MethodsX 5 (2018) 395–402

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Extraction and cleansing of data for a non-targeted analysis of high-resolution mass spectrometry data of wastewater

Yaroslav Verkh[a], Marko Rozman[a,b], Mira Petrovic[a,c,*]

[a] Catalan Institute for Water Research (ICRA), Carrer Emili Grahit 101, 17003 Girona, Spain
[b] Ruđer Bošković Institute, Bijenička cesta 54, 10000, Zagreb, Croatia
[c] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010, Barcelona, Spain

A B S T R A C T

We provide a workflow to extract unidentified signals from chromatography-high resolution mass spectrometry (LC-HRMS) data of wastewater samples as a pre-step of a non-targeted analysis of dissolved organic matter (DOM). We provide detailed methodology on data processing and cleanup using MS processing software MZmine 2 and an own set of functions in R developed for wastewater analysis. The processing involves signal extraction, linear mass correction, reduction of noise, grouping of isotopologues, molecular formula assignment and merging of replicates. The article contains software settings and reasoning behind the choice of data extraction options. The supplementary information contains a script for the correction of signal masses using internal standards and templates of internal standard lists. We included a reproducible example as an R notebook with data cleansing workflow and data exported from MZmine. The data were used according to the described methodology in the article "A non-targeted high-resolution mass spectrometry data analysis of dissolved organic matter in wastewater treatment" by Verkh et al., 2018.

- Includes a linear mass correction algorithm for LC-HRMS signals.
- Describes a pipeline of non-targeted processing of LC-HRMS data of wastewater using free software.
- Provides tests and reasons for parameter choice in non-targeted LC-HRMS wastewater data extraction.

A R T I C L E I N F O

Method name: Combined MZmine 2.26 and R extraction workflow of LC-HRMS wastewater data
Keywords: Water screening, Molecular formula prediction, R, MZmine
Article history: Received 1 March 2018; Accepted 9 April 2018; Available online 17 April 2018

* Corresponding author at: Catalan Institute for Water Research (ICRA), Carrer Emili Grahit 101, 17003, Girona, Spain.
E-mail address: mpetrovic@icra.cat (M. Petrovic).

https://doi.org/10.1016/j.mex.2018.04.008

Specifications Table

Subject area	Chemistry
More specific subject area	Environmental chemistry
Method name	Combined MZmine 2.26 and R extraction workflow of LC-HRMS wastewater data
Name and reference of original method	Not applicable
Resource availability	http://MZmine.github.io/ (MZmine user version) https://github.com/MZmine/MZmine2 (MZmine developer version) https://github.com/Tebatu/mzmine2 (MZmine customized developer version) https://bitbucket.org/Tebatu/MZminer

Method details

Extraction and cleaning up of data

The methodology was developed and tested with LC-HRMS data recorded on an Orbitrap Velos mass spectrometer. The measurement included filtered wastewater samples and blank in triplicate recorded at a resolution of 100 000 full width at half maximum. Details on the method and pretreatment can be found in [1]. LC-HRMS data of wastewater treatment samples were extracted with a custom version of MZmine 2.26 in which the atomic ratios in formula prediction were set to H/C < 3.2, O/C < 1.2, N/C < 1.3, S/C < 0.8. The formula prediction yielded output in form of "FormX IsoY MassZ", where X is neutral formula, Y is isotopic pattern score between 0 and 1, and Z is neutral monoisotopic mass of the predicted formula in Da. Within MZmine pipeline, R was used to correct *m/z* of extracted signals using a linear model based on internal standards (IS). Exemplary IS lists for *m/z* correction in both ionization modes were stored in Supplementary Files with the required order and naming of table columns within files. The files in csv-format are readable in table format into spreadsheet editors as Microsoft Excel® or OpenOffice Calc.

MZmine 2.26 parameters in electrospray positive ionization (PI) mode

- Mass Detection: "Exact mass" detector with an intensity noise level of $1{\cdot}10^3$ a.u.
- FTMS Shoulder peaks filter: The Lorenzian extended peak model function with a mass resolution of $1{\cdot}10^5$.
- Chromatogram builder: A minimum time span of 0.03 min (2 s); a minimum signal height of $1{\cdot}10^3$ a.u.; a mass tolerance of extraction 5 ppm.
- Chromatogram smoothing: Filter width of 7 points.
- Chromatogram deconvolution: Noise amplitude algorithm was applied with a signal duration of 0.05–8.00 min, a minimal signal height of $3{\cdot}10^3$ a.u. and an amplitude of noise $1{\cdot}10^3$ a.u.
- Custom database search for IS: mass tolerance <5 ppm, retention time tolerance <0.3 min. A visual check of the IS signals confirmed their validity.
- A personal script in R corrected the *m/z* of signals in exported XML peak lists using linear models based on IS. The linear models were calculated for two *m/z* ranges over and under 400 Da. The script removed IS outliers >4* interquartile range and omitted the correction where less than 3 IS signals defined the model. Exemplary R script "mzCorrectionWithIs.R" and IS list "internal_standards_-for_mz_correction_PImode.csv" that is used within can be found in Supplementary Files.
- Isotopic peaks grouper: 60 ppm mass tolerance, 0.03 min of time tolerance and a maximum charge of 2, representative peak: lowest *m/z* and monotonic shape.
- Filters: peaks in an isotope pattern ≥2, peak per row ≥1, points per peak 7–500.
- Adduct search: retention time tolerance <0.03 min, given a prevalent ionization $[M+H]^+$ the adducts in PI LC-ESI-MS are: $[M + NH_4]^+$, $[M + Na]^+$, $[M + 2Na]^{2+}$, $[M + K]^+$, $[M + 2K]^{2+}$, $[M + CH_3OH]^+$, mass tolerance 5 ppm and maximum adduct signal height of 50%. Subsequently these adducts were removed from the feature list.