Chemometrics and Intelligent Laboratory Systems 194 (2019) 103866

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

RMet: An automated R based software for analyzing GC-MS and GC×GC-MS untargeted metabolomic data

Saeed Moayedpour, Hadi Parastar *

Department of Chemistry, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Metabolomics
Chemometrics
GC×GC
GC-MS
R language

ABSTRACT

Gas chromatography-mass spectrometry (GC-MS) and comprehensive two-dimensional gas chromatography-mass spectrometry (GC×GC-MS) are powerful techniques for measurement of all metabolites in complex metabolic samples. However, analyzing GC-MS and especially GC×GC-MS metabolomic data is a major challenge to the researchers in the field of metabolomics mainly due to the complexity and large data size. In this regard, an automated R based software entitled RMet has been developed to overcome the challenges in the metabolomic analysis workflow of GC-MS and GC×GC-MS data sets. Additionally, it is able to facilitate the complex process of extracting reliable and useful biological information from these data sets. Moreover, RMet can greatly accelerate the time-consuming data analysis process of large GC-MS and GC×GC-MS datasets by the means of modern chemometric methods. In fact, RMet transforms raw GC-MS and GC×GC-MS data files into the elution profiles and mass spectra of important (significantly affected metabolites) which can be imported into NIST MS search software for the final identification of these metabolites. To show the performance of the developed software, large GC×GC-MS data sets of a previously reported environmental metabolomics study on lettuce samples exposed to contaminants of emerging concerns (CECs) were analyzed by RMet. The procedure for analyzing GC-MS metabolic data with RMet is as same as GC×GC-MS data sets but some steps can be skipped due to the lower size of GC-MS data sets. The software, its manual, sample data sets and source code are freely available on https://github.com/SUTChemometricsGroup/RMet.

1. Introduction

Metabolomics is the comprehensive study of all metabolites in a cell, tissue, or an organism in order to produce a metabolic snapshot of a biological system [1]. Metabolomic samples are mostly of high complexity due to the presence of numerous metabolites with specific physicochemical properties in the metabolome. This complexity is illustrated by the number of metabolites and phytochemicals in the plant kingdom which is estimated to be greater than 200000 [2]. As a result, measurement of all metabolites in such complex sample matrices requires the use of multiple sophisticated analytical instruments and remains an analytical challenge. Among different analytical platforms, gas chromatography-mass spectrometry (GC-MS) is the more frequently used technique for separation and identification of metabolites. However, the complexity present in most of the biological samples pushes this technique to its limits. The comprehensive two-dimensional gas chromatography-mass spectrometry (GC×GC-MS) is a great solution to overcome this challenge [3]. The GC×GC instrument separates the greatest number of metabolites with excellent sensitivity and, when combined with a fast mass spectrometry detector such as time-of-flight (TOF), provides an exceptional metabolite identification power [4]. The GC×GC technique has various advantages over GC such as improved resolution, increased separation capacity, better signal to noise ratios which lead to enhanced analyte detectability, and the ability of chemical class ordering in the 2D total ion chromatogram (TIC) [3]. However, there are significant challenges in the process of obtaining desired information from GC×GC-MS data, mostly due to the large volume of the produced data (e.g. typically in gigabytes (GB) per sample) [4]. This issue will certainly emerge for untargeted metabolomics studies in which there is a need to run many samples from at least two sample classes with a minimum of three replicate runs for each sample.

Chemometric techniques based on multivariate data analysis can properly tackle the problems surrounding large metabolomics data sets

* Corresponding author.
E-mail address: h.parastar@sharif.edu (H. Parastar).
URL: http://sharif.edu/%98h.parastar (H. Parastar).

https://doi.org/10.1016/j.chemolab.2019.103866
Received 10 June 2019; Received in revised form 20 September 2019; Accepted 6 October 2019
Available online 7 October 2019
0169-7439/

[5]. Multivariate curve resolution-alternating least squares (MCR-ALS) is a frequently used multivariate resolution method for decomposition of measure mixed analytical signals into the contribution profiles of pure constituents using a bilinear data decomposition. Combination of an appropriate compression strategy such as wavelet transform with MCR-ALS is a perfect solution for approaching the problems surrounding metabolomics data volume [5].

Different software tools have been developed for analysis of metabolomic data and for obtaining qualitative and quantitative information. Currently available omics tools such as Metabolyzer [6], PlantMat [7], MetExpert [8], MetExtract II [9,10], Lipostar [10], IMMA [11], FlavonQ-2.0v [12], and MetaboliteDetector [13] are only able to analyze Liquid Chromatography-Mass Spectrometry (LC-MS) and GC-MS metabolomic data. Although a number of data analysis methods are developed for processing GC×GC-MS data, to date there is no software that combine all required steps for analyzing GC×GC-MS metabolomics data. Thus, there is an essential necessity for a comprehensive user-friendly software which is specifically designed for omics studies in order to facilitate and speed up time-consuming data mining process of complex GC-MS and GC×GC-MS metabolomic data. Such software can popularize the application of GC-MS and GC×GC-MS techniques combined with novel chemometric algorithms and modern statistical approaches among the researchers in the field of metabolomics. In this regard, it can provide them with a large amount of new useful information about their studied biological system.

In order to meet this demand, we have developed RMet, an automated R based user-friendly graphical user interface (GUI) that aims to overcome challenges during the analysis of complex and big metabolomic GC-MS and GC×GC-MS data sets for transforming them to proper biological information. This software includes all steps of a complete untargeted metabolic data analysis work including preprocessing, segmentation, data compression, multivariate curve resolution (MCR), important metabolites identification, and metabolites classification.

Briefly, RMet applies MCR-ALS [14–17] as its resolution algorithm which is one of the most efficient ways to handle fundamental challenges that occur during GC×GC analyses such as elution time shifts, baseline/background contribution, peak overlap, and peak shape changes [3]. Then, it performs metabolite classification by building a partial least squares-discriminant analysis (PLS-DA) [18,19] and finally introduces the significantly affected metabolites using variable importance in projection (VIP) [20] scores. Afterward, metabolic pathway analysis should be performed using metabolic pathways databases in order to identify affected metabolic pathways.

