Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data

Yahya Izadmanesh a,b, Elba Garreta-Lara b, Jahan B. Ghasemi c, Silvia Lacorte b, Victor Matamoros b, Roma Tauler b,a

a Faculty of Chemistry, K. N. Toosi University of Technology, Tehran, Iran
b Institute of Environmental Assessment and Water Research, Spanish Council for Scientific Research (CSIC), Jordi Girona 18, Barcelona 08034, Spain
c Faculty of Chemistry, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 3 October 2016
Received in revised form 25 December 2016
Accepted 22 January 2017
Available online 23 January 2017

Keywords:
GC×GC-TOFMS
Wavelet compression
Chemometrics
Multivariate curve resolution-alternating least squares (MCR-ALS)
Metabolomics

ABSTRACT

Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOFMS) is a powerful tool for separation and identification of analytes in complex natural samples. In this paper, different chemometric methods were compared for fast non-targeted GC×GC-TOFMS metabolomic profiling of the crustaceous species Daphnia magna and a general chemometric strategy and workflow is proposed. The strategy proposed in this work combined the compression of GC×GC-TOFMS data matrices in the retention time direction using wavelets and the appropriate column-wise data matrix augmentation arrangement, and its modeling by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). Using the proposed strategy, eighty different D. magna metabolites were resolved and identified. After calculation of the peak capacities of different columns and peak area changes of these metabolites, the best instrumental configuration and column combination for the GC×GC-TOFMS metabolomic study of D. magna are proposed and discussed. The procedure described in this work can be applied as a general method for untargeted GC×GC-TOFMS data processing and metabolomic profiling.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Comprehensive two-dimensional gas chromatography, GC×GC, is a well-established and useful instrumental platform for analyzing biological samples [1–4]. In comparison to one dimensional (one column) chromatographic approaches, using comprehensive two-dimensional gas chromatography-mass spectrometry (GC×GC-TOFMS) allows for higher separation efficiencies, improved resolution, better identification and quantitation of coeluting analytes [5–7]. However, some difficulties are still present even when using GC×GC-TOFMS: (1) complete separation of all detectable analytes cannot be achieved in complex biological samples, because of existing limitations in experimental and instrumental conditions; (2) baseline/background changes often cause problems in GC×GC-TOFMS analysis due to the presence of two chromatographic columns; and (3), every GC×GC-TOFMS run produces a very large, dense, information-rich three-way array of data, typically of 500–1500 megabytes (MB) per sample, depending on run time and data collection rate. When multiple samples are simultaneously analyzed and compared, current laboratory computer technologies have still difficulties to manage and process adequately the huge amounts of data produced.

The first two problems can be solved to some extent with application of multivariate chemometric methods such as bilinear multivariate curve resolution-alternating least squares (MCR-ALS) [5–9], and in some particular circumstance by parallel factor analysis (PARAFAC) [10,11], PARAFAC2 [12], and trilinear MCR-ALS [13] methods. While each of these chemometric methods used for analyzing GC×GC–TOFMS data have their own strengths and weaknesses, their analytical goals are generally the same; essentially, combination of analyte resolution, identification, and quantification. However, progressive and random elution time changes in position or shape which cause limitations in the structure of the GC×GC-TOFMS data, and restricts application of some of these chemometric methods [10].

In order to address the third challenge related with the huge size of GC×GC-TOFMS data sets, several data handling strategies such as pixel level [14–16], peak table [17], or peak region basis [18] have been proposed. These strategies propose a systematic way to GC×GC-TOFMS data analysis, but the main challenge (large-size of the data sets) is still present. Data compression methods

