New Algorithms for Reducing the Rate of False Positive and False Negative Compounds Detected From Mass Spectrometry Metabolomics Data

Owen Myers[1], Susan Sumner[2], Shuzhao Li[3], Stephen Barnes[4], Xiuxia Du[1]

[1]University of North Carolina at Charlotte, [2]University of North Carolina at Chapel Hill, [3]Emory University, [4]University of Alabama at Birmingham

Overview

- We have developed new extracted ion chromatogram (EIC) construction and EIC peak picking algorithms which produce substantially less false positive peaks compared to XCMS and MZmine 2.
- These new algorithms are integrated into the MZmine 2 framework so users can easily implement the algorithms as well as use MZmine 2's visulization tools.
- We also show that the the new algorithms perform at least as wel as XCMS and MZmine 2 in terms of detecting peaks known to be present in the data.

Introduction

False positive and false negative peaks detected from extracted ion chromatograms (EIC) are a serious problem with existing software packages that preprocess untargeted liquid or gas chromatography-mass spectrometry (LC/MS or GC/MS) metabolomics data. False positives can translate downstream into spurious or missing compound identifications makeing the automated detection of metaboloites imposible without sigificant human intervention.

We have developed new algorithms that carry out the sequential construction of EICs and detection of EIC peaks. We compare the new algorithms to two popular software packages, XCMS and MZmine 2, and present evidence that these new algorithms detect significantly fewer false positives. Regarding the detection of compounds known to be present in the data, the new algorithms perform at least as well as XCMS and MZmine 2. The new algorithms have been developed as part of the Automated Data Analysis Pipeline (ADAP) workflow [1-3] and worked into the MZmine 2 framework so users can easily use them.

Methods: ADAP EIC Construction

Define ε to be the mass tolerance parameter

1) Take all the data points in a data file, sort them by their intensities, and remove those points (mostly noise) below a certain intensity threshold.

2) Starting with the most intense data point, the first EIC is created.

3) For this EIC, establish an immutable m/z range that is the data point's m/z plus and minus ε where ε is specified by the user.

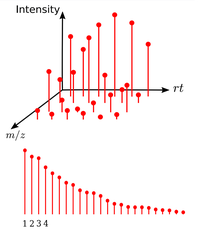
4) The next data point, which will be the next most intense, is added to an existing EIC if its m/z value falls within its m/z range.

5) If the next data point does not fall within an EICs m/z range, a new EIC is created. New EICs are only created if the point meets the minimum start intensity requirement set by the user.

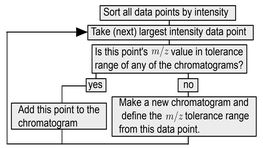
6) An m/z range for a new EIC is created the same way as in step (3) except the boundaries will be adjusted to avoid overlapping with pre-existing EICs. As an example consider an existing EIC with m/z range (100.000,100.020) for ε=0.01. If the new EIC is initialized with a data point having an m/z value of 100.025, then this new EIC will have a m/z range set to (100.020,100.035) rather than (100.015,100.035).

7) Repeat steps (4)-(6) until all the data has been processed.

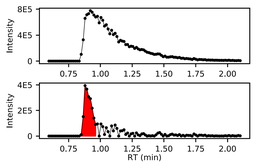
8) Finally, a post processing step is implemented. Only EICs with a user defined number of continuous points above a user defined intensity threshold are kept.



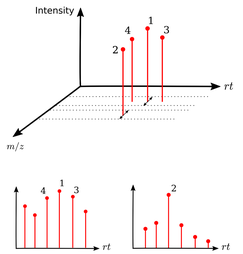
ADAP does not perform any type of baseline correction, because we have found that imperfections in baseline correction methods can produce convincing false positives in the data



Summarized ADAP EIC construction workflow diagram.



Example of baseline correction creating a convincing false positive peak.



Methods: ADAP EIC PeakPicking

A real peak in an EIC should create a local maxima in the wavelet coefficients at multiple scales. The wavelet scale for which the wavelet most closely matches the shape of the peak, the best scale, will create the largest coefficient. Scales close to the best scale should also have reasonably similar shapes to the peak and therefore create adjacent maxima between those scales. Ridgelines are the series of connected local maxima across scales which are indicative of a real peak. Our procedure for detecting the ridgelines is similar to that described by Du et al. [4] and Wee et al. [5] and is as follows.

1) Begin with the coefficients corresponding to the largest wavelet scale.

2) Find the largest coefficient at this scale and initialize a ridgeline.

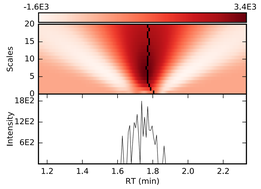
3) Remove all coefficients that are within half the estimated compact support of the Ricker wavelet (2.5 times the current scale).

4) Find the next largest coefficient discounting all removed coefficients and initialize another ridgeline.

5) Repeat steps (3)-(4) until there are no more coefficients remaining for this wavelet scale.

