Overview

• An algorithm and software tool, DtaRefinery, has been developed that significantly reduces and practically eliminates systematic mass measurement error in parent ion 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 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 parent ion mass entries by removing the estimated systematic error components.


Pacific Northwest National Laboratory, Richland, WA

Updates to the MS/MS processing pipeline

DtaRefinery process flowchart

Results

Mass measurement error improvements

Improvements in FDR

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 increased use of hybrid instruments, 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

Sample preparation and acquisition were performed in the Environmental Molecular Sciences 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-AC05-76RL01830.

References

2. Petyuk, V.A. et al. DtaRefinery (http://omics.pnl.gov/software/). Pacific Northwest National Laboratory E-mail: vladislav.petyuk@pnl.gov

CONTACT: Vladislav 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

Example of correcting highly pronounced systematic parent ion MME

The example is an actual LC-MS/MS analysis, read in a de-coupled TQD 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, which results in scan charge effects that cause noticeable systematic MME. However, after preprocessing the LC-MS/MS data using DtaRefinery, the systematic MME predicted by the model was removed using the scan parameter.

Table 2: Modifications of maximum number of true positive unique peptides with 2+ charge produced by SEQUEST within allowed FDR values. The green curve represents the results of SEQUEST searches of 72 LC-MS/MS datasets preprocessed by DtaRefinery. The expected number of true peptide identifications is shown as the blue line. For a given number of true peptide identifications, the green curve represents the results of non-preprocessed datasets.

<table>
<thead>
<tr>
<th>Modification</th>
<th>All peptide IDs</th>
<th>False discovery (%) allowed FDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>max. # of true unique peptide IDs</td>
<td>allowed FDR (%)</td>
</tr>
<tr>
<td>Preprocessed</td>
<td>max. # of true unique peptide IDs</td>
<td>allowed FDR (%)</td>
</tr>
</tbody>
</table>

Estimates of maximum number of true positive unique peptides with 2+ charge that can be identified by SEQUEST within allowed FDR values. The green curve represents the results of SEQUEST searches of 72 LC-MS/MS datasets preprocessed by DtaRefinery. The expected number of true peptide identifications is shown as the blue line. For a given number of true peptide identifications, the green curve represents the results of non-preprocessed datasets.

ΔCn from 0 to 0.4 (288,000 combinations in total).

Note that for any given FDR value, the results from refined datasetsad search 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.