The workflow is designed in such a way that enables individuals to perform a complete analysis and obtain appropriate results without having any knowledge of chemometrics, but simultaneously provides detailed statistics for experts who want to customize and optimize their analysis. Preprocessing and Segmenting of large GC×GC-MS data is a tedious process; a great deal of attention is required while segmenting data or removing redundant areas such as column bleeding and derivatization agents from the TIC if it is desired to manually modify the matrices, but all these operations are easily done in RMet by just a few clicks. Also, RMet is low-size software written in R programming language which is an open source language with high popularity among the statisticians and data scientists [21]. These novel features make RMet a dominant automated computational tool for analyzing GC×GC-MS metabolomics. It should be pointed out that RMet can be used for the analysis of GC-MS metabolomic data too. In the following sections, the RMet's function is demonstrated in data processing of a previous environmental metabolomics study on lettuce samples exposed to contaminants of emerging concern (CECs) by GC×GC-TOFMS which aims to investigate the effect of CECs exposure of lettuce on its metabolic pathways [3].

2. Experimental

2.1. Software development

RMet is developed under RStudio version 1.1.383 using R core version 3.4.3, its execution file can be run in Windows environment without any limitation on the version and it is available free of charge at https://github.com/SUTChemometricsGroup/RMet along with a manual, source codes and sample data sets. In order to use RMet on the Linux and Macintosh operating systems, one should install R core (freely available at https://cran.r-project.org) and run the RMet.R code which is available at the previously mentioned Github link. For this GUI, R packages of MASS, ALS, wavelets, rgl, hash, fields, Matrix, irlba, mixomics, gWidgets2, gWidgets2RGtk2, RGtk2, RNetCDF, and abind were utilized. Minimum required computer configuration depends on the size of input data, but a desktop computer with 8 GB of RAM and a 3 GHz processor can easily analyze up to 20 GB of GC×GC-MS data. The installation steps of RMet are shown in Figs. S1–S8 of supporting information (SI).

2.2. Real GC×GC-TOFMS data sets

To test the performance of the developed software, GC×GC-TOFMS data sets of extracted metabolites from four control lettuce samples and four lettuce samples exposed to eleven CECs were used [3]. Data sets are available at https://github.com/SUTChemometricsGroup/RMet. In this study, the lettuce (*Lactuca sativa* L) samples were irrigated with water contaminated with 50 μg/L concentration of 11 CECs (pharmaceuticals, personal care product, anticorrosive agents, and surfactants) for 34 days under controlled conditions. Then, the crops were harvested and their metabolome was extracted. In order to extract the metabolites from the lettuce leaves, 60 mg of plant materials, 400 μL methanol, 30 μL of a 50 μg/mL D-glucose solution, and 30 μL of a 50 μg/mL salicylic acid solution were vortexed and sonicated in an Eppendorf tube. Next, the solution was vortexed with 200 μL of chloroform and 400 μL of water respectively. Finally, it was centrifuged and the aqueous phase was transferred to a vial. The extracted metabolome of leaf samples was derivatized with tetramethylsilane (TMS) and analyzed by GC×GC-TOFMS (HP 6890 N (Agilent Technologies, Palo Alto, CA)) in optimum conditions. The readers encourage to read ref. [3] for more details.

The analysis of eight metabolome samples leads to the generation of eight CDF files with a total size of 8.6 gigabytes (GB). All datasets were directly imported to RMet in order to perform analysis and identify the important (significantly affected) metabolites.

3. Results and discussion

RMet's data processing strategy is shown in Fig. 1. It is a specifically designed data processing platform for analyzing both GC×GC-MS and GC-MS metabolomic data sets. Following, the applied approaches and algorithms in each step of the RMet workflow will be discussed in detail during the analysis of GC×GC-MS data of control and CECs exposed lettuce samples. These data sets were used to demonstrate the functionality and output of each data mining step.

3.1. Data import

In this section, the user can upload both GC×GC-MS and GC-MS data in CDF, CSV, and Rdata formats. In order to create the data matrix for GC×GC-MS data from raw CDF file, the instrument modulation period and the detector frequency is required. After importing data, two- and three-dimensional (2D and 3D) visualization of TIC are available. For example, Fig. 2 shows the 3D visualization of one of the control lettuce sample. As previously mentioned, four control and four exposed samples were recorded with a modulation time of 4 s and a detector frequency of

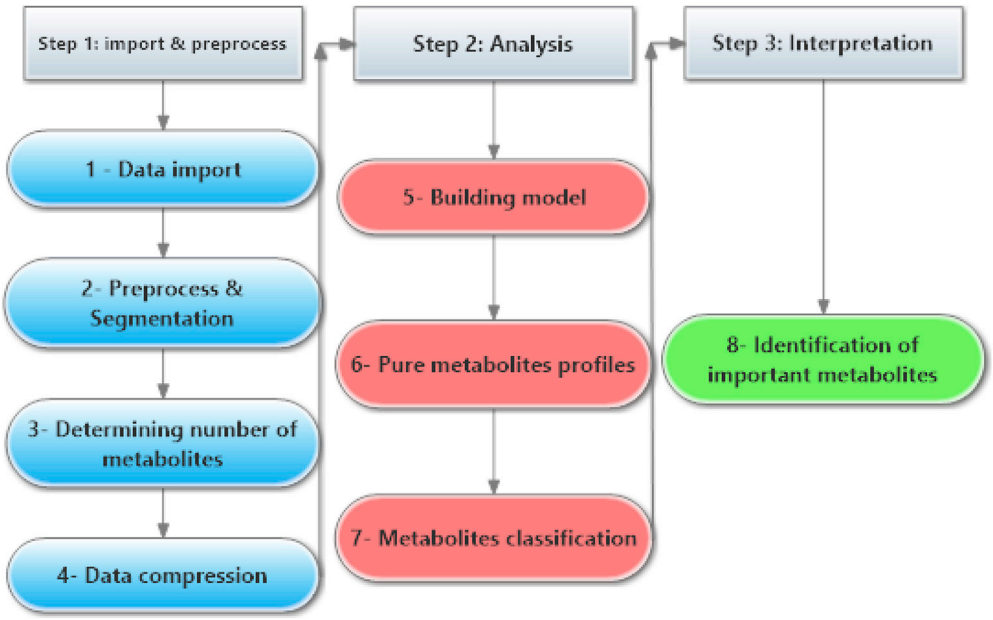

Fig. 1. Schematic diagram of RMet workflow.