* Corresponding author.
E-mail address: roma.tauler@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.chroma.2017.01.052
0021-9673/© 2017 Elsevier B.V. All rights reserved.
can be used to reduce data sizes without missing relevant information, so that reasonable low capacity storage is maintained and computational demands are alleviated [19–21]. Data compression algorithms based on wavelet transform have been shown to be very effective methods in chemistry [22] and they have the advantages of simplicity, speed, and multi-resolution capability. Wavelet compressed data are amenable to secondary processing such as multivariate curve resolution modeling methods [23]. Metabolomics was defined in 1990s to describe techniques aimed at measuring metabolites present within a cell, tissue or organism during a genetic alteration or physiological stimulus [24]. Metabolomics strategies have been divided into two distinct approaches, untargeted and targeted metabolomics, each with their own inherent advantages and disadvantages. The objective of untargeted metabolomics is to comprehensively survey as many metabolites as possible and maximize number of detected peaks to construct a global profile. Due to its comprehensive nature, untargeted metabolomics must be coupled to advanced chemometric techniques, to reduce the extensive datasets into a smaller set of manageable signals. Major challenges of this approach lie in the protocols and time required to process data, difficulties in identifying and characterizing unknown small molecules, reliance on the intrinsic analytical coverage of the employed platform, and bias towards detection of high-abundance molecules. By contrast, targeted metabolomics is the measurement of predefined list of chemically characterized and biochemically annotated metabolites.

*Daphnia magna*, a freshwater crustacean, is used extensively as an aquatic test species [25], being the object of standardized testing guidelines from the Organization for Economic Co-operation and Development (OECD) [26]. Acute and chronic tests of *D. magna* are among the most frequently performed studies in aquatic toxicology because these animals are relatively easy to culture, have a short lifecycle, and can be maintained at high population densities in relatively small volumes and thus are cost-effective [27].

In this paper, a general workflow is proposed for metabolomic profiling of *D. magna* by GC×GC-TOFMS and chemometric analysis. A comparison is made on the performance of different chemometric methods and on the best GC×GC-TOFMS instrumental configurations for general metabolomics work.

2. Experimental

2.1. Chemical and reagents

D-glucose (U-13C6, 99%), used as internal standard (IS), was supplied by Cambridge Isotope Laboratories, Inc. (Andover, MA, USA). Triphenylamine (98%) used also as IS, was obtained from Sigma-Aldrich (St. Louis, MO, USA). Pyridine (anhydrous, 99.8%), chlorotrimethylsilane (TMCS), methoxyamine hydrochloride (98%) (MeOX) and N-methyl-N-trimethyl-silyl trifluoracetamide (>98.5%) (MSTFA), used as derivatization agents and the saturated Alkane standard mixture for the performance test of GC systems from C1 to C30 were also obtained from Sigma-Aldrich (St. Louis, MO, USA). Analytical reagent grade hexane, methanol and chloroform were supplied from Merck (Darmstadt, Germany).

2.2. Sample preparation and metabolite extraction

Metabolites were extracted from the whole organism of *D. magna* using the same method described in the previous work [28]. Polar metabolites of the whole organism were extracted with 400 µL methanol, vortexed and sonicated for 15 min, and then 200 µL water and 400 µL chloroform were added before centrifugation at 10,000 g during 15 min, in order to separate the aqueous and the lipid phase. The aqueous phase of every sample was transferred to a new eppendorf and 10 µL of d-glucose were added as Internal Standard (IS) at a concentration of 50 µg mL⁻¹. Afterwards, the extract was evaporated to dryness with a Speedvac (Thermo Scientific) at 40 °C during 3 h. When samples were completely dry they were stored at −80 °C until analysis.

The obtained dried extracts were derivatized as follows. 65 µL of MeOX (20 µg µL⁻¹) in pyridine were added to each sample. After mixing for 1 min, the mixture was incubated for 90 min at 30 °C. Thereafter, 30 µL of MSTFA (1% TMCS) were added, vortex mixed for a minute and then incubated for another 30 min at room temperature. Prior to injection, 10 µL of triphenylamine were added as IS in a concentration of 50 µg mL⁻¹ and, daphnid derivatized extracts were finally filtrated through a 0.22 µm filters (Ultrafree–MC, Millipore). Then, the final polar fraction was transferred to a chromatographic vial and injected into GC×GC-TOFMS instrument. GC×GC-TOFMS analysis was performed within 12 h from derivatization.

