ADAP: A Suite of Computational Algorithms and Software Tool for Preprocessing Mass Spectrometry-Based Metabolomics Data

Xiuxia Du[1], Aleksandr Smirnov[1], and Susan Sumner[2]
[1] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, USA
[2] Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

Untargeted mass spectrometry (MS)-based metabolomics, the unbiased detection and relative quantitation of ideally all metabolites in a biological system, has become a powerful discovery tool in many scientific disciplines. The informatics pipeline for making sense of the resulting data involves preprocessing of the raw peak data to detect unique chemical species, assignment of specific metabolites to these species, and integration of these metabolites into a coherent and physiologically meaningful integrated multi-omics framework that can yield a holistic understanding of the biological system (Figure 1A). As the first step of this informatics pipeline, data preprocessing (Figure 1B) is critical for the success of the metabolomics study. Data preprocessing generally consists of three computational steps: peak detection, peak grouping and annotation for LC-MS and spectral deconvolution for GC-MS, and peak alignment. While many existing software tools have performed admirably considering the complex nature of the data, the underlying algorithm in each step of the data preprocessing are being seriously challenged as many more metabolites can now be detected and data has become much more complex than before, due to the unprepredented sensitivity of the analytical platforms that has been made possible by recent technological advances in chromatography and mass spectrometry.

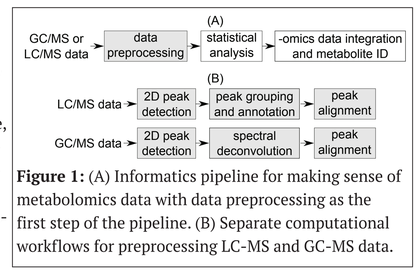


Figure 1: (A) Informatics pipeline for making sense of metabolomics data with data preprocessing as the first step of the pipeline. (B) Separate computational workflows for preprocessing LC-MS and GC-MS data.

Peak Detection

Peak here refers to a unique pair of mass and retention time (RT) that corresponds to an ion. It is also called a feature, but "peak" is used here to avoid possible confusions because features mean different things for different people. Peak detection is the first step of data preprocessing and is critical for successful extraction of metabolite information from raw LC-MS and GC-MS data. Raw LC-MS and GC-MS data has three dimensions: *m/z* (mass-to-charge ratio), retention time (RT), and intensity. In existing software tools including the current release version of ADAP, two steps constitute the peak detection process: (1) construction of extracted ion chromatograms (EIC) in the 2D plane of *m/z* and RT, and (2) detection of chromatographic peaks in the 2D plane of intensity and RT.

Construction of EICs

Figure 2 shows the summarized workflow. Specifically, first define ε to be the mass tolerance parameter, then
(1) Take all the data points in a data file (Figure 3A), sort them by their intensities (Figure 3B), and remove those points (mostly noise) below a certain intensity threshold.
(2) Starting with the most intense data point (Figure 3C), the first EIC is created (Figure 3D).
(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 (Figure 3E). 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.

Figure 2: Summarized ADAP EIC construction workflow diagram.

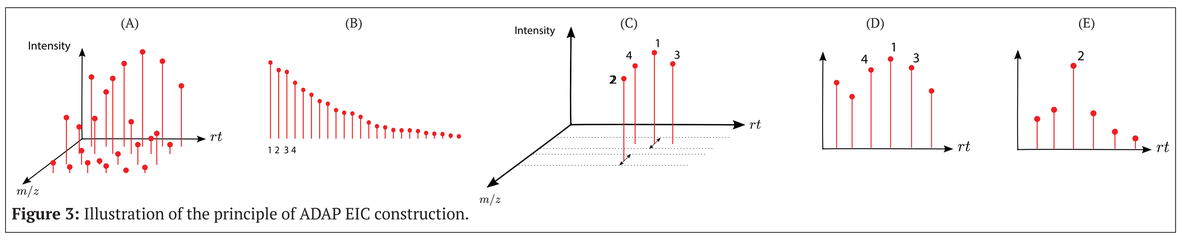


Figure 3: Illustration of the principle of ADAP EIC construction.

Detection of Peaks from EICs

Peak detection uses continuous wavelet transform (CWT), a widely used signal processing technique. 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. The requirement that a ridgeline must exist for a peak makes CWT-based peak detection robust (Figure 4). Ridgelines are constructed according to the following procedure.

(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.

