EIC + , an algorithm for automatic and unsupervised extraction of ion chromatograms in high resolution mass spectrometry

Jean-Claude Boulet

EIC+, an algorithm for automatic and unsupervised extraction of ion chromatograms in high resolution mass spectrometry

Jean-Claude Boulet[1]

SPO, INRAE, Univ Montpellier, Institut Agro, Montpellier, France

INRAE, PROBE Research Infrastructure, Polyphenol Analytical Facility, Montpellier, France

Abstract

Within the feature extraction process in high-resolution mass spectrometry (HRMS), we focused on the identification of EICs-Extracted Ion Chromatograms. A proposed algorithm tuned with its default parameters yielded EIC matrices which recovered most of the 836 reference features identified within a set of public and fully described Orbitrap data, already benchmarked with XCMS, MZmine2, MS-Dial and Compound Discoverer. A good accuracy on the calculation of the m/z values was also observed. Tools are available on the Galaxy Test Toolshed.

Keywords: extracted ion chromatograms; features; m/z values; high resolution mass spectrometry; EIC+ algorithm

1. Introduction

A mass spectrum is a beam of abundances of ions represented by their m/z values [3]. Within each mass spectrum, the regions of m/z values with high abundances, or signals, form mass peaks, or peaks. HRMS-high resolution mass spectrometry consists in liquid chromatography coupled to mass spectrometry. The analysis of a sample in HRMS yields a MS data, which is a series of mass spectra, each acquired at a different retention time (RT). A collection of several MS data will form a MS dataset. Within each MS data, a feature is defined when a same peak is identified in several mass spectra which are consecutive by their RTs. A feature is a triplet containing a m/z value, a retention time and a signal intensity. Extraction of features, performed by proprietary or free softwares, is the beginning of HRMS preprocessings. It can be summarized as follow for a MS1 acquisition (full scan). Let ms_i be the mass spectrum corresponding to the i^{th} retention time. Each ms_i is a matrix of two columns, m/z

Email address: jean-claude.boulet@inrae.fr (Jean-Claude Boulet)

values and signal intensities, which form peaks with the profile option of the mass spectrometer, or which have been converted to centroids with the centroid option. Successive ms_i, e.g. ms_i, ms_{i+1} and ms_{i+2}, may contain the same m/z value. From a MS1 data and a given m/z value, an Extracted Ion Chromatogram or EIC is obtained by collecting the signals associated to a given m/z value (with a tolerance) along all the RTs. The next process is the identification of features or chromatographic peaks, sometimes refered to as peak picking. But the calculation is tricky, and several softwares, or even the same software with different parameters, do not yield the same solution. Benchmarks give clues on their respective performances, according to different figures of merit. A first approach was to compare the features obtained by different softwares, using their distances calculated with their m/z and RT values. The figure of merit was the number of features identified by several softwares, considered to be true positives. Venn diagrams were used by Tautenhahn et al [8] to compare Centwave, MZMine and Matchfilter algorithms, and by Myers et al [4] to compare XCMS [7] and MZmine2 [6]. But an outlier on Venn diagrams is simply an algorithm which performs differently; its performances can be worse, or better. Then, another and more reliable approach was based on a ground truth, reference values to which experimental results could be compared. Li et al [2] compared several softwares on samples for which hundreds of compounds had been manually quantified. The figure of merit was the recovery rate of features, based on the proximity between reference and experimental values for three variables: RTs, m/z and signal intensities. Such approach targets final users, because it gives them clues to choose a processing application. But it does not explain why an application performed better than another. In other words, we have seen that features extraction is a two-steps process: EIC extraction from raw MS data then features identification from EICs. The recovery rate was not 100% for any software: where was information lost? At the EIC step? At the features step? Both? Moreover, the accuracy of m/z prediction was not discussed, while it should be a mandatory figure of merit when the goal of the processing is to propose CHONS chemical formula.

Taking those remarks in consideration, the design of the following algorithm was focused on the EIC extraction and the accuracy of the prediction of the m/z values, with a parsimonious number of tuning parameters. The aim of this article was to show that almost all information from the dataset published by Li et al could be recovered by the EICs, and beyond, that it remains possible to improve the features extraction process.

