ADAP-GC: A software package for preprocessing gas chromatography-mass spectrometry metabolomics data

Aleksandr Smirnov[1], Owen Myers[1], Wei Jia[2], Douglas I. Walker[3], Shuzhao Li[3], Dean P. Jones[3], Xiuxia Du[1]*

[1]University of North Carolina at Charlotte, [2]University of Hawaii Cancer Center, [3]Emory University

Overview

ADAP-GC (**A**utomated **D**ata **A**nalysis **P**ipeline) is a computational workflow developed for extracting metabolite information from raw gas chromatography mass spectrometry (GC/MS) metabolomics data. It carries out a sequence of tasks including

- Construction of extracted ion chromatograms (EICs)
- Detection of peaks from EICs
- Spectral deconvolution
- Alignment of analytes across samples (in development)

ADAP-GC is implemented as a part of ADAP 3.2 library package written in Java and incorporated into the framework of MZmine 2 (https://github.com/mzmine/mzmine2/releases).

Introduction

The algorithms of ADAP-GC workflow are based on the works by O. Myers et al. [1] on construction of EICs and detection of EIC peaks and on the spectral deconvolution algorithm developed by Y. Ni et al. [2]. The main objectives in development of the current version of ADAP-GC include:

- **Ability to process high mass resolution data.** With ever more GC/MS metabolomics data acquired at high mass resolution, it has become necessary to update ADAP algorithms to handle such data.
- **Flexibility of algorithms.** ADAP algorithms have been first developed in 2010 to process GC/MS data. The new algorithms are designed to process both GC/MS and LC/MS data. Moreover, the algorithms can be used on both high and low mass resolution data.
- **Accessibility to the research community.** The previous versions of ADAP-GC workflow have been mostly prototyped in R and were not ready for an extensive use. The new version is developed to be user-friendly and easily accessible to researchers (Figure 1).

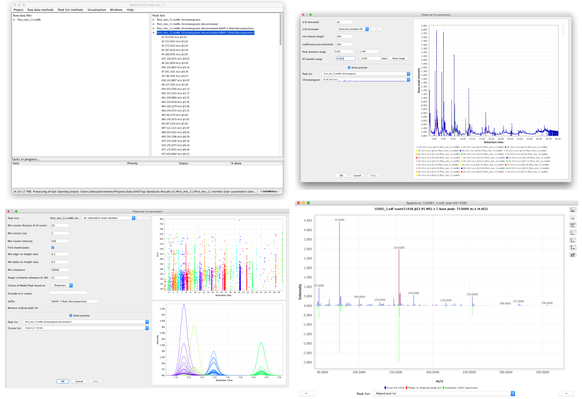

Figure 1: MZmine 2: (Top-Left) Main window; (Top-Right) EIC peak detection parameters window with a preview; (Bottom-Left) Spectral deconvolution parameters window with a preview; (Bottom-Right) Visualization of a mass spectrum constructed by spectral deconvolution.

Methods

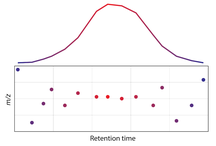

Figure 2: Accuracy and precision of m/z measurements at different intensities.

EIC construction is designed to process high mass resolution data. EICs are constructed by detecting small m/z ranges that contain relevant sequences of data points. Those points form an EIC. Each data point either falls into an existing range or forms a new range.

The algorithm is similar to the algorithms included in XCMS and MZmine 2. However, while those implementations process data points chronologically, ADAP algorithm parses data points from the highest intensity to the lowest intensity in the entire data file. This helps to find more accurate m/z ranges since high-intensity points have higher accuracy and precision of m/z values.

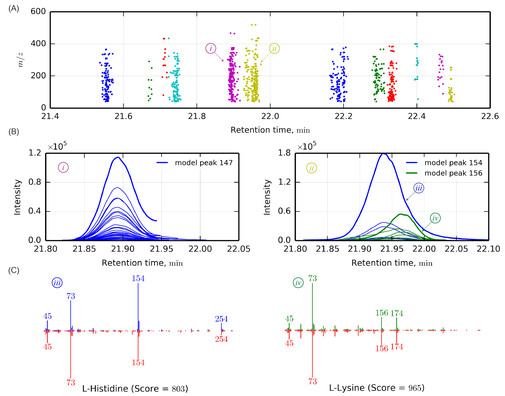

Figure 3: Results of Spectral Deconvolution: (A) Clustering based on the retention time; (B) Clustering based on the elution profile; (C) Constructed fragmentation spectra.

Spectral Deconvolution consists of several steps:

1. Determine dense **clusters** of EIC peaks with similar retention times (Figure 3A);
2. In each cluster, **filter** out composite EIC peaks and peaks with low sharpness;
3. Refine each cluster by finding **groups** of EIC peaks with similar elution profiles (Figure 3B) — each group corresponds to one analyte;
4. In each group, choose a **model peak** that represents the elution profile of an analyte;
5. Decompose each EIC peak into a linear combination of model peaks and combine the decomposition coefficients to form the **fragmentation spectra** of all analytes.

Compared to ADAP-GC 3.0, the new algorithm eliminates the step of determining deconvolution windows and improves the overall runtime by using DBSCAN-clustering in Step 1 and optimizing decomposition in Step 5.