2.3. GC×GC-TOFMS

The comprehensive two-dimensional gas chromatography was coupled to time-of-flight mass spectrometry (GC×GC-TOFMS) using a Pegasus 4D (LECO, St. Joseph, MI, USA), system which has a 6890 gas chromatograph (Agilent Technologies, Palo Alto, CA, USA) equipped with a split/splitless injector, a secondary oven to fit the secondary column, and a ZX1 (Zoex, Houston, TX, USA) two-stage thermal modulator. Liquid nitrogen was used to cool nitrogen gas for cold pulses. The column combinations and method parameters used for analysis of *D. magna* samples are shown in Table 1.

The six chromatographic configurations used for analysis of metabolomic samples of *D. magna* are named S1, S2, S3, S4, S5 and S6 in Table 1. Three different column configurations were tested: A, B and C (see Table 1) and the best chromatographic configuration for *D. magna* metabolomic study was investigated based on theoretical peak capacities. The first columns (1D) used in all column configurations were coated with 5% diphenyl, 95% dimethylpolysiloxane (TRB-5MS (30 m × 0.18 mm ID × 0.18 µm film thickness) from Teknokroma (Sant Cugat del Vallès, Spain)) in configuration A, and 2B-5MS from Phenomenex (30 m × 0.25 mm ID × 0.25 µm film thickness) in configuration B and C. The used second columns (2D) were coated with 50% diphenyl, 50% dimethylpolysiloxane, (TRB-50HT (2 m × 0.10 mm ID × 0.10 µm film thickness)) in configuration A and B, and TRB-50 (20 m × 0.25 mm ID × 0.25 µm film thickness) from Teknokroma in configuration C.

The secondary oven temperature was 5 °C higher than preliminary oven and with a similar ramp. The helium carrier gas flow rate, modulation period, hot pulse duration, and temperature offset of modulator versus primary oven temperature are also given for each column combination in Table 1. The MS transfer line was held at 250 °C, and TOFMS instrument, equipped with an electron ionization ion source, was set to scan mass range of 60–700 m/z. The ion source temperature was 200 °C. Detector operation was at 1800 V and applied electron energy was 70 eV. Acquisition rate was 100 spectra per second with unit mass resolution. 2.0 µL of each sample was injected via an auto-sampler (Agilent 7890) in splitless injection mode.

3. Methods

3.1. GC×GC-TOFMS data structure

Comprehensive two-dimensional gas chromatography (GC×GC) is a technique with two columns of different selec-
tivity connected in series by a modulation device [29]. A general scheme of a GC × GC system is presented in Fig. 1 (a). First column or mode (1D column) is a conventional column, while the second column (2D column) is a shorter and efficient column as those commonly used in fast GC [30]. 1D column is connected to injector and 2D column is fitted into detector. Both columns can be placed in the same oven or each one separated. At the beginning of 2D, a modulator (M) is placed, which is the main part of the GC × GC instrument and functions in three steps: (i) Collection of continuously narrow fractions of the 1D effluent, ensuring that separation in the first column is maintained; (ii) focus or refocus the effluent and (iii) quickly transfer the 1D collected fraction sample and inject it as a narrow pulse to 2D column. The combination of these three steps is called modulation cycle, which is repeated throughout the chromatographic run. Required time to perform a cycle is called modulation period (MP), which is typically 2–10 s and depends on the time required for the compounds to be eluted in 2D. The MP should be as small as possible in order to avoid separation losses in 1D. Therefore, the entire first column chromatogram is modulated into a series of high-speed short secondary chromatograms of a length equal to MP, which are continuously recorded by the detector. K is the total number of modulations in one GC × GC-TOFMS run.

Every modulation produces a data matrix, X (I, J) where I (1, 2, . . . , l) is number of mass spectra scans in each modulation and J is the predefined m/z range to be recorded for every scan. If all modulations (k = 1, 2, . . . , K) are considered simultaneously, the whole set of acquired data of GC × GC-TOFMS analysis of a single sample, gives a three-way data array with dimensions equal to X (I, J, K). If several samples are analyzed, a four-way data array will be obtained with dimensions equal to (I, J, K, Z), where I, J, and K are same as before, and Z is the number of analyzed samples (Fig. 1 (a)). As shown in Fig. 1 (b), the three-way data array obtained in the analysis of every sample can be converted into a large two way array, Xaug, in a column-wise or in a row wise augmentation manner [31,32]. For column-wise augmentation, Xaug (I ×K, J), the different individual data matrices (or slices of three-way array) from every modulation can be put on top of each other, in a way that, the m/z direction is common among all modulations [7]. For row-wise augmentation, Xaug (I, K×J), the different individual data matrices (slices of three-way array) of modulation are set one besides the other, in a way that, the retention times in 2D column are common among all different modulations. If three-way data arrays of more than one sample are available, for simultaneous analysis, the column or row-wise augmented data matrices of different samples can be super-augmented (Xaugsup) in the same manner [7].