2. Material and methods

2.1. The EIC+ *algorithm*

Contrary to ADAP [5] or Centwave [8] which process each data once, our algorithm perfoms several processings of the original centroid data. The goal was to determine more accurately the experimental m/z values leading to the future EICs. At each loop, m/z values were recalculated, converging towards

stable values. The algorithm called *EIC+* is schematized in figure 1. The three tuning parameters were: *mztol*, *eicsig* and *dw*, i.e. a tolerance around a given m/z value, a minimum of signal and a Durbin-Watson threshold. The starting point was a centroid high resolution MS data.

- Identification of a first m/z list
  All the signals of the MS data were gathered into the same matrix of two columns: m/z values and signal intensities. The m/z values were sorted in ascending order, identical m/z values were merged, signal intensities added. The maximum of signal was associated to a m/z value called mz_{max} which was stored in a list of m/z. The m/z between the boundaries [mz_{max}-*mztol*, mz_{max}+*mztol*] were dropped. Then, the algorithm checked for the next mz_{max} and the previous calculations were repeated until the maximum of the remaining signals reached *eicsig*.

- Filling a first EIC matrix
  The first EIC matrix was initiated as a matrix of zeros; the rows were the retention times, and the columns the list of mz_{max} previously identified. The MS data was scanned again, each m/z value was compared to the m/z of the EIC matrix. If an absolute difference was lower than *mztol*, the signal was added into the EIC matrix at the corresponding retention time and m/z; otherwise it was discarded. Some statistics were also collected.

- Converging towards a last EIC matrix
  The statistics allowed a new barycentric calculation of each m/z value of the previous EIC matrix. Two EICs which became closer than *mztol* were merged into a new and temporary EIC whose m/z value was recalculated by weighting the m/z with the signals used to fill the matrix. After exploring all of the possibilities of merging EICs, a new EIC matrix was initiated to zeros and filled as previously described. Several loops were processed until a stop, no more than 5 m/z values merged in 3 successive loops. The last EIC matrix was obtained.

- Cleaning up the last EIC matrix
  The cleaning was achieved using a slightly modified version of the Durbin-Watson (DW) test. For each EIC forming a vector of length N_{RT} (the number of retention times), a vector of length $N_{RT} - 1$ containing the differences between two successive elements was calculated. The DW value was the ratio between the norm of the difference vector and the norm of the EIC vector. It ranged from 0 for continuous signals to 2 for very discontinuous signals. EICs with DW values higher than *dw* were supposed to be noise, so they were dropped.

2.2. Feature extraction from the EICs

Regular outputs from e.g. XCMS or MZMine2 are features, while regular outputs for *EIC+* are matrices of EICs, the step before. Our goal was to show

