Exploring Proteomics Metadata Using Spotfire[1] and a Companion User Interface

Roman M. Ženka; Kenneth L. Johnson; H. Robert Bergen, III
Proteomics Core
Mayo Clinic, Rochester, MN

Abstract

Introduction: The instruments and proteomics software provide large amounts of metadata that can be used to track and improve performance of the pipeline. It is cumbersome to analyze the metadata manually, while automatic processing can provide only limited insight.

Objective: Our goal was to enable in-depth understanding of the data acquisition process for optimization, quality control and troubleshooting. Since the amount of human attention devoted to each experiment is limited, we designed a tool to minimize time needed to make informed decisions. We aimed to maximize the amount of data that gets routinely reviewed to gain better insight into the process and ensure quality outputs.

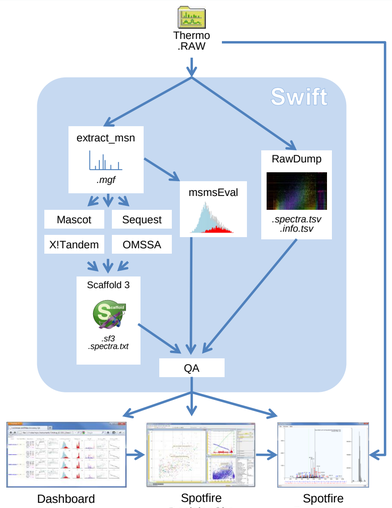
Methods: We have implemented a software tool that combines spectrum level metadata from our LTQ-Orbitrap proteomics pipeline including: the original .RAW file, a peak-picked version, results from a spectrum quality assessment tool (msmsEval[2]), search results (Mascot, Sequest, X!Tandem, Scaffold[3]) and polymer detection (in-house). We utilize our proteomic pipeline "Swift[4]" to automate acquisition of relevant metadata and their alignment. The resulting data are stored in a large tab-delimited file.

Spotfire DecisionSite software is utilized for visualization and analysis of the metadata. Spotfire is a very fast visualization platform, enabling rapid hypothesis forming and testing. Since Spotfire is not proteomics-specific, a specialized plugin was written to provide extra functionality, namely spectrum viewing and annotation.

We developed an application called Spotfire Proteomics that is integrated with Spotfire. While Spotfire provides interactive plots, Spotfire Proteomics adds detailed raw spectrum plots overlaid with the collected metadata (such as ion ladders, polymer peaks, survey spectrum preview). The application caches the extracted spectra for faster viewing and utilizes keyboard shortcuts.

The application is written in Microsoft .NET framework, and is installed and updated automatically over the network using one-click deployment. A Thermo Scientific library provides direct access to data in instrument .RAW files.

Overview of the System



Extracted Metadata

.RAW

• Full path.RAW file • Parent m/z • TIC • RT • MS Level • Parent Scan • Child Scans • Ion Injection Time • Cycle Time • Elapsed Time • Dead Time • Time To Next Scan • Lock Mass Found • Lock Mass Shift • Conversion Parameter I,A-E • Source Current • Vacuum Ion Gauge • Vacuum Convectron Gauge • FT Penning Gauge • FT Pirani Gauge 1 • Multiplier 1 • Multiplier 2 • FT CE Measure Voltage • FT Analyzer Temp

.mgf

• m/Z • Z • file name

Scaffold

• peptide/protein IDs • search engine scores • mass accuracy • NTT • modifications

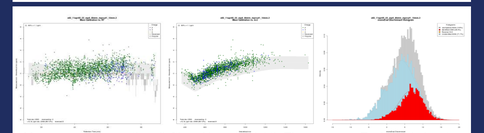
msmsEval

• Npeaks • NormTIC • GoodSegs • IntnRatio1% • IntnRatio20% • Complements • IsoRatio • H2ORatio • AADiffRatio • discriminant • P(+|D) • Z_prob

Swift

• Dissociation type • Polymer segment • Polymer offset • Polymer score

Dashboard



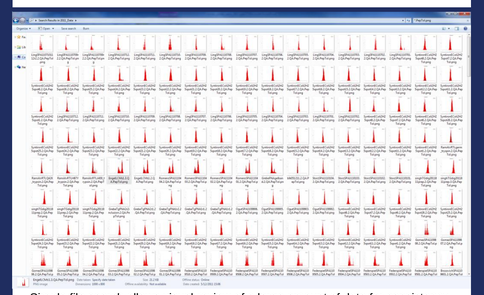
Mass accuracy over time (left) and by m/z (right). The red/gray line on the left side depicts lockmass correction (red - lockmass not found). The gray area is a 2D LOESS fit containing 95% of peptide identifications. The check-mark is a typical calibration pattern for one of our instruments.

msmsEval discriminant score plot. The identified spectra are in red. Pinpoints high-quality spectra that were not identified.