6) Move to the next scale (decrease by one) and repeat (1)-(6). Add new coefficients to an existing ridgeline if they are close in RT. We define close to be a difference in their indices that is less than or equal to the current scale being investigated.

7) After all scales have been processed, ridgelines must have a length, i.e., the total number of scales represented in the ridgeline, greater than or equal to 7, and not more than 2 gaps (missing scales) total.



Example Ridgeline and the corresponding peak

Determination of Peak Location and Boundaries

- The location of the peak is taken to be the RT of the largest coefficient in the ridgeline.
- The left (right) boundary of the peak is taken to be the RT of the peak minus (plus) the best scale multiplied by the time between scans.
- Peak boundaries should be close to local minima. However, the boundaries determined above often do not coincide with the local minima. We correct the boundaries to the first local minima, closest to the boundry determined in the way described above, on each sides of the EIC.

Signal-to-Noise Ratio Estimation

Define S to be the signal
Define N to be the noise
Define PW to be the peak width where the peak width is defined to be the number of scans between the two boundaries of a peak

Method 1:

1) Set two windows, one on each side of the peak in the EIC. The windows begin at the left and right peak boundaries and end at the peak boundaries ± 2xPW.

2) Calculate the standard deviation of the intensities in the two combined windows and store it as one possible value of the noise.

3) Expand both windows out from the peak by a single scan. The boundaries closest to the peak remain the same. After the first expansion, each window has a length of 2xPW+1.

4) Calculate and store the standard deviation of the intensities in the combined windows.

5) Repeat steps (3)-(4) until each window has a length of 8xPW.

6) Incrementally shrink each window by one scan, calculating and storing the standard deviations of the combined windows. The windows are shrunk by moving the boundary closest to the peak toward the boundary furthest from it.

7) Repeat step (6) until the window size is 2xPW.

8) The final noise estimate is taken to be the smallest stored standard deviation.

Method 2:

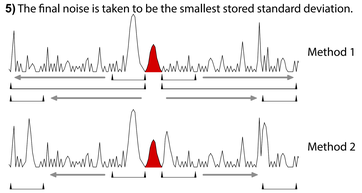
1) Same as (1) in method 1.

2) Same as (2) in method 1.

3) Shift each entire window away from the feature by one scan; the window lengths do not change.

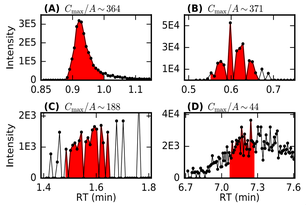
4) Repeat steps (2)-(3) until each window's boundary furthest from the feature is 8xPW from the closest boundary of that feature.

5) The final noise is taken to be the smallest stored standard deviation.



Using the wavelet coeficients to filter false positives

The magnitude of the wavelet coefficient can be used to help determine how good the peak shape is but it alone is not sufficient for determining whether or not a peak is real. This is because the wavelet coefficient has a strong dependence on the intensities of the data points in the peak. By dividing the coefficient of a peak by the peak's area we produce a number wich can be used to help filter out false positives based on their shape. We call this the coefficient-over-area (C/A) value.



To the left are several example peaks in the YP01 data file shown with their corresponding (C/A).

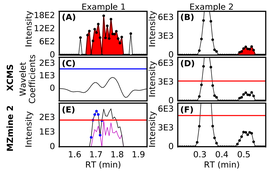
- An important property of this measure, is that intermittent dips in the intensity can increase the value due to the reduced area as seen in panel (B) when compared with panel (A)
- C/A is beneficial for finding messy low intensity peaks.
- Can be problematic if the area is so small it results in the detection of a peak with a very bad shape.

Additional Peak Property Filters

1) After peaks have been detected through CWT and ridgeline detection, it could be useful to discount low intensity peaks or keep only the highest intensity peaks. ADAP includes a second (in addition to the EIC construction threshold) intensity threshold. Peaks with heights less than this threshold will be discarded.

2) It is not uncommon to see EIC peaks that contain zero intensity points because of missing mass values in some scans. We assume these zero intensity points are missed because of the instrument and, in general, would like to detect peaks with missing points as long as their overall profile suggests that they correspond to real compounds.

Two Examples of Peaks Found by ADAP but missed by XCMS and MZmine 2



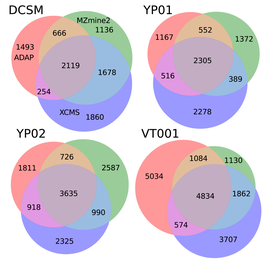
Shown in panles (A) and (B) of the figure to the left, we show two peaks colored in red that are detected using the ADAP algorithms but missed by both XCMS and MZmine 2 when using *centWave*. Though the peak in panel (A) does not have a smooth profile, it has been manually confirmed to be the first isotope (^{13}C) of a compound whose monoisotopic mass produces a clear high intensity EIC.