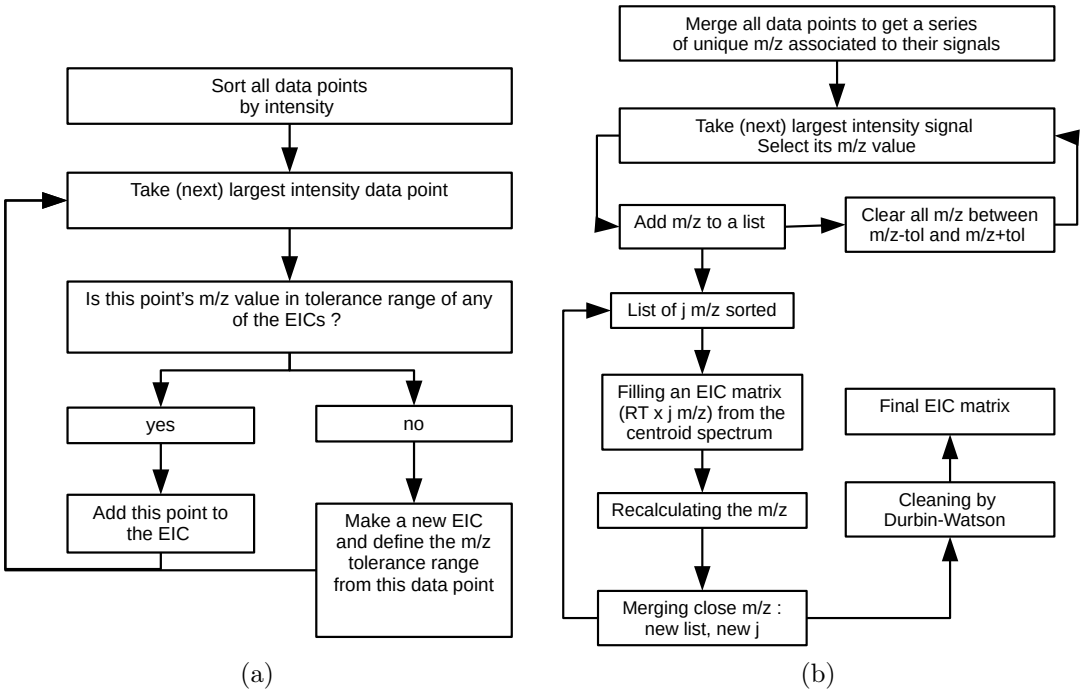


Figure 1: Flow diagram of MZmine2-ADAP (a) and of *EIC+* (b). The ADAP flow diagram was reproduced from figure 1 of Myers et al, Anal.Chem., 2017. The box "Filling an EIC matrix" in (b) performs almost the loop of the ADAP algorithm in (a), with the difference that, in case of "no", ADAP makes a new EIC with the data point whereas *EIC+* drops it.

that a large part of the information concerning the reference features had been captured within the EIC matrix. For comparison with the results obtained by Li et al [2], features extraction from our EIC matrices was mandatory. It was performed with a home-made tool, designed to be plugged into the workflow, not described because we do not claim any novelty with it. The results were good enough to use it, even if other algorithms may have performed better. To save calculation time, a selection of the EICs corresponding to the reference m/z values was performed on the EIC matrix before features extraction.

2.3. The MS datasets

Two MS datasets were considered. MS dataset 1 consisted in a single Orbitrap MS data, named *VI2016_AC_4A.raw* acquired at the University of Barcelona in 2016. The sample was a red wine spiked with acetaldehyde prior to its analysis. Polyphenols had been targeted, 117 compounds had been previously manually identified in this wine sample, following a previous work on model solutions [9]. Nevertheless, as several chemical compounds had exactly the same raw formulae, the number of different m/z values was 109. MS dataset 2 consisted in ten Orbitrap MS datasets described by Li et al [2], and available with their metadata thanks to the Metabolights project at: *https://www.ebi.ac.uk/metabolights/MTBLS733*. The spectra were named *SA1.raw* to *SA5.raw* and *SB1.raw* to *SB5.raw*. The metadata consisted in a list of 836 chemical compounds manually identified by Li et al into the 10 spectra, with their RTs, m/z values and signal intensities. Only MS1 spectra were considered. Before processing, two successive conversions were performed: from .raw to .md5 by MSConvert, then from .md5 to .h5 by HDFView.

2.4. The processings

All tools were coded in Scilab 6.0.2 *(https://www.scilab.org)* with the Fact toolbox, then wrapped into Galaxy tools. The centroid calculation was performed according to a previous work [1].

2.5. Tuning the EIC+ *with dataset* 1

The workflow concerning dataset 1 involved EIC extraction only, features were not considered yet. The development and the very first application of *EIC+* was based on dataset 1 and the figures of merit described hereafter. Preliminary experiments had concluded that the minimum of signal *eicsig* and the Durbin-Watson threshold *dw* could be fairly set to 30000 and 1 respectively. Variations around these values did not strongly modify the results (not shown). Our attention was focused on *mztol*. Values were chosen from 0.0012 to 0.0060 by steps of 0.0012. The algorithm was run with each of these values, yielding for each *mztol* a list of experimental m/z values. Then, each reference m/z value was seeked within these lists with m/z tolerances set to 0.0005, 0.0100, 0.0030, 0.0050 and 0.0100.

2.6. The benchmark, with dataset 2

The processing consisted in two successive steps: 1) extracting the EICs with *EIC+*; 2) extracting the features from the previously determined EICs.

2.6.1. Step 1: extracting the EICs with EIC+

The workflow concerning MS dataset 2 is described in figure 2. Compression was a means to save computational resources and time. It consisted in merging 5 successive RTs. When EICs had been extracted from each of the 5 datasets, they were averaged, yielding a single EIC file from SA and a single EIC file from SB. Then, features could be extracted.

2.6.2. Step 2: extracting the features

The features were extracted with a home-made tool using 4 parameters: i) the minimum and maximum sizes of the RT window, $0.3mn$ and $0.05mn$ respectively; ii) thresholds for the maximum of signal in each selected feature/peak and for the Durbin-Watson value, 100000 and 1 respectively.