Mass accuracy histogram. The darker portion denotes spectra acquired with lockmass found.

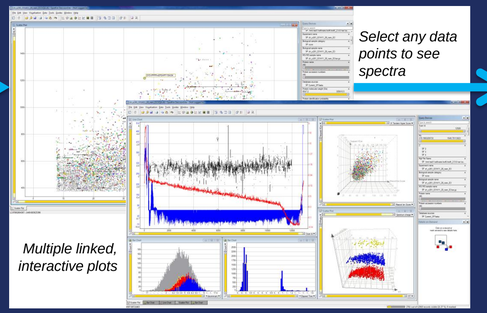
Oversampling errors (one peptide being identified several times in a row) result in horizontal lines in this plot. Polymers are marked as well. Unidentified spectra show as dots.

Source current over RT allows visual differentiation of methods, detecting chromatography problems and ion source instability.



Simple file search allows visual review of a large amount of data for consistency.

Spotfire DecisionSite



Discussion

Our experience has shown that our tool speeds up the review process. Users can quickly focus on the spectra of interest, and then rapidly review each one at rates of up to 1000 spectra per minute (when scrolling through the list of spectra to spot an anomaly). The tool is fast and easy to use, encouraging exploration.

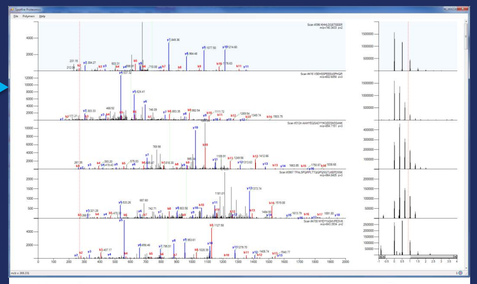
The simplified static dashboard was the most successful component due to its low cognitive demands. Our technicians habitually review the dashboard for every run.

The main disadvantage of the interactive Spotfire visualization is the slow startup time of Spotfire (20+ seconds per file). This reduces the amount of interactive analyses that can be performed in given amount of time.

Conclusions

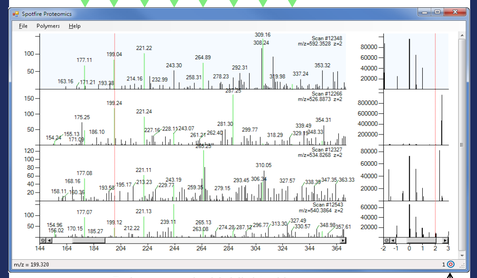
- the tool has been integrated into our daily operations
- technicians developed a habit of checking the static dashboard for each experiment, even if they did not study the data in detail
- new ways of utilizing the tool are still being discovered whenever a new problem arises
- most often used for troubleshooting
- instrumentation issues made more visible
- helped to quantify anecdotal observations

Spotfire Proteomics

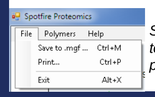


Up to 5 zoomable spectra (rapidly mouse wheel through larger list)

Precursor preview



Polymer peaks highlighted in green (realtime polymer detection on displayed spectra)



Save selected spectra to .mgf for further processing

Multiple Spotfire connections for data comparison

References

1. http://spotfire.tibco.com/products/decisionsite.aspx

2. Jason WH Wong, Matthew J Sullivan, Hugh M Cartwright and Gerard Cagney. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics* 2007,8:51doi:10.1186/1471-2105-8-51

3. http://www.proteomesoftware.com/

4. http://goo.gl/aYJkJ - shortened version of http://informatics.mayo.edu/svn/trunk/mprc/swift/index.html