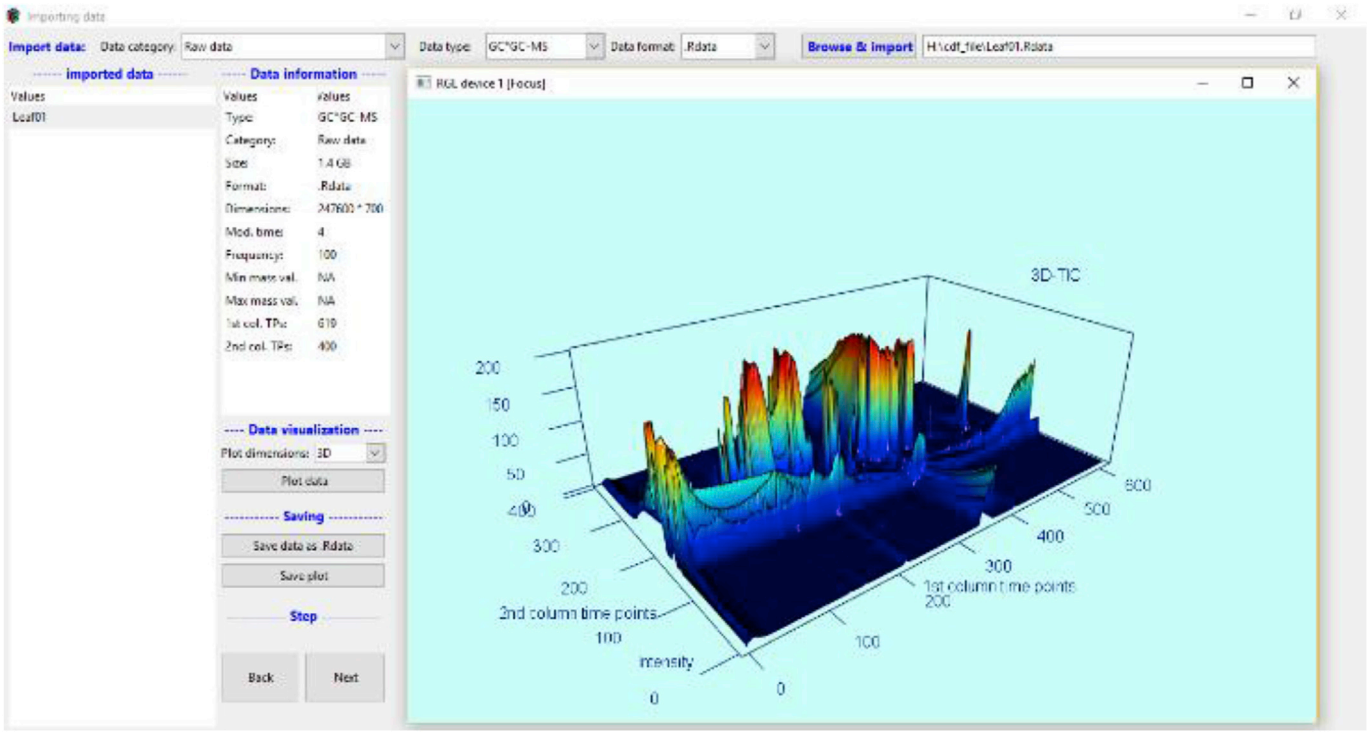

Fig. 2. RMet data importing page.

100 Hz. Considering these parameters and the CDF vector's length, the TIC of each sample was a 400×619 matrix. It is worth mentioning that the total size of imported datasets was 8.6 GB. Also, Figs. S9–S12 show how to start analysis and import data into RMet.

3.2. Preprocess & segmentation

This section greatly facilitates the data segmentation and removal of the redundant parts in the chromatogram including bleeding and/or derivatization agents parts. The user can view the TIC of a selected matrix with the different intensity ceiling for better area selection. Segmentation is included in the software for decreasing the data size. Additionally, the calculation efficiency can be improved by selecting segmentation operation. Removing the column bleeding, column overload, and derivatizing agent areas from TIC is also available by selecting the vertical and horizontal removal operations. All selected matrices will be modified by clicking on the "operate" button (Fig. 3). In the case under study in this work, the 8 imported datasets were divided into two segments. Both segments cover the same range of second column time points (from time points 177 to 400 in order to exclude the column bleeding area). In

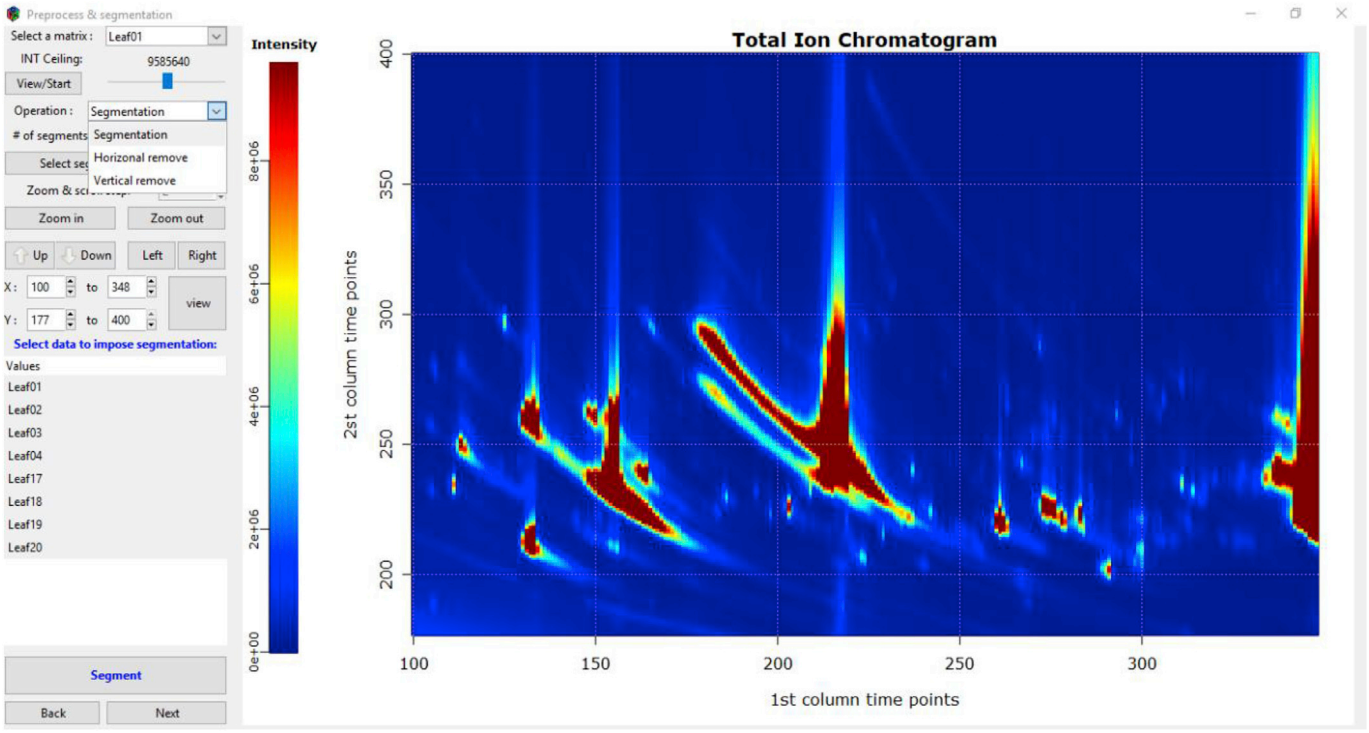

Fig. 3. Preprocess and segmentation step in RMet.

addition, the coverage of first column time points in the first segment (i.e., segment A) was from 100 to 348, and from 358 to 619 for second segment (i.e., segment B). Fig. 3 shows the 2D contour plot for segment A of the GC×GC-MS TIC of one of the control lettuce samples. More details about the preprocessing and segmentation part of the RMet software can be found in supporting information (SI) (section S2.2, Figs. S13–S19).