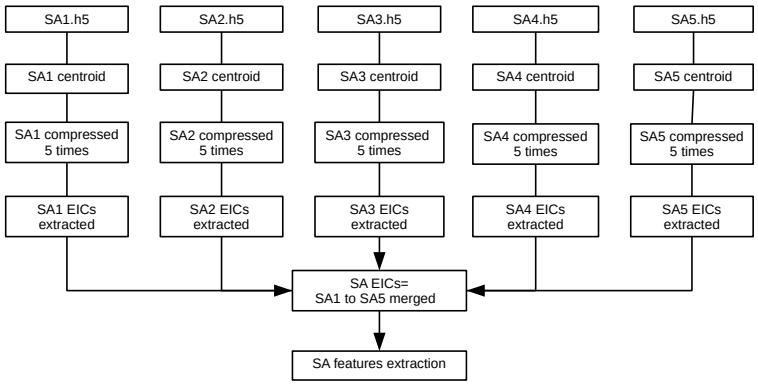


Figure 2: Workflow applied to SA1-SA5. SB1-SB5 were processed similarily.

2.7. The figures of merit

Five figures of merit were calculated. The EICs m/z recovery and the EICs m/z unicity ratio were applied to dataset 1 for tuning the parameters, and to dataset2 for assessing the quality of EICs extraction. The true features, the accurately quantified true features and the RMSEP-root mean square error of prediction of the true features m/z values were applied to dataset 2 after features extraction.

- The EICs m/z recovery
  The EIC recovery is N_R the number of chemical compounds identified by the processing. It should be compared to the total number of reference m/z values.

- The EICs m/z unicity ratio
  The i^{th} m/z value from the m/z reference list of chemical compounds can be associated to N_u^i m/z values of the EIC matrix, under a certain tolerance. The not-null N_u^i values were added and divided by N_R, yielding the unicity ratio:

$$unicity\ ratio\ =\ \frac{\sum N_u^i}{N_R}$$

  A ratio close to 1 suggests that there is non ambiguity of identification between the m/z values from the reference list and the m/z values from the EIC matrix.

- True features and accurately quantified true features
  Our experimental features were compared to the reference features provided by Li et al on the basis of their m/z values, their retention times and their signal intensities. According to the same authors, true features verified simultaneously shifts in m/z values and RTs lower than thresholds set to 10 ppm and 0.3 minutes respectively. One of our first motivations was to determine chemical compositions from the molecular weights of the ions, so we compared the m/z values on the basis of a mass unit, rather than of a percent which depends on the m/z values. For a comparison as fair as possible with the previous benchmark of Li et al, the median of the 836 reference m/z values was calculated, the value of 386 obtained. Therefore, the threshold was set to 10 ppm of 386, that is 0.00386 mass units. For the retention times, no threshold was used. Our algorithm provided the start and stop RTs, the test was passed when the reference RTs fell within this range. Then, accurately quantified true features corresponded to true features for which the experimental SB/SA signal ratio did not differ from more than 20% from the reference ratio determined manually by Li et al and refered to as fold changes in their data.

- The RMSEP of the true features m/z values
  Let $\mathbf{m}_{ref}$ and $\mathbf{m}_{exp}$ be vectors containing the N reference m/z values and the N experimental m/z values associated to the N true features defined above. The RMSEP-root mean square error of prediction was calculated for SA then for SB according to Equation 1, with $\mathbf{d} = \mathbf{m}_{ref} - \mathbf{m}_{exp}$.

$$RMSEP = \sqrt{\frac{\mathbf{d}^T\mathbf{d}}{N}} \tag{1}$$

3. Results and discussion

Results concerning the tuning of the parameters for dataset 1 were gathered into table 1. The best solution should present the highest recovery percent and the lowest unicity ratio, while *mztol* being as low as possible: $mztol = 0.0024$ was chosen. Then a processing of dataset 2 was performed with the parameters defined above, i.e. $mztol = 0.0024$, $eicsig = 30000$ and $dw = 1$, which are also

EICs m/z recovery					
mztol	m/z tolerance, $\times 10^{E4}$ with reference values				
	5	10	30	50	100
0.0012	23	50	93	94	95
0.0018	24	55	94	98	99
0.0024	26	55	100	103	103
0.0030	26	59	98	102	103
0.0036	27	63	101	105	106
0.0048	31	66	101	105	106
0.0060	33	62	103	106	108