- Panel (C) illustrates that the panel (A) peak was missed by XCMS becuase a check perfomed by *centWave* does not find a large enough (blue line) coefficient at the smallest wavelet scale in the ridgeline.
- Panel (E) depicts that MZmine 2 missed the panel (A) peak becuase the number of points above the estimated baseline was not adaquate.
  - Red line: estimated baseline
  - Magenta line: original EIC
  - Black line: Distorted (bug in MZmine 2/centWave interface) EIC that MZmine 2 passes into centWave.
- Panel (D) shows that XCMS missed the panel (B) peak because there are no points above the baseline. Baseline is over-estimated due to the presence of the leftmost peak.
- Panel (F): For the same reason as above the panel (B) peak was missed by MZmine 2 using centWave.

Results

We have run EIC construction and EIC peak detection on four data files DCSM, YP01, YP02, and VT001 using XCMS, MZmine 2, and ADAP. Each software package detects the majority of monoisotopic peaks of the compounds manually identified to be present in the data.

We randomly sample peaks from each lobe of the Venn diagram and visually inspect each of the samples. Using several criteria we count the number of "good" peaks in each sample. The count of good peaks in the random sample gives us an estimate of the proportion of good peaks to false positives in each lobe. We use the Clopper-Pearson method [6] to determine the 95% confidence interval (CI) for each estimate.



(top number) Percentage, shown in red, of good peaks in the ADAP, XCMS, or MZmine 2-only lobe of the respective Venn diagrams.
(bottom range) 95% confidence interval, shown in black.

Data File	ADAP (%)	XCMS (%)	MZmine 2 (%)
DCSM	94.5 91.8-96.5	16.5 13.0-20.5	43.3 38.3-48.3
YP01	67.3 62.4-71.8	18.0 14.4-22.1	6.5 4.3-9.4
YP02	52.3 47.2-57.2	36.8 32.0-41.7	3.0 1.6-5.2
VT001	46.8 41.8-51.8	3.5 1.9-5.8	2.0 0.9-4.0

Percentage of good peaks (top) and the 95% confidence interval (bottom) in two of the important overlapping reagions of the YP01 Venn diagram.

Data File	XCMS and MZmine 2 Overlap (%)	ADAP, XCMS and MZmine 2 Overlap (%)
YP01	67.8 62.9-72.3	83.3 79.2-86.8

Conclusion

- Accurate construction of EICs and detection of peaks from EICs are critical for the success of any untargeted, mass spectrometry-based metabolomcis studies because false positive and false negative EIC peaks can turn into false and missing compound identifications.
- Motivated to come up with new EIC conastruction and EIC peak picking algorithms by the high rate of false positive and false negative peaks detected by existing software packages.
- Compared the performance of the new algorithms with results found with XCMS and MZmine 2 and demonstrated that the percentage of false positive peaks detected by ADAP is significantly lower.
- Showed that the reduction of false positives does not come at the cost of poor sensitivity.
- New algorithms have been incorporated into MZmine 2 to take advantage of the its strengths including modularity, visualization, and flexibility.

With LC and GC/MS platforms becoming increasingly sensitive, more compounds are now detectable in biological samples. The development of automated data preprocessing pipelines that can distinguish real peaks in the data from noise or other artifacts is important for obtaining a clear picture of the role of metabolites in biological processes. Currently no software package performs well enough that the results can be trusted without a significant amount of human oversight and involvement. We hope that the ADAP algorithms are one step forward in the direction of reducing false positive and false negative chromatographic peaks and in the direction of fully automated data preprocessing.

Acknowledgement

We thank National Science Foundation Award 1262416 for funding this research and development.

References:

[1] Jiang, W.; Qiu, Y.; Ni, Y.; Su, M.; Jia, W.; Du, X. J Proteome Res 2010, 9, 5974-81.
[2] Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. Anal Chem 2012, 84, 6619-29.
[3] Ni, Y.; Su, M.; Qiu, Y.; Jia, W.; Du, X. Anal Chem 2016, 88, 8802-11.
[4] Du, P.; Kibbe, W. A.; Lin, S. M. Bioinformatics 2006, 22, 2059-65.
[5] Wee, A.; Grayden, D. B.; Zhu, Y.; Petkovic-Duran, K.; Smith, D. Electrophoresis 2008, 29, 4215-25.
[6] Newcombe, R. Confidence Intervals for Proportions and Related Measures of Effect Size; Chapman & Hall/CRC Biostatistics Series; CRC Press, 2012.

Overview of Data Files

DCSM: A standard mixture file that was generated from a mixture of 22 standard compounds 21 of which were manually confirmed to exist in the data file. Equipment was OrbiTrap Velos mass spectrometer and Waters Acquity HSS T3 column using a reverse phase chromatographic method.

YP01, YP02, and VT001 were all generated from NIST Standard Reference Material (SRM) 1950, a representation of human plasma. Specific equipment used is as flollows:
YP01: LTQ Orbitrap Velos mass spec. and a ThermoFisher reverse phase anion exchange column.
YP02: LTQ Orbitrap Velos mass spec. and a ThermoFisher reverse phase C18 column.
VT001: Thermo Q Exactive HF Orbitrap mass spec. and a ThermoFisher reverse phase anion exchange column.

For more information:
http://www.du-lab.org dulab.binf@gmail.com