3.3. Determining the number of metabolites

Here different segments for control and exposed samples can be augmented in the column-wise or row-wise way for simultaneous analysis by MCR-ALS model. Since determining the number of components is required for MCR, singular value decomposition (SVD) is used to determine the number of metabolites in the augmented data matrix. Fig. 4

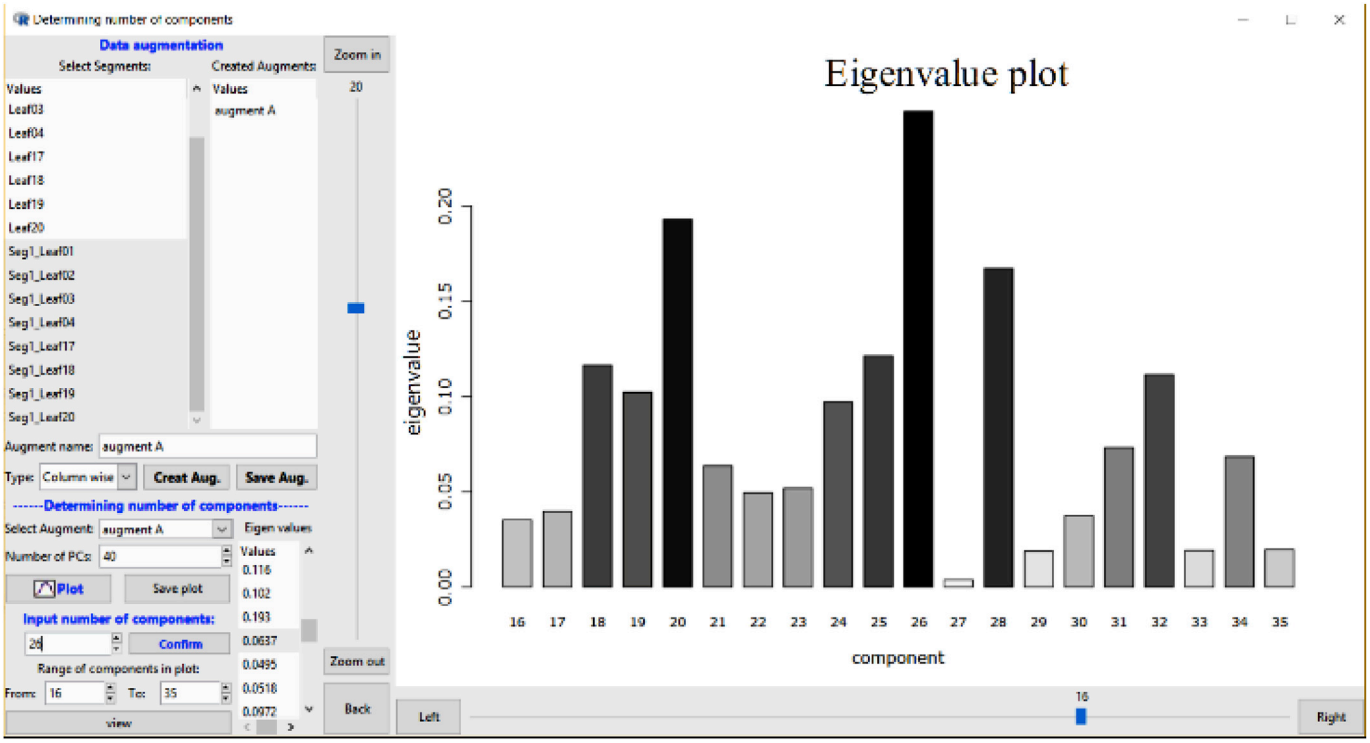

Fig. 4. Metabolites number determination in RMet.

depicts the component determination page of RMet. By creating and selecting an augmented matrix, the user can see the plot containing logarithm of singular values ratios. In this plot, zooming and scrolling are provided for precise estimation of the number of components (Fig. 4). Both segments A and B were arranged into column-wise super-augmented matrices named augment A and augment B. According to the SVD ratio plot (Figs. 4), 26 metabolites were considered for augment A and 32 metabolites for augment B. More details can be seen in Figs. S20–S21.

3.4. Data compression

As it has already been discussed, the big size of GC×GC-MS data sets is a challenging issue. Although segmentation can help, it is not an ideal solution because increasing the number of segments can lead to errors in the classification step (due to the separation of some metabolites' chromatographic profiles into two different segments). Therefore, a proper compression algorithm is required to reduce data size and make the analysis more feasible. RMet performs a discrete wavelet transform [22,23] in the time direction in order to compress the augmented data matrix. Indeed, increasing the level of compression leads to losing a greater portion of data. Therefore, finding an optimized compression level is clearly important. In this regard, RMet provides the user with the final data size and chromatogram created by imposing different compression levels. By doing so, one can select the proper compression level based on the pattern of the resulted chromatogram, the portion of negative values, and the system's computing power. In our study, both augmented data matrices were compressed by level 3. In other words, their size was reduced from 2.76 GB to 345 MB for segment A and from 2.91 GB to 364 MB for segment B. As an instance, Fig. 5 demonstrates compressed GC×GC-MS data sets for segment A. Also, Figs. S22–S24 depicts more details about data compression in RMet.