Fig. 2 shows the proposed chemometrics workflow for analysis of the GC × GC-TOFMS metabolomic data of D. magna. In step 1, data were first explored for their internal structure. To check whether GC × GC-TOFMS three-way data structure conforms to the trilinear model, rank analysis (by singular value decomposition, SVD, [33]) of the column and row wise augmented data matrices was performed. Chemometric methods based on different model assumptions were compared and evaluated. The tested chemo-
metrical methods were PARAFAC and MCR-ALS with the trilinear constraint (step I (a)), PARAFAC2 and non-synchronized trilinear MCR-ALS methods which allow for peak shift deviations (step I (b)), and bilinear MCR-ALS method (step I (c)) [10,34,35]. In step II, the column-wise augmented data set was compressed along the elution time mode using wavelets [21] to reduce computer storage and computation time (see below). In step III (a), compressed data were analyzed using the most suitable chemometric approach to estimate the mass spectra of the resolved metabolites. In step III (b), the elution (concentration) profiles of these metabolites in the two columns were recovered (Section 3.5). In step III (c), retention indexes (RI) were calculated. Finally, in step IV, metabolites were identified using their RI and mass spectra in $S^{'T}$, using NIST data base library. More detailed explanation of all these different steps are given below.

3.2. Data structure and modeling testing of GC×GC-TOFMS data

When GC×GC-TOFMS data conforms with the trilinear [10] model, for each component, the same elution profiles and spectra in the two columns are recovered for all modulation data matrices, only differing in a scale factor [7]. A fast way, for testing fulfillment of the trilinearity for a particular chromatographic run data set, is to perform singular value decomposition (SVD) of the corresponding row- and column-wise augmented data matrices [36] (see Fig. 1(b)). If data are trilinear, then, in absence of noise, both arrays will share the same row- and column- vector spaces and give the same number of large singular values (larger than those due to noise). In this conditions, the chemical rank (mathematical rank in absence of experimental noise [36]) should be the same and equal to the total number of eluted components [32]. In case of fulfillment of the trilinear model, and of absence of rank deficiencies in any of the data modes (i.e. every component gives a distinct profile in every mode), the same ‘chemical’ rank (mathematical rank in absence of experimental noise) should be obtained. If for instance, SVD analysis shows a higher number of components in row-wise augmented data matrix than column-wise augmented data matrix, then data structure is not in agreement with the trilinear model, probably due to changes in the elution profiles caused by time shifts and shape changes between modulations. This situation has been shown to be present in many cases as shown in several previous studies of chromatographic data and described in more detail in previous works [6–9,34].

3.3. Bilinear MCR-ALS

MCR-ALS is based on a bilinear model, which decomposes the measured data, collected in a data matrix, into a set of pure component contributions in two factor matrices. For GC×GC-TOFMS data, the bilinear MCR-ALS decomposition model can be described by Eq. (1)

$$X = CS^{'T} + E$$

Where $X$ is the data matrix collected in a single modulation, $I$ is retention times in $S^{'T}$ column and $J$ is measured $m/z$ values. $C$ is the $2D$ elution profiles of the N eluted chemical components, and $S^{'T}$ is the pure mass spectra, which can be used for identification of possible metabolites. NIST MS spectra library was used for this purpose [28]. $E$ is the residual matrix having the data variance unexplained by the model [5,8]. $C$ and $S^{'T}$ matrices can be estimated by an iterative algorithm based on two constrained alternating linear least-squares steps [37]. This bilinear model for MCR-ALS data decomposition can be easily extended to simultaneous analysis of multiple modulations of the same GC×GC-TOFMS run and to multiple GC×GC-TOFMS runs (Eqs. (2) and (3)) [5–9,38].

