A System for Creating, Viewing, and Editing Precursor Mass Spectrometry Ground Truth Data

Introduction: Mass spectrometry (MS) is a powerful vector for the analysis of molecular components (such as proteins, peptides, lipids, and metabolites) in biological samples across a broad range of applications [1]. MS experiments generate datasets consisting of millions of 3-D points consisting of mass-to-charge (m/z), retention time (RT), and intensity. MS experiments require the mapping of all or some of these points to signal groups that correspond to a single (or multiple, in the case of isomers) molecule at a given charge state.

This process, called feature detection, has been addressed by numerous algorithms, commercial software, and public software such as MaxQuant [2], MZMine 2 [3], CentWave (XCMS) [4], MatchedFilter (XCMS) [4], and Massifquant (XCMS) [5]. Unfortunately, many of these and other algorithms for MS1-aware analysis have not been quantitatively evaluated [6], mostly due to the fact that ground truth data is very difficult to generate.

A system for producing precursor ground truth annotations requires several functions:

- It must parse, load, store, and retrieve precursor data.
- It must efficiently display many points on the screen.
- It must display points in representative subsets, as not all points can be rendered on the screen at once.
- It must output the data in easy-to-port formats.
- It must provide the user with efficient navigation of the data (zoom and shifting to the right, left, up, or down).

It should also be designed in such a way as to allow easy cross-platform install without the need for excessive dependencies or onerous compilation. With such a software, ground truth datasets can be created and used to evaluate many aspects of mass spectrometry data processing, including precursor mapping and signal extraction algorithms.

Proposed Solution: I present JS-MS, a software suite that provides a dependency-free, browser-based, one click, cross-platform solution for creating precursor ground truth. The software retains the first versions capacity for loading, viewing, and navigating MS1 data in 2- and 3-D, and adds tools for capturing, editing, saving and viewing isotopic envelope and extracted isotopic chromatogram features. The software can also be used to view and explore the results of feature finding algorithms.

Development and Validation: To demonstrate the efficacy of JS-MS, we created the first fully annotated quantitative ground truth MS1 dataset. This dataset is an untargeted protein identification sample consisting of 48 Universal Proteomics Standard 2 (UPS2) proteins. Isotopic envelopes are a unique 3-d signal group and are created for each type of ionized molecule. Each envelope is also made up of one or more extracted ion chromatograms (XICs). XICs typically represent a roughly Gaussian shape. Therefore, to annotate the UPS2 data set, we segmented all signals into XICs or noise. To group points into XICs, JS-MS allows users to draw a rectangle over the desired XIC area. The points within the area will then be highlighted to clearly show they have

been categorized. Variations in peak intensity within XICs was allowed if it did not violate an on-average Gaussian shape. XICs were then clustered into isotopic envelopes, by selecting the XICs with the mouse in envelope mode, where valid envelopes have a m/z distance of 1/n where n is a real positive integer. Isotopic envelopes are then highlighted a different color to show their categorization. Overlapping XICs were grouped together by comparing intensity apex in retention time, m/z distance, and similarity of intensity distributions.

After segmentation was done, a worklist of each isotopic envelope was created for validation. The annotation of this dataset required more than 1,000 hours of manual curation and consists of more than 62 million points. Over 1.2 million of these points have been grouped in 57,518 XICs, and those were grouped into over 14,000 isotopic envelopes.

The UPS2 dataset can be used to quantify many evaluations such as XIC extraction algorithms, XIC clustering into isotopic envelopes, MS1-based quantification methods, MS2 quantification methods, and false detection estimations. JS-MS will allow for the creation and validation of more ground truth datasets that will assist further evaluation and algorithm creation for mass spectrometry data.

Intellectual Merit: I am very well acquainted with the details and properties of mass spectrometry data. I have worked extensively in the Smith lab over the past three years using software to visualize and interact with hundreds of thousands of signals from MS1 data. The Smith lab has had substantial success in integrating traditional computer science solutions with domain specific properties from mass spectrometry to exploit computational and analytical efficiency beyond general applications.

Broader Impacts: The quantitative ground truth dataset, UPS2, is designed to advance the capability of quantitative evaluations in mass spectrometry data processing. The creation of more benchmark datasets for precursor-aware mass spectrometry algorithms with JS-MS will enable a new workflow for precursor MS algorithm evaluation that includes quantitative evaluation. New algorithms can be designed using information derived from ground truth annotations created with JS-MS. Once implemented, their performance can be evaluated in terms of, for instance, m/z accuracy of traces annotated, to demonstrate clear improvement over existing algorithms. These evaluations will demonstrate strengths and weaknesses to reviewers and users alike.

References

[1] Cole, R.B.: Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation, and Applications. Wiley, N.p. (1997)

[2] Cox, J., Mann, M.: MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26(12), 13671372 (2008)

[3] Pluskal, T., Castillo, S., Villar-Briones, A., Oresic, M.: Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11(1), 395 (2010)

[4] Tautenhahn, R., Bottcher, C., Neumann, S.: Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9(1), 504 (2008). doi:10.1186/1471-2105-9-504

[5] Conley, C.J., Smith, R., Torgrip, R.J., Taylor, R.M., Tautenhahn, R., Prince, J.T.: Massifquant: open-source Kalman filter based XC-MS isotope trace feature detection. *Bioinformatics* 30(18), 359 (2014)

[6] Smith, R., Ventura, D., Prince, J.T.: Novel algorithms and the benefits of comparative validation. *Bioinformatics* 29(12), 15831585 (2013)