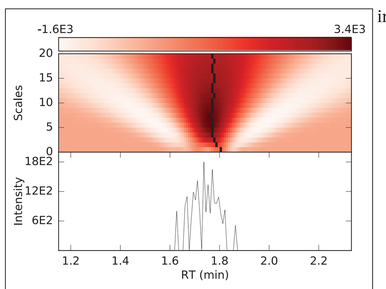


Figure 4: Principle of CWT-based peak detection from EICs. (Top) Wavelet coefficients with respect to wavelet scales and RT. (Bottom) The corresponding EIC.

m/z vs ppm As Mass Tolerance for EIC Construction

Mass tolerance is a critical user-defined parameter in EIC construction that can have a huge impact on the constructed EICs and the peaks detected from them. This parameter can be specified in terms of either *m/z* unit or ppm. Due to the importance of this parameter, we investigated what unit of mass tolerance should be preferred by examining 441 high quality EIC peaks in a data file and the data points that form these EICs. It turns out that a mass range in *m/z* of 0.02 could ensure that most of the EIC peaks would include the majority of the data points forming each peak, whereas a mass range in ppm needs to be about 100 ppm to achieve the same goal (Figure 5). However, such a huge ppm value will almost certainly cause the issue of merging two or more EICs for large masses. On the other hand, a much smaller ppm tolerance will almost certainly cause the issue of splitting two or more EICs for small masses. This investigation demonstrated that a mass tolerance in *m/z* is more appropriate for the construction of EICs if one single mass tolerance is to be used for all of the *m/z* values in a data file.

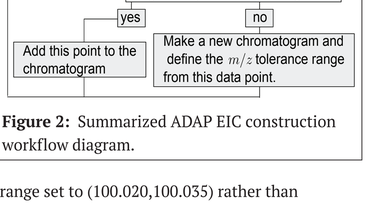


Figure 5: Varying mass tolerance for detecting EIC peaks of different *m/z* illustrated using 441 peaks in a data file.

Spectral Deconvolution

Spectral deconvolution examines all of the ions detected in the step of EIC peak detection, collects ions produced by the same analytes, and constructs the fragmentation spectrum for each analyte. Each constructed spectrum is expected to contain ions of one single analyte because spectral deconvolution separates ions that are in the same raw mass spectrum but belong to different coeluting analytes. The constructed fragmentation spectra are used later to identify the analytes.

Spectral deconvolution in ADAP (Figure 6) is achieved by determining the number of eluting analytes, choosing model peaks representing the elution profile of each analyte, and decomposing chromatographically unresolved EIC peaks into a linear combination of the model peaks. To simplify and to speed up the deconvolution procedure in ADAP, the entire retention time range is first split into a number of deconvolution windows. These windows are chosen so that (i) each window contains the entirety of EIC peaks produced by the same analyte or by coeluting analytes and (ii) each window contains a much smaller number of EIC peaks in comparison with the total number of EIC peaks in the entire data file. Subsequent deconvolution steps are carried out separately in each window so that the deconvolution algorithms are not overwhelmed with a large number of EIC peaks.

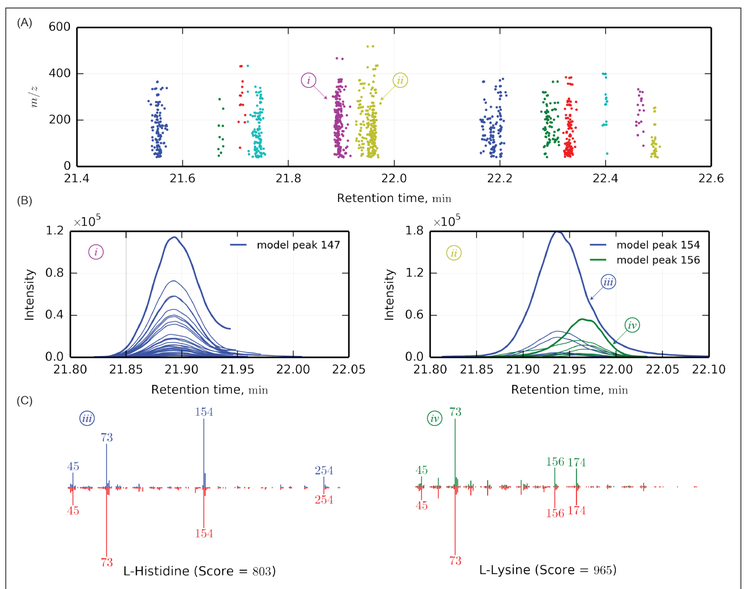


Figure 6: Deconvolution workflow illustrated using a data file from urine sample acquired at the unit mass resolution: (A) DBSCAN clustering of the apex retention times of all EIC peaks in the entire data file. Each color represents a cluster. (B) Hierarchical clustering of the elution profiles of EIC peaks in clusters i and ii in panel A. Cluster i results in one cluster and cluster ii results in two smaller clusters. (C) Constructed fragmentation spectra (top) and in-house library spectra (bottom) acquired on the same equipment as the sample.