EICs m/z unicity ratio					
mztol	m/z tolerance, $\times 10^{E4}$ with reference values				
	5	10	30	50	100
0.0012	1	1	1.06	1.18	1.23
0.0018	1	1	1.03	1.12	1.17
0.0024	1	1	1.01	1.07	1.16
0.0030	1	1	1	1.05	1.13
0.0036	1	1	1	1.05	1.10
0.0048	1	1	1	1.01	1.08
0.0060	1	1	1	1.00	1.05

Table 1: Two figures of merit calculated for dataset 1. The number of different m/z values from the reference list is 109.

the default parameters. This choice was supported by similar results previously obtained on two other datasets, not reported here. Using the *EIC+* algorithm, two EIC matrices $\mathbf{X_{sa}}$ and $\mathbf{X_{sb}}$ of dimensions (2484×19958) and (2475×18773) were obtained from the five SA and the five SB MS data of Li et al, respectively. The first value was the number of retention times, the second the number of EICs. $\mathbf{X_{sa}}$ presented 800 EICs, and $\mathbf{X_{sb}}$ 799, whose m/z values corresponded to reference features, see table 2.

Dataset	$\mathbf{X_{sa}}$	$\mathbf{X_{sb}}$
Total number of EICs	19958	18773
Number of selected EICs	800	799
Unicity ratio	1.015	1.011
RMSEP of m/z values	0.00044	0.00044
DW value	0.27	0.27

Table 2: EICs extraction from SA and SB. The threshold on the m/z values between selected EICs and reference features was 0.0030.

Note that the EICs extraction did not provide information on the reference features that could be identified, 0, 1, 2 or more by EIC. To allow a comparison with the results of Li et al, features were extracted from the selected EICs of $\mathbf{X_{sa}}$ and $\mathbf{X_{sb}}$, yielding 3757 and 3531 peaks/features respectively, an average of $4-5$ by EIC. Some integration problems were due to the different EIC shapes, as observed in Figure 3. Automatic integration was easy with well-defined peaks, as in Figure 4(a) and (b). On the other hand, as in Figure 4(d), automatic integration failed because of the complexity of the signal -here a long RT-. Thus, the visualization of an EIC can be helpful to assess the quality of integration of its features. Manual integration would have recovered as accurately quantified 11 features which failed the automatic integration. However this result was not considered any more, Li et al based their benchmark on automatic integration

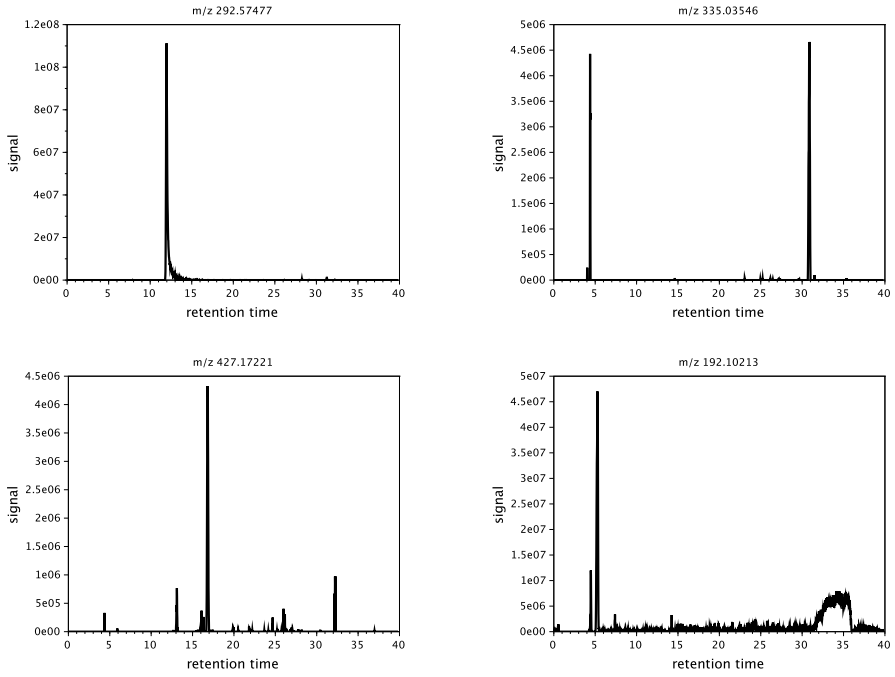


Figure 3: Dataset 2, the diversity of EICs illustrated with four examples from SA