$$X_{aug} = C_{aug}S^{{'T}} + E_{aug}$$

$$X_{superaug} = C_{superaug}S^{{'T}} + E_{superaug}$$

To assess the quality and reliability of MCR-ALS solutions, lack of fit, LOF (Eq. (4)) and percent of explained variance, $R^2$ (Eq. (5)) values were calculated according to equations defining these two parameters:

$$LOF(\%) = 100 \frac{\sum_{ij} e_{ij}^2}{\sum_{ij} x_{ij}^2}$$

$$R^2 = 100 \frac{\sum_{ij} x_{ij}^2 - \sum_{ij} e_{ij}^2}{\sum_{ij} x_{ij}^2}$$

where $x_{ij}$ is an element of the experimental data matrix, $\hat{x}_{ij}$ is the corresponding value calculated by MCR-ALS, and $e_{ij} = x_{ij} - \hat{x}_{ij}$ is the residual value [37].
The number of components in MCR-ALS modeling was first estimated by visual inspection of the number of singular values [33] of the considered data matrix \((X, X_{\text{aug}}\text{ or } X_{\text{superaug}})\) which were larger than those already associated to noise (i.e. when the size of singular values are not changing anymore and they are small). Increasing the number of components was considered appropriate, only if a diminution in the lack of fit (Eq. (4)) and an increase in the explained variance (Eq. (5)) after MCR-ALS were observed. Also, visual inspection of the shapes of resolved elution and spectra profiles can help to ascertain the reliability of the selected components and to discern between those with chromatographic and MS signal features and those that are just noise or experimental artifacts. MS raw signal intensities were initially divided by 10^4 to make computations more manageable and to facilitate evaluations, graphical representations and comparison of the results.

### 3.4. Data compression and MCR-ALS analysis

Due to huge size in one single GC×GC-GC-TOFMS data set, especially in the time direction and for \(X_{\text{aug}}\) and \(X_{\text{superaug}}\) matrices, a data compression strategy based on the use of wavelets [20,39,40] allowed an easier chemometric analysis and reduced computer storage requirements considerably. Wavelet compression consists of two algorithms: first, there is a compression algorithm that takes the input data matrix, \(X\), and generates a representation of it, \(\hat{X}\), which requires fewer computer storage (bits), and there is also a reconstruction algorithm that operates on this representation, \(\hat{X}\), and generates reconstructed matrix [21]. Reconstructed matrix contains the same information as \(X\), but requires much less computer storage.

In this work, the discrete wavelet transform compression/filtering algorithm was used to reduce the size of the huge time mode of GC×GC data without missing relevant information [20]. This wavelet function can be described by Eq. (6) [19].

\[
(f(t) , \Psi_{m,n}(t)) = \int_{-\infty}^{+\infty} f(t) \Psi_{m,n}(t) dt
\]

where \(f(t)\) is the signal to be compressed, \(\Psi_{m,n}(t)\) is the mother wavelet function, \(m\) is the scale and \(n\) is the shift in time. The used mother wavelet in this study was the simple Haar function (Eq. (7)).

\[
\Psi_{m,n}(t) = \begin{cases} 
1 & \text{if } 0 < t < 0.5 \\
-1 & \text{if } 0.5 < t < 1 \\
0 & \text{elsewhere}
\end{cases}
\]

By changing size of the mother wavelet function, or scaling and translating it, other mother functions can be easily obtained, which provides the multi-resolution property [39].

Multi-resolution property of wavelets which means adapting the signal resolution makes it applicable to stationary and non-stationary signals [41].

Wavelet decomposition and compression was applied independently on every column \((m/z)\) of the \(X_{\text{aug}}\) matrix. Compression reduces the size of data \(2^n\) times, which, \(n\) is the compression level [20,39,40]. The compressed matrix contains the same information as \(X_{\text{aug}}\), but needs much lower computer storage. For the datasets under study in this paper, it is demonstrated that level-4 wavelet compression is reliable without significant loss of relevant information in the elution time direction. Data compression was performed individually on augmented matrix of each samples \((X_{\text{aug}})\). The compressed data of all samples were then super-augmented to obtain \(X_{\text{compress}}\) for MCR-ALS analysis.