Deconvolution within each window starts with two sequential clustering phases applied to EIC peaks. The first-phase clustering is based on the proximity of peak apexes in the time domain, and each resulting cluster indicates the presence of at least one analyte. Because coeluting analytes are in close proximity of each other and could fall in the same cluster, simple comparison of retention times cannot detect all coeluting analytes. Detection of these analytes could be achieved by using elution profiles. Toward this end, a second-phase clustering that is based on the elution profiles of EIC peaks is carried out to group unique EIC peaks from each first-phase cluster. As a result, each first-phase cluster can be split into one, two, or more smaller clusters, and each resulting cluster indicates the presence of one single analyte. From each second-phase cluster, a model peak is selected that can best represent the elution profile of the corresponding analyte. Because an observed EIC peak can be produced by two or more coeluting analytes, the fragmentation spectra of the detected analytes are constructed by decomposing every observed EIC peak into a linear combination of model peaks.

ADAP spectral deconvolution has been developed and its performance has been evaluated using unit mass resolution data acquired on TOF-MS from standard-mixture and urine samples and high mass resolution data acquired on GC-Orbitrap from environmental pollutants samples. In addition, the identification and quantitation results from ADAP-GC for the unit mass resolution data were compared with those produced by AMDIS, AnalyzerPro, and ChromaTOF (Table to the right). ADAP-GC (both 3.0 and 3.2) and ChromaTOF produce similar results in terms of the number of identified compounds, their matching scores, and R2 values, whereas AMDIS and AnalyzerPro tend to miss certain compounds.

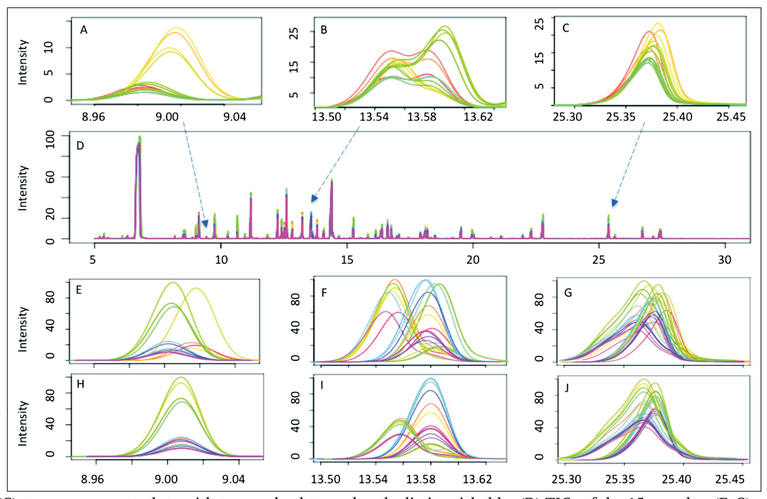
	standard mixture (196 compounds)			urine (224 compounds)		
	identified	score	R^2	identified	score	R^{2a}
ADAP-GC 3.2	188	921	0.997	223	917	0.929
ADAP-GC 3.0	189	917	0.998	224	903	0.926
AMDIS	179	899	0.990	221	906	0.803
AnalyzerPro	172	869	0.996	217	884	0.914
ChromaTOF	188	902	0.996	220	906	0.919

[a] R^2 values for the urine samples are approximate since precise concentrations of the compounds could not be determined a prioi.

Alignment

RT for the same compound shifts from run to run and alignment is needed to correct the RT so that subsequent statistical analysis can be performed. The principle in existing alignment algorithms can be roughly divided into two categories. One category uses warping to find a nonlinear function to correct RT. The other category creates a reference list of peaks and align peaks in individual data files to the reference peak list. However, warping functions can only capture system-level variations in RT and is incapable of capturing analyte-level variation. Alignment algorithms that use a reference peak list check for proximity of peaks in terms of *m/z* and RT and could shift peaks that correspond to the same analyze differently, resulting in mis-alignment.