3.5. Building a model

In this step, the MCR model is built to obtain pure chromatographic and spectral profiles of metabolites. The MCR-ALS is one of the most powerful resolution methods to resolve complex hyphenated and multi-dimensional chromatographic data [17,24] (e.g., GC-MS and GC×GC-MS), while also handling fundamental challenges during chromatographic analyses such as elution time shifts and peak shape changes. The MCR-ALS method is based on the fulfillment of the bilinear model and therefore, chromatographic data should be arranged in an augmented data matrix. For example, GC×GC-TOFMS data sets of control and exposed lettuce samples can be arranged in a column-wise super-augmented data matrix with mass-to-charge ratios (m/z) as columns and elution times in first- and second-chromatographic columns as rows of this data matrix. This data augmentation provides the MCR-ALS method with two outstanding advantages. First, it can perfectly handle unavoidable chromatographic challenges such as shifts in retention time within and between GC×GC-TOFMS chromatographic runs, as the m/z values are similar among all measured spectra in all second-column modulations. The second advantage is the capability of performing the simultaneous analysis. In fact, MCR-ALS benefits from a great flexibility to consider all samples (standard, unknown and replicates) in a single super-augmented data matrix, even if the number of rows (retention times) varies among the individual data matrices. Furthermore, adding extra components to the MCR-ALS model enables the modeling of baseline/background contributions. In this section of the RMet software, the user can select the desired configuration such as constraints, convergence criterion, and the maximum number of iterations to build MCR model. After choosing the desired configuration, MCR modeling can be done by pushing the "Run" button in RMet software (Fig. S25). After performing MCR analysis, the output R file can be easily uploaded to the software in order to proceed to the next steps. The output of the MCR-ALS modeling step is the resolved pure elution and spectral profiles for metabolites and some statistical parameters such as lack of fit (LOF) and noise level. In our study, the resulted LOF and noise level for MCR model of augment A was 2.1% and 2.2% respectively. For segment B, these values were 2.5% and 2.8% respectively. All of these values were acceptable according to the agreement between LOF and noise level.

3.6. Pure components profiles

In this step, the user can view the output of MCR-ALS including

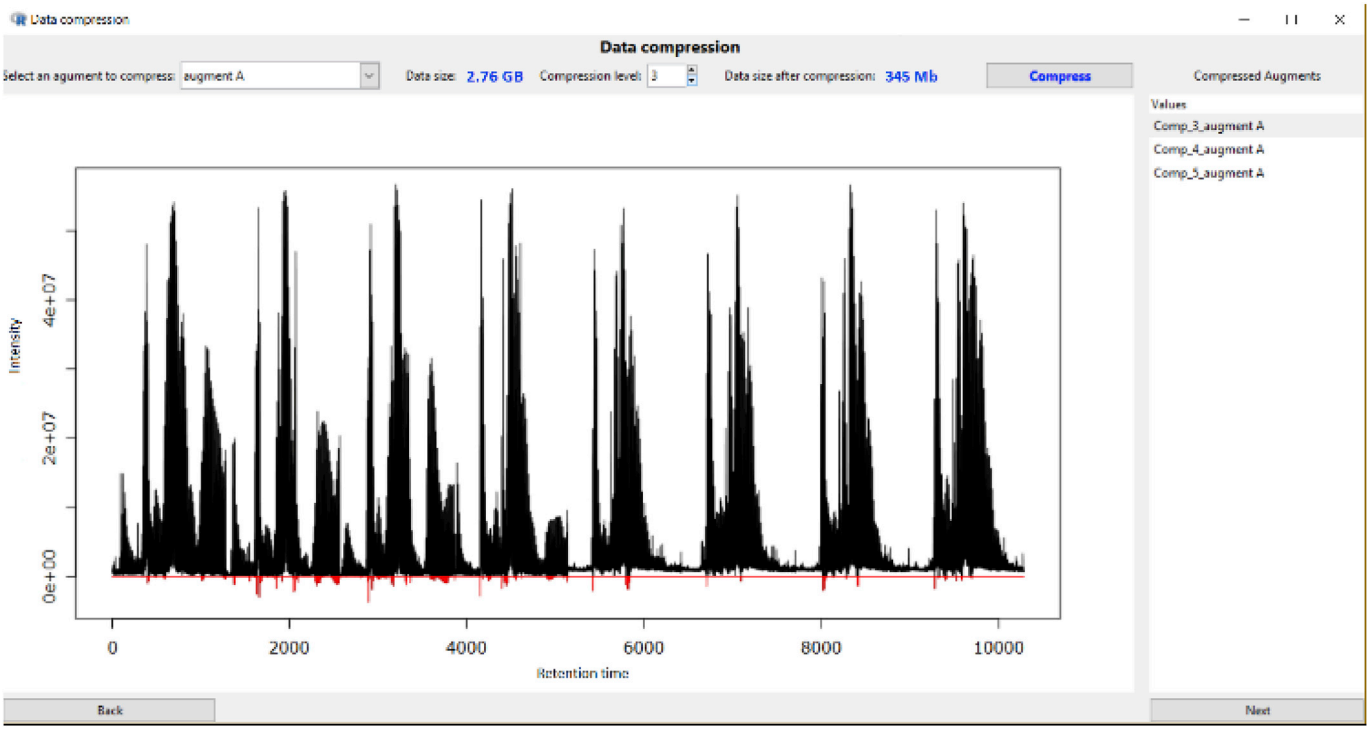

Fig. 5. Data compression step in RMet.

resolved elution profiles in chromatographic dimension for GC-MS and in two chromatographic dimensions for GC×GC-MS data along with resolved mass spectra. Fig. 6 shows the resolved mass spectra. The software can also display the elution profile of a single metabolite in different samples (e.g., control and exposed) in order to see the chromatographic variations of selected metabolite among the control and exposed samples (see Figs. S26–S27 for more details).

3.7. Supervised classification of samples based on their metabolites

After obtaining pure elution profiles for metabolites, the peak areas of the resolved components in different samples can be calculated in RMet. Therefore, a new data matrix is created in RMet with samples as rows and resolved metabolites as columns of this data matrix. Among different linear and non-linear classification methods, it is possible to use different methods like PLS-DA and/or orthogonal projection to latent structure-discriminant analysis (OPLS-DA) as a linear models and kernel based methods and/or support vector machine (SVM) as non-linear model [25]. However, as PLS-DA can be considered as a more frequently used method, therefore, this method is included in RMet to classify the identified metabolites based on their relative concentration (i.e., peak areas). The user should define the numbers of classes by selecting all samples belonging to a determined class and pressing the "create class" button. By doing so, the software automatically generates the **X**-block which contains the relative concentration of all resolved metabolites in all samples and the **y**-block that is a vector indicating which samples are in the defined classes. After selecting the proper preprocessing (mean--centering, scaling, auto-scaling), the PLS-DA classification model is built for training set and evaluated using cross-validation (Fig. 7).

