searches using exhaustively translated genomes (e.g., in all six reading frames from stop-to-stop codons as a set of putative protein sequences).

Acknowledgements

Samples were analyzed using capabilities developed under the support of the NIH National Center for Research Resources (RR18522) and the U.S. Department of Energy Biological and Environmental Research (DOE/BER).

Significant portions of the work were performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

References

1. Petyuk, V.A. et al. DtaRefinery, a software tool for elimination of systematic errors from parent ion mass measurements in tandem mass spectra data sets. *Mol Cell Proteomics* **9**, 486-496 (2010).
2. Petyuk, V.A. et al. Elimination of systematic mass measurement errors in liquid chromatography-mass spectrometry based proteomics using regression models and a priori partial knowledge of the sample content. *Anal Chem* **80**, 693-706 (2008).

CONTACT: Vlad Petyuk, Ph.D.
Biological Sciences Division, K8-98
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: vladislav.petyuk@pnl.gov