2013-08-21

Annotated Chromatographic Isotope Features from a Highly Complex Mass Spectrometry Proteomic Dataset (MOUSE) for Feature Detection Algorithm Evaluation

John T. Prince
Brigham Young University - Provo, jtprince@chem.byu.edu

Christoper Conley

Guide to MOUSE annotation data

Annotated chromatographic isotope features from a highly complex mass spectrometry proteomic dataset (MOUSE) for feature detection algorithm evaluation

1. Purpose and Contents

These files contain human annotation of features in one section of a complex proteomic mass spectrometry run ("11_110506125210"). A feature is considered the chromatographic signal from a single isotope of a single peptide at a particular charge state. The set also includes annotation of which features belong together as isotopes.

This data set and accompanying annotations, referred to as the MOUSE data set, were used for feature detection algorithm evaluation (FDAE) as described in "Massifquant: open-source Kalman filter based XC-MS feature detection". *Bioinformatics*. 2013. (additional citation details available after publication). Existing and future feature detection and quantitation methods designed for LC-MS are welcome to make use of the data. Since the annotation has isotopic information, users are encouraged to develop other meaningful algorithm evaluations at the isotopic-level. Finally, it serves as an example to be improved upon for future manual-annotation efforts undertaken by the XC-MS community.

directory	description
bonafideFeatures/	reliable features for FDAE
questionableFeatures/	unreliable features unfit for FDAE
dataSourceFile/	subset of experimental data for MOUSE annotation
questionableFeatures/finalQCFilteredOut/	features excluded based on several QC checks

Experimental description. This data set was one fraction of a larger mouse brain phosphoproteomic analysis. 408.8 mg of brain tissue was homogenized/boiled in SDS-lysis buffer and clarified. Proteins were digested and peptides purified following the FASP protocol to yield an estimated 7.3 mg of peptides. 25 mg of Titanspere TiO2 beads (GL Sciences) were used to enrich for phosphorylated peptides. 3M Empore Anion Exchange disks were packed into a 200 l pipette and Britton & Robinson buffer was used to elute at pH 11 (the fraction represented by this data set), 6, 5, 4, and 2. MS analysis was perfomed on an

LTQ-Orbitrap XL fed by an Eksigent NanoLC UHPLC system. A Nano Acquity (1.7m, 130 C18 bead BEH, 75m m x 150mm) column run at 375 nL/min in a linear gradient from 2.5% to 10% ACN (with water and 0.1% formic acid as the second buffer) for 60 minutes, then to 28% ACN for an additional 220 minutes. The tune file, method file and all log files are included. The entire raw file (∼1.5GB) is available upon request. The relevant parameters are: MS1 data collected between 375–1800 m/z at 60,000 resolution with an MS/MS data dependent scan collected after each MS scan. The section chosen for hand-annotation generally spans retention time 5429.5–7306.2 seconds and 600.0003–637.3923 m/z. In total, this area contained 589 annotated features which show variation in length, shape, and variance.

Annotation guidelines. We generated and followed these guidelines for feature annotation:

To be called a chromatographic peak, a series of centroids should typically exhibit the following properties:

Within
(1) The m/z error variance structure is influenced by intensity. Toward the tails of a feature, the m/z observations show mostly symmetric and expanding deviations from the mean. The body and apex centroids deviate less. From a bird's eye view (i.e., looking down the intensity axis), the m/z-time projection has the shape of a string fraying at the edges.
(2) The collective centroids should have a chromatographic peak shape. Dramatic oscillations in intensity from scan to scan could disqualify an annotation.

Between
(1) The detected features should have approximately the same m/z ppm variance.
(2) Within an an isotopic envelope, features should have very similar mode and shape, although length typically varies.

In each case, great effort was made to balance the benefits of the systematic application of these rules with human judgement. Each feature was individually annotated (based on all criteria) and then wrapped into appropriate isotopic distributions.

We used Topp-View for annotation as follows: From a global 2-D view, the annotator identified mass traces satisfying mentioned properties. After zooming, a 3-D inspection confirmed similar chromatographic length and shape for a given isotopic distribution. Once confirmed, the feature was saved as its own .mzML file. Candidate mass traces that did not sufficiently satisfy all the criteria, but still had some resemblance to a feature, were labeled as questionable and saved as .mzML files; these were excluded from the algorithm performance analysis since they were deemed liable to interfere with true algorithmic specificity and sensitivity. Objectively determining a peak's chromatographic boundaries is difficult, especially since there is so much diversity among peak shape and length. Generally, we tried to include as much of each peak tail as possible and to be as consistent as possible across each data set.

A few caveats. For those interested in gleaning isotopic information, note that different isotopic envelopes may have some features included in the FDAE, while others may have been excluded. Thus, these particular isotopes fall into two different directories. Also some

features included in the questionableFeatures/ directory may not be found in the MOUSE experimental data because the original boundaries of annotation were redefined to improve FDAE quality.

2. Annotated Features File Description

The annotated features have file names with a particular structure and can be found in bonafideFeatures/ and questionableFeatures directories..

Structure:
f<intensity size> < feature id > 'dot'< isotopic peak index > < distance between isotopes > <optional comment on peak quality> 'dot' mzML

Explanation of structure:
<intensity size> : (no letter, 'm', 's')
If there is no letter 'm' or 's' following the 'f' (e.g. f24.2.2.mzML) then these are features corresponding to very large intensity isotopes that was part of a first pass in the annotation. The letter 'm' corresponds to 'medium' sized isotopes, where this is somewhat of a subjective classification with no scientific cutoff. The letter 's' is for small intensity isotopes.

<feature id>: (1,2,3,4,....)
These give the isotope/feature an identification. For instance, the collection of features with id '6' corespond to files: { f6.2.3.mzML, f6.3.3.mzML , f6.4.3.mzML, f6.end.3.mzML, f6.start.3.mzML } and is the 6th isotope of the largest intensity class.

< isotopic peak index >: (start,2,3,4,....,end)
The key word *start* and sometimes abbreviated as simply 's' is the first feature in the isotope (starting left to right in the m/z axis); *2* is the second feature in the isotope (starting left to right in the m/z axis) . . . ; and the key word *end* , sometimes abbreviated as simply 'e' is the last feature.

< distance between isotopes > (1,2,3,4,5,6,7)
1 => distance of 1 m/z between features; 2 => distance of 1/2 m/z b/w features; etc.

<optional comment on peak quality> : (c,chroma,etc) Sometimes the chromatography is poor or some other note is appended.

Example
f6.4.3.mzML, is the isotope with feature id 6, and is the 4th feature left to right in the m/z axis, and all features are spaced 1/3 m/z apart.

A caveat. Some of the features that had no identifiable isotope were documented as simply f<intensity size> < feature id > 'dot' mzML. For example, the feature bonafideFeatures/fs339.mzML had no isotopic counterpart features.

3. References

Please cite "Massifquant: open-source Kalman filter based XC-MS feature detection". *Bioinformatics*. 2013 (additional citation details available after publication).