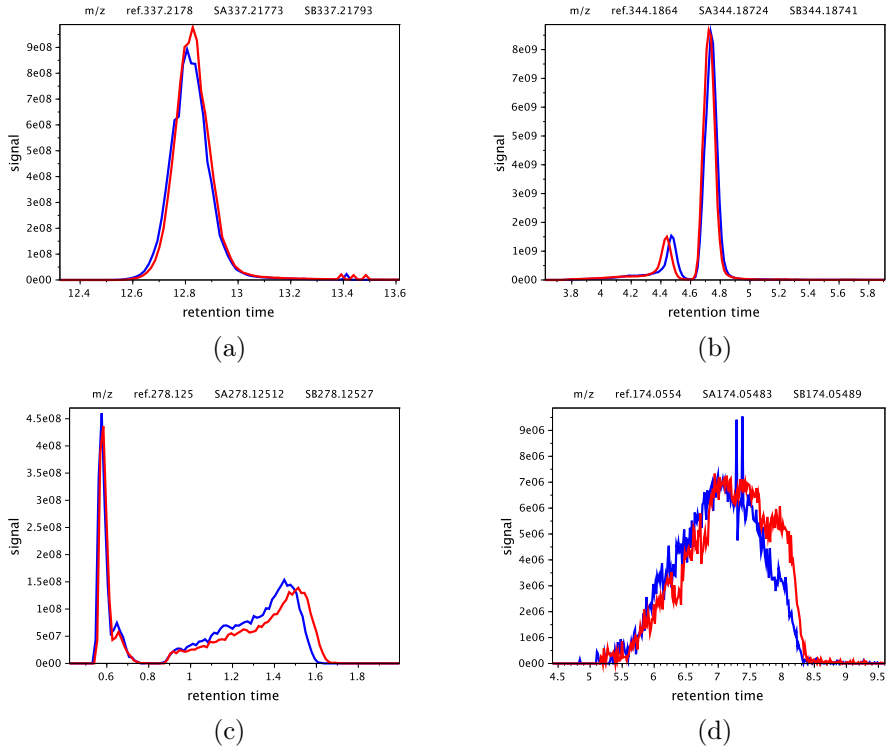


Figure 4: Dataset 2, diversity of features shapes for SA and SB in blue and red, respectively. Features (a), (b) and (c) were correctly integrated , but not for feature (d) whose RTs did not match between SA and SB.

and so did we. All the results concerning the features were made available as a spreadsheet file in the supplementary material. A summary is presented in table 3. The numbers of true features and accurately quantified true features were 820 and 787 respectively. These scores ($820 - 787$) were good compared to the results obtained by Li et al: $748 - 482$ for Compound Discoverer, $799 - 654$ for MS-Dial, $820 - 731$ for XCMS and $769 - 761$ for MZMine2. Then it is important to point out that Li et al tuned their algorithms with the same datasets analyzed in their paper, while we used another dataset. Their approach led to the best performance of each software, relevant for their benchmark but not in most non-targeted analysis, the reference compounds being unknown. Its simplicity of use constitutes a real added-value for the *EIC+* algorithm, even if the calculation takes time.

number of reference features	836
number of true features	820
number of accurately quantified true features	787
RMSEP of m/z values from SA ($N = 820$)	0.00042 u
RMSEP of m/z values from SB ($N = 820$)	0.00045 u
shift in RT	0.3 ± 0.1 mn

Table 3: Summary of the features extraction issued from datasets 2. Inspired from Li et al [2], the true features verify: $\Delta\ m/z < 0.00386$ and the reference RT fall within the experimental RT range, for both SA and SB. True features become accurately quantified when their experimental SB/SA signal ratio does not vary from more than 20 p.cent from the reference ratio. RMSEP is the root mean square error of prediction.

Some attention was also paid to the errors of prediction for the RTs and for the m/z values. RTs showed a shift of 0.30 minutes between our estimations and the reference RTs, see Figure 5(a), certainly due to a different calculation; our estimations were based on the median RTs. The RMSEP on m/z values were 0.00056 and 0.00052 for SA and SB respectively, but errors of prediction were lower for a majority of samples. From figure 5(b) corresponding to SA, more than 650 m/z values present differences under or equal to 0.0002. With such accuracy, the determination of raw formulae from experimental m/z values becomes possible.