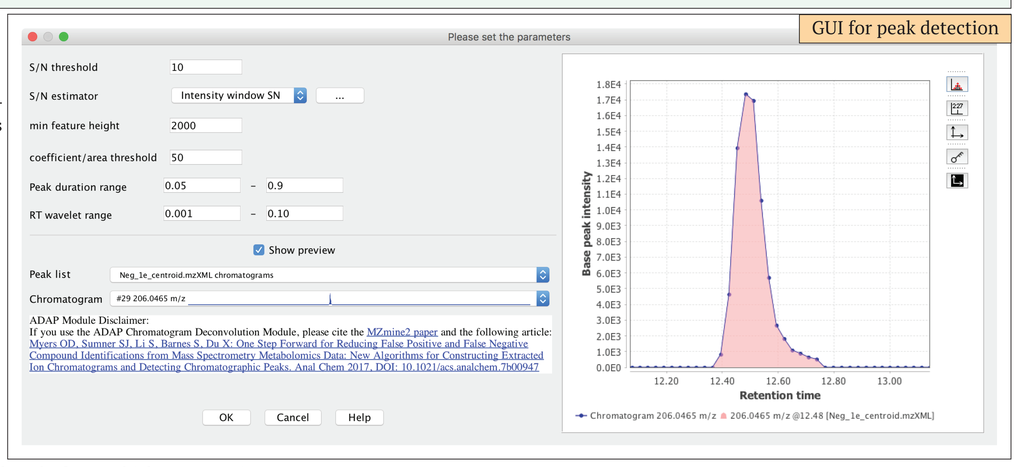
Conceptually, alignment should be formulated as a correspondence problem by finding the same analyte across data files. The alignment algorithm in ADAP is precisely analyte-based. Figure on the right shows the result of aligning 15 GC-MS data files. Each data file is represented by one unique color. (A-C) Total ion chromatograms (TICs) within three different time intervals. (A) one component elutes with two distinct TIC peaks; (B) two components elute with two distinct TIC peaks; (C) two components elute with two peaks that are barely distinguishable; (D) TICs of the 15 samples. (E-G) EICs before alignment; (H-J) EICs after alignment. EIC pairs E-H, F-I, and G-J correspond to TIC segments (A), (B), and (C), respectively. For the two EIC pairs (F and I) and (G and J), two co-eluting analyses became distinguishable only after alignment.