The **X**-block cumulative variances for selecting the proper number of latent variables (LVs) and other required statistics for model validation is also provided by the software (Fig. 7). Please see section S2.7 in SI to get more details about PLS-DA model and its features. In our study, auto-scaling was selected as the preprocessing method and the classification model was built using 4 LVs. As shown in Fig. 7, control and exposed samples are completely separated in PLS space. Also, Fig. 8 shows some of the statistical parameters (**X**-residuals, regression coefficients, correlation circle plot, and cumulative variance) to evaluate the validity of the classification model for training set. As it can be seen from this Figure, all of these plots confirm the validity of the developed classification model. Also, RMet enables users to perform classification on test data after building the model and it provides related statistical information [26]. See section S2.7 and Figs. S28–S32 in SI for more details.

3.8. Identification of important metabolites

Here we introduce the last step of the RMet workflow which is the identification of important metabolites. The variable importance in projection (VIP) scores show the influence of each metabolite on the PLS-DA model, therefore they are very appropriate for determining the important (significantly affected) metabolites among the resolved ones. In this section, VIP scores are computed and represented for each metabolite using the PLS weights associated with each LV. In this software, the "greater than one rule" is used as a criterion for selection of the important metabolites [20]. Finally, the *NIST MS search* compatible text files containing mass spectrum information of the important metabolites is exported to a user-selected folder that enables the identification of the important metabolites (Figs. S33–S35). At this stage, RMet has perfectly performed its duty, which was transforming raw GC×GC-MS data file into a recognition of metabolic changes in the system. With important metabolites in hand, one can use metabolic pathways databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [27,28]. database in order to perform a pathway analysis and determine the functional modules that included at least two of the identified metabolites so that the affected metabolic pathways will be identified. In our study, 26 metabolites were found to be significantly affected as a result of exposure of the plant to the CECs. Table 1 shows these important metabolites and their VIP scores. Alterations in concentrations of these important metabolites reveal that exposure of lettuce to the mentioned CECs cause significant changes in various metabolic pathways of the plant such as carbohydrate metabolism, the citric acid cycle, pentose phosphate pathways, and glutathione pathway.

As it has already mentioned in the text, the same procedure can be performed for the analysis of GC-MS data using RMet and the only different section is the "Segmentation & preprocessing" section. For GC-MS data matrices, it can only select a range of retention times by inputting the desired range in spin buttons and clicking on the "Impose button" (Fig. S36). All the following steps are the same as GC×GC-MS procedures but the compression can be skipped since GC-MS data sets have usually low sizes. As an example, GC-MS data of metabolic profiles of daphnia magna exposed to salinity [29] were tested using RMet and can be found in https://github.com/SUTChemometricsGroup/RMet.

4. Conclusion

Development of new integrated software for fast and accurate analysis of large GC-MS and GC×GC-MS untargeted metabolic data set can

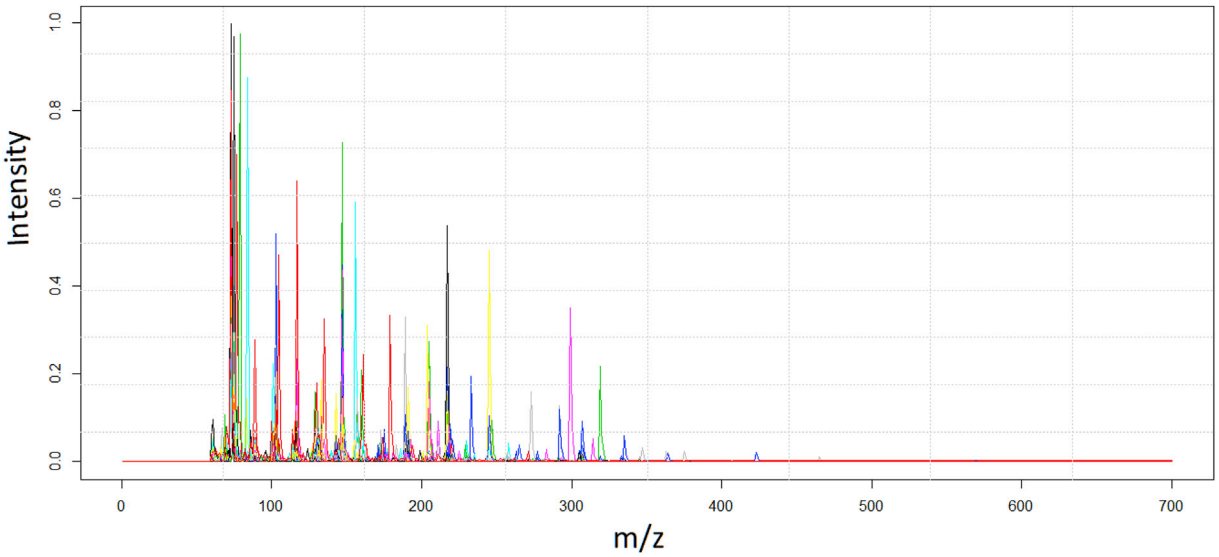

Fig. 6. Resolved mass spectra of all metabolites.