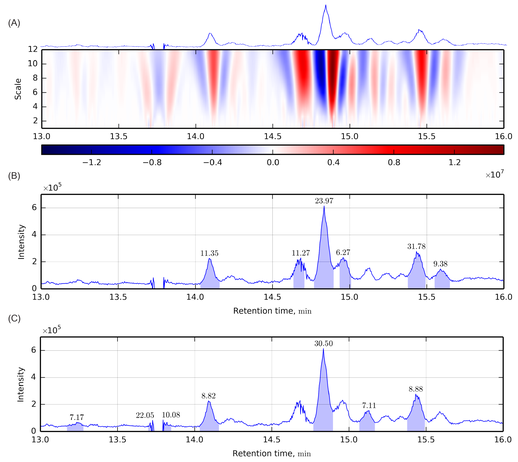

Figure 4: EIC Peak Detection: (A) Coefficients of the continuous wavelet transform; (B) and (C) Detected peaks with their signal-to-noise estimates.

EIC Peak Detection is based on the continuous wavelet transform (CWT):

1. **Convolution** of an EIC and the Ricker wavelet is calculated at different scales.
2. **Ridgelines** of the continuous wavelet transform coefficients are detected (Figure 4A). Each ridgeline of sufficient length gives rise to a peak.
3. **Signal-to-noise** (SNR) ratio is estimated for every peak, and peaks with a small SNR are filtered out. Users have an option to choose one of two SNR-estimators:
   - CWT-based estimator (Figure 4B) — effective when EIC peaks are composed of a large (50 and more) number of data points — the signal and noise levels are estimated as coefficients of CWT at the best-fit scale and 95-quantile of the absolute values of CWT coefficients at the smallest scale respectively;
   - Intensity-based estimator (Figure 4C) — effective when EIC peaks are composed of a small (5–10) number of data points — the signal and noise levels are estimated as the peak intensity and the minimum standard deviation of the signal around the peak.

Compared to the existing peak detection algorithms in XCMS and MZmine 2, the new algorithm detects significantly less false EIC peaks and a comparable number of true EIC peaks.

[In development] Alignment of analytes is performed by detecting similar analytes in samples. The similarity is estimated by the formula

$$S(a_1, a_2) = \alpha S_{spec}(a_1, a_2) + \beta S_{prof}(a_1, a_2)$$

where α, β are weighting coefficients, S_{spec} spectrum similarity of two analytes, and S_{prof} similarity of the elution profiles of two analytes. The value S_{prof} is estimated after the alignment of two elution profiles by minimizing their cross-correlation.

Results

A series of 16 mixtures containing 256 compounds, including brominated flame retardants, dioxins, furans, polychlorinated biphenyls, organonitrogen pesticides, pyrethroids, organophosphorus pesticides, and organochlorine pesticides were analyzed. Analyte separation was accomplished by Trace 1310 GC (Thermo Scientific) and accurate mass detected by Q-Exactive GC hybrid quadrupole-Orbitrap GC-MS/MS (Thermo Scientific).

Sample	Total	Identified	Ave Score	Std Div
PCB_Content_Eval_Mix1	6	6	889	50.06
PCB_Content_Eval_Mix2	3	3	878	24.57
PCB_congener_calibration_mix	14	14	884	42.74
PBB153	1	1	851	0.00
PBDE_Tech_Mixes	6	6	845	29.82
Dioxins	5	5	772	71.39
Furans	5	5	834	96.97
Pest_mix_08	16	16	840	100.63
Pest_mix_09	40	40	846	72.95
Pest_mix_10	25	24	785	101.14
Pest_mix_11	28	25	888	86.87
Pest_mix_12	35	33	897	52.02
Pest_mix_13	27	25	852	95.79
Pest_mix_14	9	9	919	33.09
Pest_mix_15	25	24	869	52.66
Pest_mix_16	8	8	882	46.91
			858	59.85

Conclusion

A software package, ADAP-GC, for preprocessing GC/MS metabolomics data is presented. ADAP-GC features:

- Accurate construction of EICs for high mass resolution data and reduced number of false EIC peaks during EIC peak detection;
- Ability of EIC construction and peak detection algorithms to work with both GC/ and LC/MS data;
- Seamless integration with MZmine 2 and its advanced visualization and data import tools.

In order to make ADAP more accessible to the metabolomics community, the developed algorithms are incorporated into MZmine framework (version 2.25 [3]). Thus, MZmine 2 acquires three new computational modules for processing high mass resolution data and can now process GC/MS data as well as LC/MS.

Acknowledgement

We acknowledge the NSF Award 1262416, NIEHS: P50ES026071, P30ES019116, U2CES026560 and EPA: 83615301 for funding this research and development.

For more information ...

http://www.du-lab.org, dulab.binf@gmail.com

References

[1] O. Myers, S. Sumner, S. Li, S. Barnes, and X. Du. *Analytical Chemistry*, (submitted).

[2] Y. Ni, M. Su, Y. Qiu, W. Jia, and X. Du. *Analytical Chemistry*, 88(17):8802–8811, 2016.

[3] MZmine 2. `https://github.com/mzmine/mzmine2/releases`. [Accessed May 15, 2017].