Fig. 3 shows the general procedure followed for simultaneous MCR analysis of wavelet compressed GC×GC-GC-TOFMS data of multiple chromatographic runs and subsequent recovery of elution
profiles in 1D and 2D columns using fast non-negative least squares [42] post-processing step. MCR-ALS on $X_{\text{compr}}$ gave the compressed elution profiles ($C_{\text{compr}}$) and mass spectral profiles ($S^1$). Once $S^1$ was properly estimated, it was possible to obtain the uncompressed elution profiles ($C_{\text{superaug}}$) using a non-negativity constrained least squares postprocessing step [42]. $C_{\text{aug}}$ was then extracted from $C_{\text{superaug}}$ taking into account the order used for superaugmentation of the different samples in $D_{\text{superaug}}$ matrix (see Section 3.3).

3.5. Elution profiles in first and second columns

$C_{\text{aug}}$ is the column-wise augmented concentration matrix of one sample and contains the concatenated 2D elution profiles. The elution profiles of metabolites in 1D and 2D columns can be recovered by refolding the augmented elution profile of every component in $C_{\text{aug}}$. It contains second-column elution profiles for all N components in all modulations. To obtain first and second-column elution profiles of every component in each sample, every column in $C_{\text{aug}}$ should be appropriately refolded to give a matrix with ($I$, $K$) dimensions, where $I$ is the number of elution times in the second column and $K$ is the number of modulations, i.e. the number of elution times in the first column. The sum along the columns of this refolded data matrix gives an estimation of the corresponding first-column elution profiles. Therefore, for every sample, a matrix of first-column elution profiles of dimensions ($N$, $K$) is obtained, $N$ being the number of resolved components.

In case of several samples ($X_{\text{superaug}}$), the concentration matrix of every sample ($C_{\text{aug}}$) is extracted from $C_{\text{superaug}}$, by keeping in mind the arrangement of samples in ‘superaugmentation’ step (Fig. 3). Then the procedure for obtaining 1D and 2D elution profiles is the same as explained before.

3.6. Metabolites identification

MCR-ALS resolved elution and spectra profiles were assigned to metabolites and identified by comparing the retention indexes (RI) of the peak 1D profiles on one side and from the mass fragmentation patterns associated to the MCR-ALS resolved mass spectra profiles on the other side. Standard mass spectral database of the National Institute of Standards and Technology (www.nist.gov/srd/nist1a.htm) was used to identify spectra fragmentation patterns. For each mass spectrum, 100 hits were retrieved by the NIST Mass Spectral Search 2.2 software distributed with the NIST 2014 library. A reverse match factor (RMF) based on the correlation coefficient between the MCR-AS resolved and experimental mass spectra reported by NIST software was used for selection of the best identified compound for MCR-ALS resolved mass spectra. This match factor is reported between 0 (no match) and 1000 (perfect match). As a general guide, a value of 900 or greater was considered to be a very good matching; between 800 and 900, a good match; between 700 and 800, a fair match; and less than 600 a poor or very poor match. To increase the reliability of the identification, NIST internal linear RI markers of first column were also included in the evaluation of the library hits, using the injection of a saturated Alkanes standard mixture ($C_7$ to $C_{30}$) [43].

3.7. Calculation of peak capacities

The most suitable chromatographic configuration (column and instrumental setup) was chosen by comparison of the peak capacities of all chromatographic configurations given in Table 1. Peak capacity in the first column was calculated by Eq. (8) [44]:

$$CP_1 = \frac{(t_{R,1}) - t_{M,1}}{4\delta}$$

(8)

where $CP_1$, $t_{M,1}$ and $t_{R,1}$ are respectively peak capacity, column holdup and retention times of the last peak eluting in the first column, and $4\delta$ is the average peak width. In the second column, the separation is approximately isothermal because of the short length of 2D column, and peak capacity was calculated using Eq. (9) [45]:

$$CP_2 = 1 + \frac{\sqrt{N}}{4} \times \ln \left(\frac{t_{R,2} \times t_{M,2}}{t_{R,1} \times t_{M,1}}\right) \times 20$$

(9)

where $CP_2$, $t_{M,2}$ and $t_{R,2}$ are the peak capacity, column holdup and maximum retention times for the last peak eluting from 2D column, respectively [44,45]. The peak width and retention times of glucose were used to estimate the number of plates ($N$) in the second column. Overall peak capacity was obtained by multiplying $CP_1$ and $CP_2$.