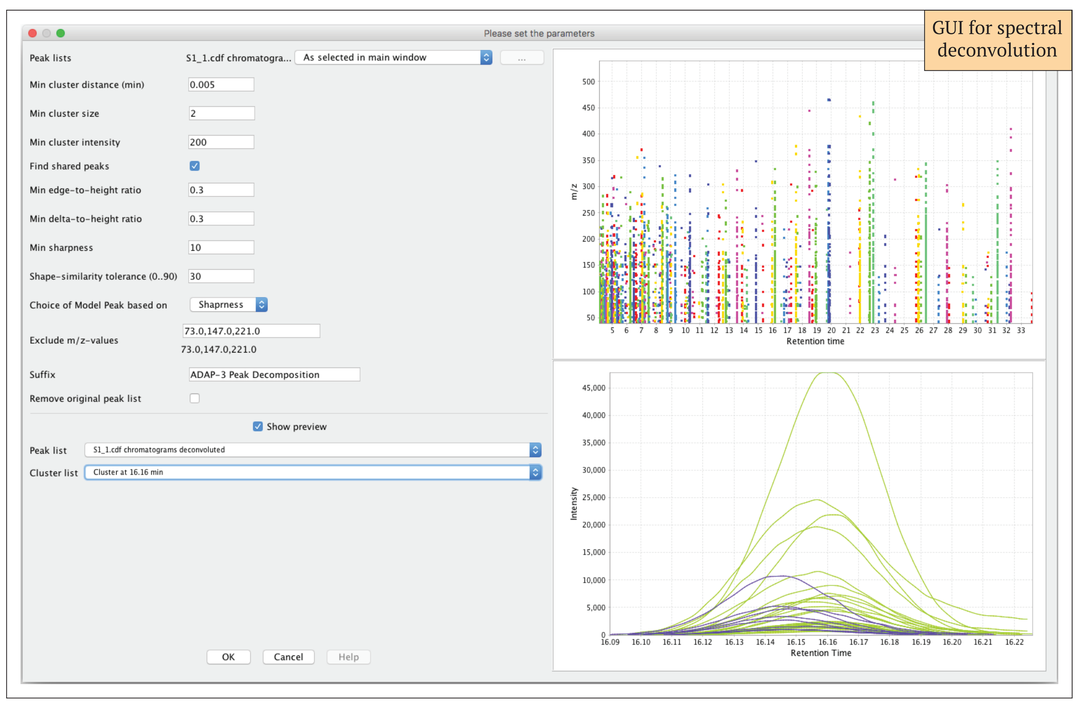
Software

ADAP algorithms have been implemented in Java and incorporated into MZmine 2, a graphical software framework tool that is used by thousands of researchers around the world. This incorporation makes it possible for ADAP to take advantage of the strengths of MZmine 2. These strengths include: (1) platform independence due to Java technology, (2) modular framework, which simplifies incorporation of new algorithms, and (3) rich visualization capabilities including display of spectra, chromatogram, and results from multiple preprocessing steps.

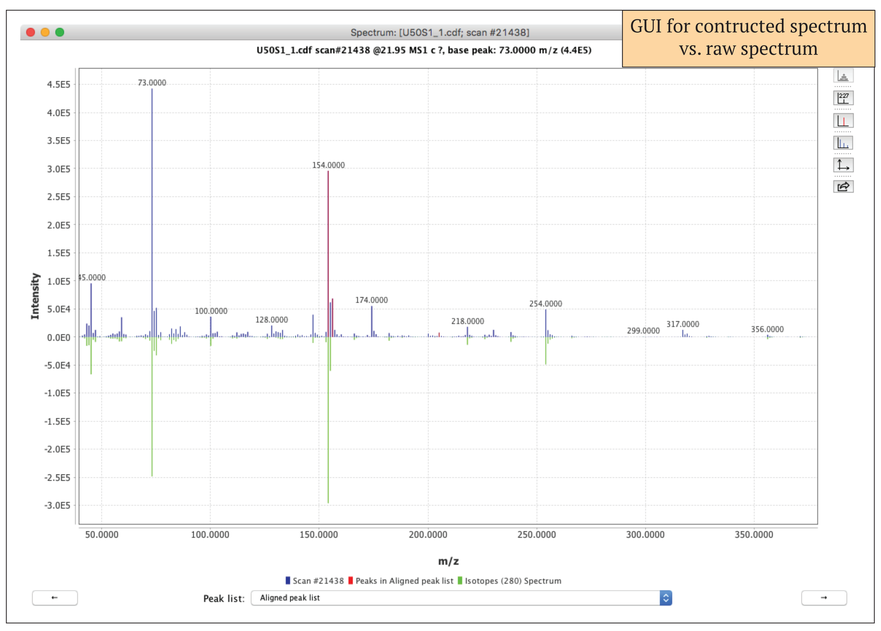
GUI for peak detection



GUI for spectral deconvolution



GUI for contructed spectrum vs. raw spectrum



References

[1] Aleksandr Smirnov, Wei Jia, Douglas I. Walker, Dean P. Jones, and Xiuxia Du*: **ADAP-GC 3.2: Graphical Software Tool for Efficient Spectral Deconvolution of Gas Chromatography-High Resolution Mass Spectrometry Metabolomics Data.** *Journal of Proteome Research* 2018, 17 (1):470-478.
[2] Owen Myers, Susan Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du*: **One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks.** *Analytical Chemistry* 2017, 89(17):8696-9703.
[3] Owen Myers, Susan Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du*: **A detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data.** *Analytical Chemistry* 2017, 89(17):8689-8695.
[4] Yan Ni, Mingming Su, Yunping Qiu, Wei Jia, and Xiuxia Du*: **ADAP-GC 3.0: Improved Peak Detection and Deconvolution of Co-eluting Metabolites from GC/TOF-MS Data for Metabolomics Studies.** *Analytical Chemistry* 2016, 88 (17):8802-8811.
[5] Xiuxia Du*, Steven Zeisel: **Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives.** *Computational and Structural Biotechnology Journal* 2013, 4, e201301013.
[6] Yan Ni, Yunping Qiu, Wenxin Jiang, Kyle Suttlemyre, Mingming Su, Wenchao Zhang, Wei Jia, and Xiuxia Du*: **ADAP-GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies.** *Analytical Chemistry* 2012, 84 (15), 6619-29
[7] Wenxin Jiang, Yunping Qiu, Yan Ni, Mingming Su, Wei Jia, and Xiuxia Du*: **An automated data analysis pipeline for GC-TOF-MS metabonomics studies.** *Journal of Proteome Research* 2010, 9 (11), 5974-81.

Acknowledgement

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