4. Conclusion

Our *EIC+* algorithm presented a good recovery of true features and a good accuracy of the calculated m/z values, standing the comparison with XCMS, MZMine2 or Xcalibur. But its most interesting characteristic relies on a few (3) and self-explanatory tuning parameters, a threshold on the m/z values, a threshold on the signal intensities and a threshold on the noise estimated by the Durbin-Watson value. Less parameters means easier and more reproducible tunings. To give access to Galaxy users, the algorithm was deposited in the Galaxy Test Toolshed (https://testtoolshed.g2.bx.psu.edu/) where it forms a

suite of 9 tools. The Galaxy Test Toolshed eases the installation in a Galaxy web application. The code of the 9 tools can also be downloaded via a link to a Mercurial git managed by Galaxy.

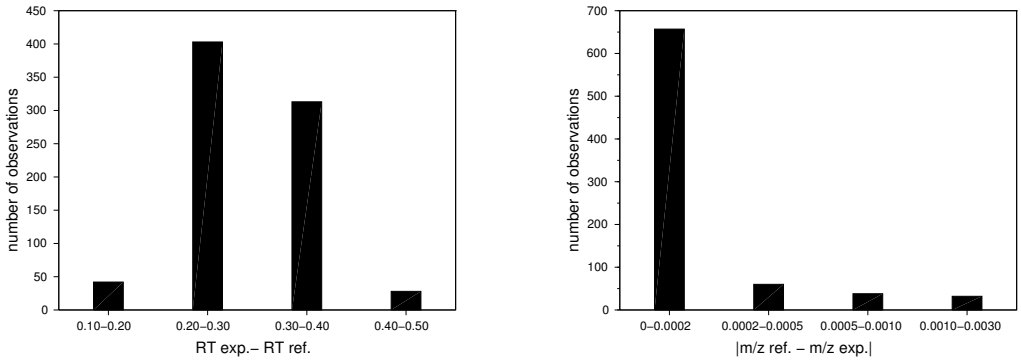


Figure 5: Number of observations in groups defined by the differences between experimental and reference values for retention times (a) and for m/z values (b)

5. Acknowledgement

This work was part of the Phenoval PRRI program funded by the Region Occitanie, France. Thanks to Emmanuelle Meudec, Pierre-Marie Allard, Arnaud Garcia, Yann Guitton and David Ropartz for useful discussions and/or providing datasets during the preliminary tests of this algorithm, and to Virginie Rossard for support about Galaxy.

6.

7. Bibliography

[1] Jean-Claude Boulet, Emmanuelle Meudec, Anna Vallverdu-Queralt, and Veronique Cheynier. Hrms: focus on the m/z values estimated by the savitzky-golay first derivative. *Rapid Communications in Mass Spectrometry*, doi:10.1002/rcm.9036, 2021.

[2] Zhucui Li, Yan Lu, Yufen Guo, Haijie Cao, and Qinhong Wang. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Analytica Chimica Acta*, 1029:50–57, 2018.

[3] Kermit K. Murray, Robert K. Boyd, Marcos N. Eberlin, G.John Langley, Liang Li, and Yasuhide Naito. Definitions of terms related to mass spectrometry. *Pure and Applied Chemistry*, doi:10.1351/pac-rec-06-04-06, 2013.

[4] Owen D. Myers, Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. Detailed investigation and comparison of the xcms and mzmine2 chromatogram construction and chromatographic peak detection methods for proprocessing mass spectrometry metabolomics data. *Analytical Chemistry*, 85:8689–8695, 2017.