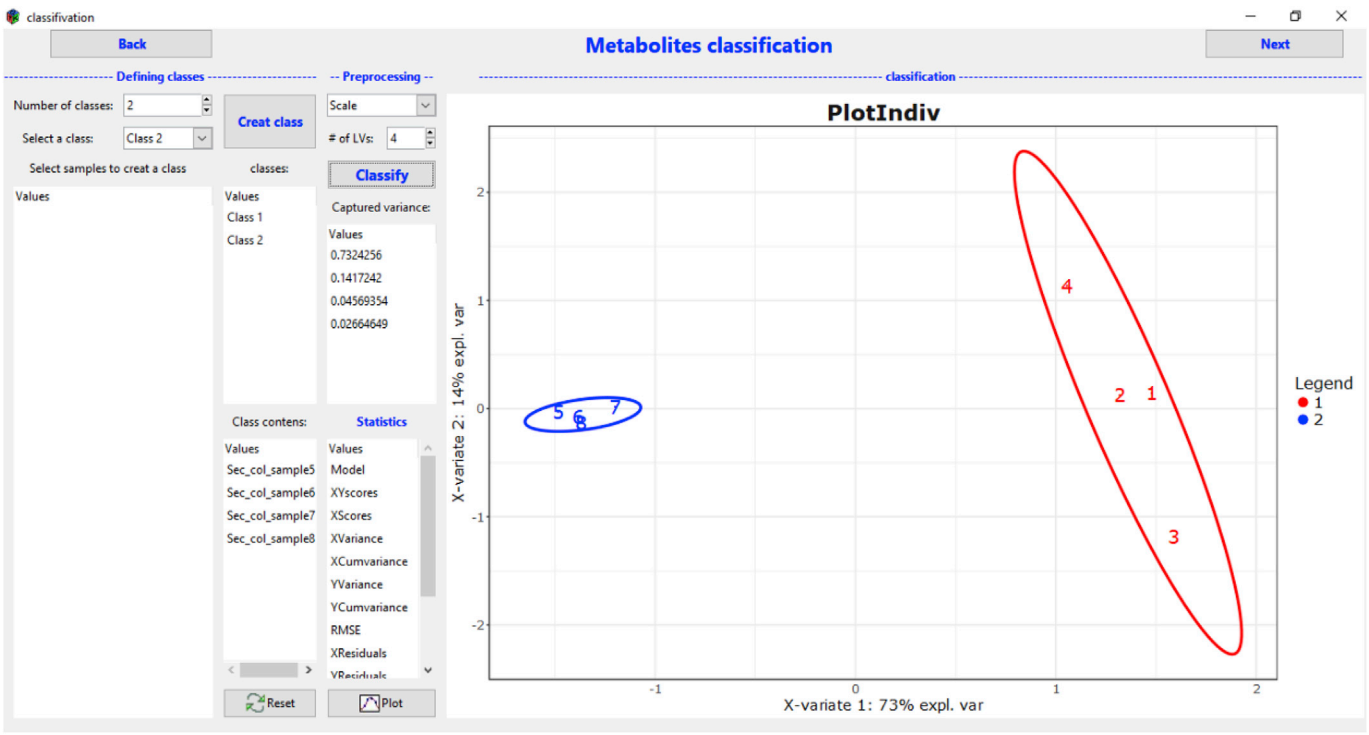

Fig. 7. Metabolites classification using PLS-DA in RMet.

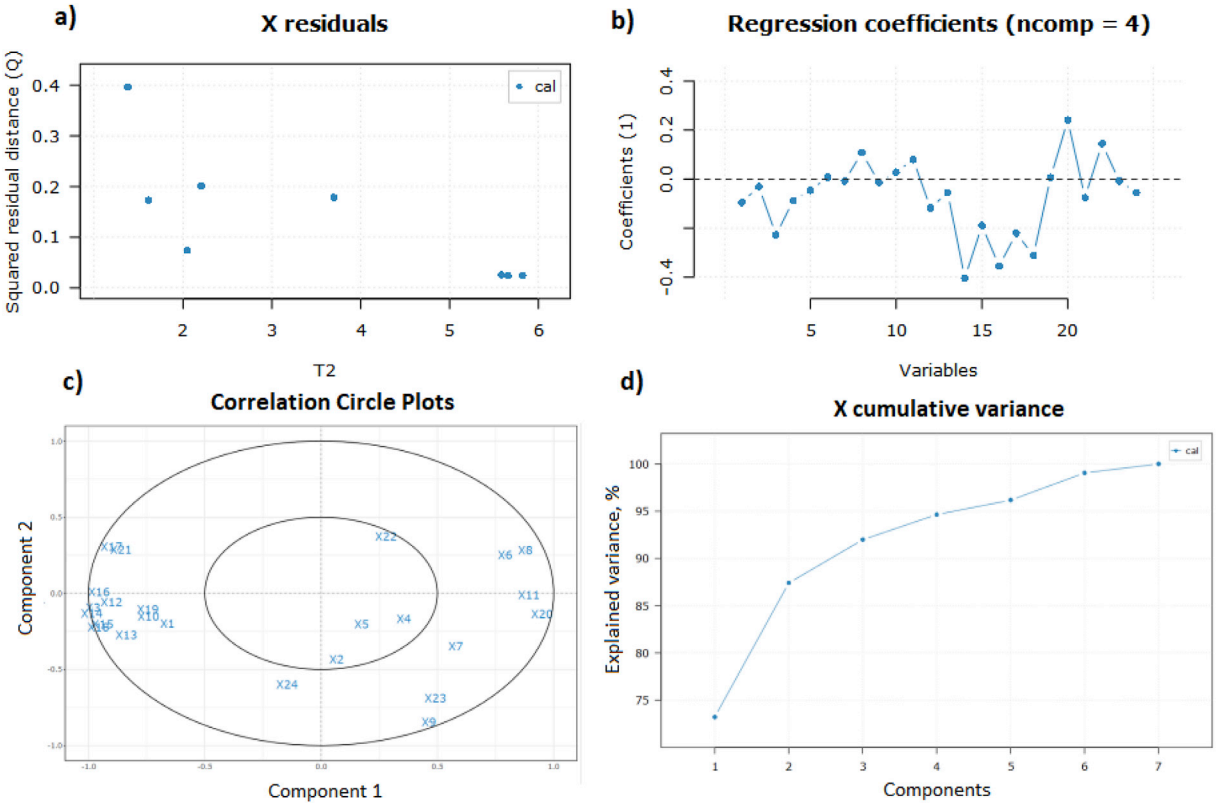

Fig. 8. Some statistical information of resulted PLS-DA model for classification of resolved metabolites. a) X-block residuals. b) Regression coefficients considering 4 LVs. c) Correlation circle plots. d) Latent variables' cumulative variance plot.

Table 1
Identified important metabolites and their VIP scores.

Metabolite	VIP score	Metabolite	VIP score	Metabolite	VIP score
L-Proline, 5-oxo-1	1.35	Tartaric acid	1.24	Prostaglandin F-2 β	1.36
Butane, 2,3-diol	1.10	Methylmalomic acid	1.16	D- mannose	1.21
Maleic-acid	1.01	Propanedioic acid	1.33	Acetic acid	1.03
Glycerol	1.21	Myo-inositol	1.16	Xylonic acid	1.13
Lyxose	1.24	Allo-inositol	1.13	Ribitol	1.12
Citric acid	1.17	Sylo-inositol	1.04	Sorbit	1.25
Benzoic acid	1.36	D-Glyceraldehyde	1.28		
Succinic acid	1.34	Ribonic acid	1.14		

extremely help the researches in the field of metabolomics. In order to meet this crucial need, we have developed RMet, a novel automated R based software for analyzing both GC-MS and GC×GC-MS untargeted metabolomic data sets in a simple, quick, and reliable manner. All required steps for completing a metabolomics data analysis workflow including data preprocessing and segmentation, data compression, determining the number of metabolites, obtaining pure elution profiles and mass spectra of all metabolites, recognition and classification of important metabolites, and exporting the mass spectra of important metabolites for further identification with mass spectrometry libraries such as NIST are perfectly covered in RMet. RMet utilizes the MCR-ALS which is a powerful method for decomposition of raw GC-MS and GC×GC-MS data into the pure elution profiles and spectra of metabolites. The greatest limitation of MCR-ALS analysis is the need for high computing power. RMet overcomes this limitation by providing a proper compression strategy before the resolution step which can reduce data size without losing important information. These novel specifications in addition to its user-friendly environment make RMet a useful data mining tool for GC-MS and GC×GC-MS based untargeted metabolomics studies.