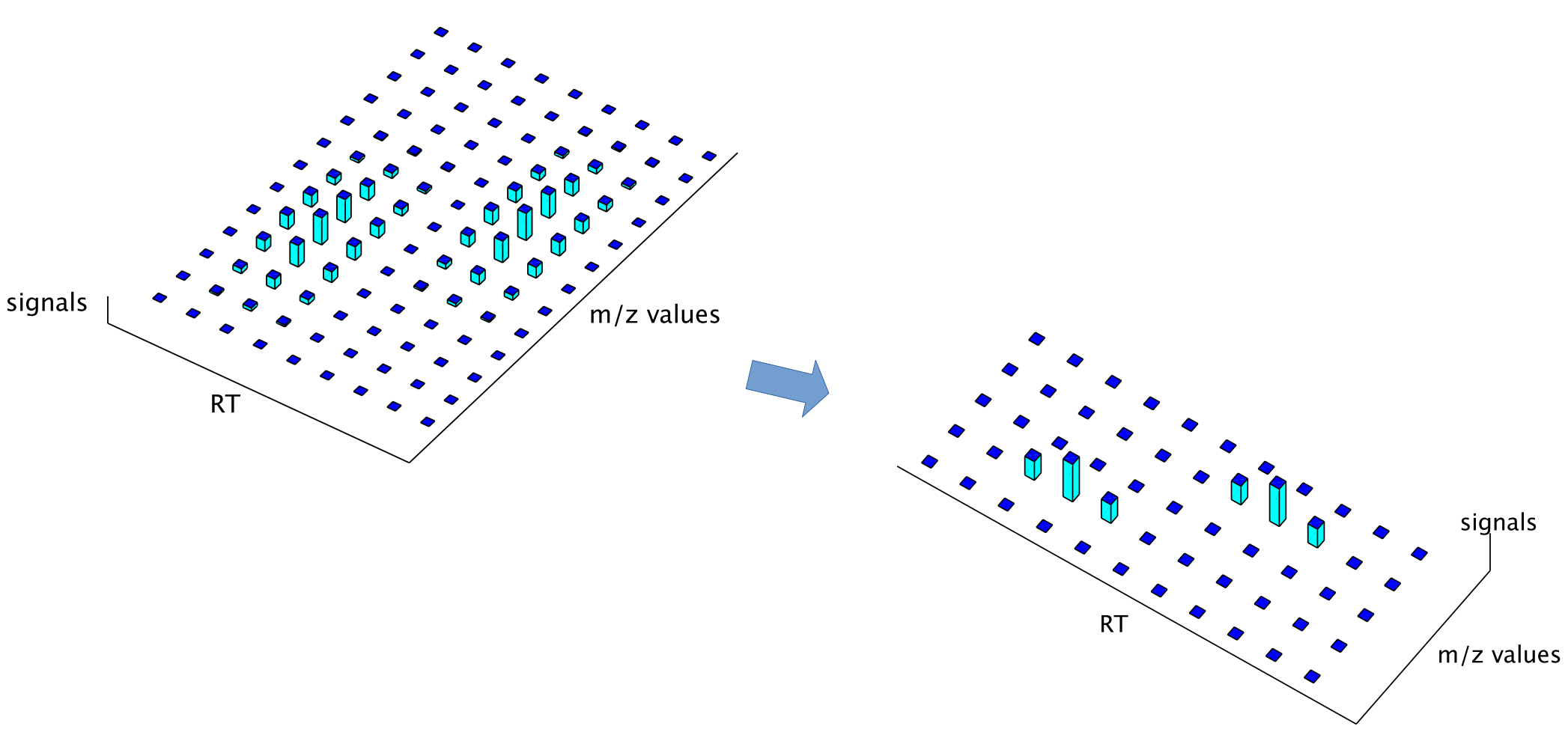
[5] Owen D. Myers, Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. One step forward for reducing false positive and false negative compound identifications form mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. *Analytical Chemistry*, 89:8696–8703, 2017.

[6] Tomas Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic. Mzmine2: modular framework for processing, visualizing and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(395):1–11, 2010.

[7] Colin A. Smith, Elisabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. Xcms: processing mass spectrometry data for metabolite profiling using non-linear peak alignment, matching, and identification. *Analytical Chemistry*, 78:779–787, 2006.

[8] Ralf Tautenhahn, Christoph Bottcher, and Steffen Newman. Highly sensitive feature detection for high resolution lc/ms. *BMC Bioinformatics*, 9:504, 2008.

[9] Anna Vallverdù-Queralt, Emmanuelle Meudec, Matthias Eder, Rosa M. Lamuela-Raventos, Nicolas Sommerer, and Véronique Cheynier. Targeted filtering reduces the complexity of uhplc-orbitrap-hrms data to decipher polyphenol polymerization. *Food Chemistry*, 227:255–263, 2017.


signals
RT
m/z values
signals
RT
m/z values