4.1. Independent testing

Prof. Mehdi Jalali-Heravi.

Chemistry and Biochemistry Department, California State University, Los Angeles, Phone: (949) 466 4766, Email: mjalali2@calstatela.edu.

RMet software is a very interesting piece of work, especially in opening a new window for carrying metabolomics studies using GC-MS and GC×GC-MS techniques. I confirm all abilities of this software as authors described in the manuscript. This software helps those researchers who are interested to work with metabolomic data. This software is able to implement the common data size reduction method of wavelet transform and data segmentation method as well. Also, MCR-ALS algorithm is included in this software for resolution of mixed chromatographic signals into the pure elution and mass spectral profiles. Furthermore, PLS-DA model was included for classification of resolved metabolites. For each model, the statistical parameters can be viewed. One of the advantages of this software is that it is very easy to be installed and its applying is very simple and as mentioned in the manuscript no needs to be expert in programming. I believe that regarding the easiness and simplicity of this software, all chemists who are interested in RMet can use it.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the Research Council of Sharif University of Technology (SUT) for the financial support of this research with grant no. G960613. They would like to thank Prof. Josep M. Bayona from IDAEA-CSIC institute in Barcelona (Spain) to get access to the metabolomic data sets used in the work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.103866.

References

[1] O.A. Jones, Metabolomics and Systems Biology in Human Health and Medicine, Cabi, UK, 2014.
[2] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding, Anal. Chim. Acta 879 (2015) 10–23.
[3] C. Hurtado, H. Parastar, V. Matamoros, B. Piña, R. Tauler, J.M. Bayona, Linking the morphological and metabolomic response of Lactuca sativa L exposed to emerging contaminants using GC× GC-MS and chemometric tools, Sci. Rep. 7 (2017) 6546–6558.
[4] S. Yang, J.C. Hoggard, M.E. Lidstrom, R.E. Synovec, Gas Chromatography and Comprehensive Two-Dimensional Gas Chromatography Hyphenated with Mass Spectrometry for Targeted and Nontargeted Metabolomics in "Metabolomics in Practice: Successful Strategies to Generate and Analyze Metabolic Data, Wiley, USA, 2013.
[5] H. Parastar, R. Tauler, Big (Bio) chemical data mining using chemometric methods: a need for chemists, Angew. Chem. Int. Ed. (2018), https://doi.org/10.1002/anie.201801134.
[6] T.D. Mak, E.C. Laiakis, M. Goudarzi, A.J. Fornace Jr., Metabolyzer: a novel statistical workflow for analyzing postprocessed lc–ms metabolomics data, Anal. Chem. 86 (2013) 506–513.
[7] F. Qiu, D.D. Fine, D.J. Wherritt, Z. Lei, L.W. Sumner, PlantMAT: a metabolomics tool for predicting the specialized metabolic potential of a system and for large-scale metabolite identifications, Anal. Chem. 88 (2016) 11373–11383.
[8] F. Qiu, Z. Lei, L.W. Sumner, MetExpert: an expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications, Anal. Chim. Acta (2018), https://doi.org/10.1016/j.aca.2018.03.052.
[9] C. Bueschl, B. Kluger, N.K. Neumann, M. Doppler, V. Maschietto, G.G. Thallinger, J. Meng-Reiterer, R. Krska, R. Schuhmacher, MetExtract II: a software suite for stable Isotope-assisted untargeted metabolomics, Anal. Chem. 89 (2017) 9518–9526.
[10] L. Goracci, S. Tortorella, P. Tiberi, R.M. Pellegrino, A. Di Veroli, A. Valeri, G. Cruciani, Lipostar, a comprehensive platform-neutral cheminformatics tool for lipidomics, Anal. Chem. 89 (2017) 6257–6264.
[11] H. Liu, L. Yang, N. Khainovski, M. Dong, S.C. Hall, S.J. Fisher, M.D. Biggin, J. Jin, H.E. Witkowska, Automated iterative MS/MS acquisition: a tool for improving efficiency of protein identification using a LC–MALDI MS workflow, Anal. Chem. 83 (2011) 6286–6293.
[12] M. Zhang, J. Sun, P. Chen, Development of a comprehensive flavonoid analysis computational tool for ultrahigh-performance liquid chromatography-diode array detection-high-resolution accurate mass-mass spectrometry data, Anal. Chem. 89 (2017) 7388–7397.
[13] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis, Anal. Chem. 81 (2009) 3429–3439.
[14] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC× GC-TOFMS combined to multivariate curve resolution, Anal. Chem. 83 (2011) 9289–9297.
[15] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC× GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 117 (2012) 80–91.
[16] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry using multivariate curve resolution–alternating least squares, Anal. Bioanal. Chem. 405 (2013) 6235–6249.
[17] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297.
[18] E. Szymańska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (2012) 3–16.
[19] E. Saccenti, M.E. Timmerman, Approaches to sample size determination for multivariate data: applications to PCA and PLS-DA of omics data, J. Proteome Res. 15 (2016) 2379–2393.
[20] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemometr. Intell. Lab. Syst. 78 (2005) 103–112.
[21] A. Decan, T. Mens, M. Claes, P. Grosjean, On the development and distribution of R packages: an empirical analysis of the R ecosystem, in: Proceedings of the 2015 European Conference on Software Architecture Workshops, ACM, 2015, p. 41.

[22] B. Walczak, D. Massart, Wavelets—something for analytical chemistry? Trends Anal. Chem. 16 (1997) 451–463.
[23] X. Shao, W. Cai, Z. Pan, Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis, Chemometr. Intell. Lab. Syst. 45 (1999) 249–256.
[24] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605.
[25] R.G. Brereton, Chemometrics for Pattern Recognition, Wiley, UK, 2009.
[26] A. Rizzi, A. Fioni, Virtual screening using PLS discriminant analysis and ROC curve approach: an application study on PDE4 inhibitors, J. Chem. Inf. Model. 48 (2008) 1686–1692.
[27] KEGG, Kyoto Encyclopedia of Genes and Genomes (KEGG) database. http://www.genome.jp/kegg/pathway.html, 2017.
[28] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (2000) 27–30.
[29] H. Parastar, E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Chemometrics comparison of gas chromatography with mass spectrometry and comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry daphnia magna metabolic profiles exposed to salinity, J. Sep. Sci. 41 (2018) 2368–2379.