3.8. Software

GC×GC-TOFMS data were acquired using ChromaTOF software version 3.3.2 (LECO, St. Joseph, MI, USA). Data from ChromaTOF software were imported and converted into MATLAB in the 60–700 m/z range.
range by using bioinformatics toolbox (The Mathworks, Inc., Natick, MA, USA). The compression was done by the wavelet toolbox (The Mathworks, Inc., Natick, MA, USA). NIST MS Search version 2.2 (National Institute of Standards Technology, USA) and the Golm metabolome database (GMD) of derivatized compounds [46,47] were used for metabolite identifications. Wavelet toolbox of Matlab was used [48] for data compression. PARAFAC and PARAFAC2 from PLS toolbox (http://www.eigenvector.com/) and MCR-ALS toolbox [37] were used for chemometrical analysis.

4. Results and discussion

4.1. GC×GC-TOFMS experimental profiles and selection of the best chromatographic configuration

Fig. 4 shows the 3D TIC chromatograms of the D. magna sample extract analyzed by GC×GC-TOFMS using the different column configurations set in Table 1. The peak shape and peak capacity of metabolites, as well as their distribution along the chromatogram are influenced by the internal diameter, operating conditions and film thickness of the columns [1,44]. According to Fig. 4, the column set A (see Table 1) showed an apparent worse result on the resolution of the peaks and peak shapes of the metabolites than column set B. The change in offset modulations from 30 °C to 15 °C did not bring any significant improvement in the separation conditions of samples analyzed by S1, S2 and S3 configurations (see Fig. 4(a)–(c), respectively). Taking into account these considerations, in the column set B, 1D column was changed by another with an ID of 0.25 mm and 0.25 μm of film thickness.

In column configuration S4, by the column set B, with a 1D column with ID 0.25 mm, and a 2D column with ID 0.10 mm (see Table 1), the parameters were modified with the intention of obtaining the highest peak capacity (see Fig. 4(d)). A distribution of the peaks was better than those observed for the previous three configurations. In order to test an additional configuration, column configuration C was also tested, in which, the two columns had the same internal diameter (0.25 mm ID), (see Table 1). This new configuration was intended to determine the effects of increasing diameter in the second column and, to avoid peak overload and obtain optimum flow rates in both columns (see Fig. 4(e) and (f)).

Results showed that column set C was probably the configuration providing better chromatographic results. However, to ensure that this configuration was better than others, chromatographic parameters (peak capacity and resolution) of all identified metabolites were also assessed and compared for the different column configurations. In this regard, selection of the best column configuration was done by comparison of the peak capacities of configurations (Section 3.6).

Fig. 5(a) shows the calculated first and second column peak capacities and total peak capacities of the different investigated chromatographic setups. The peak capacities of S1, S2, S3 and S4 setups were lower than peak capacities of S5 and S6 setups. In S1, S2, S3 and S4 setups, separation of metabolites achieved in 1D column was partially lost in 2D column because of longer modulation periods. In second column, the column length is short and with longer modulation periods, the probability of remixing of components and therefore overlapping of elution profiles increases. In contrast, in configurations S5 and S6, the separation efficiencies achieved in the first column were maintained in 2D column, because of their shorter modulation periods (2.5 s). The peak capacity of S4 setup was higher than peak capacity of S1, S2 and S3 setups, probably because the stationary film thickness of 1D and 2D columns in S4 was higher and the metabolites separation was more effective. When comparing S5 and S6 setups, the final temperature in the former was 310 °C, which was too high for temperature tolerance of 2D column, hence, S6 setup was finally preferred over S5 setup as the best column combination. It has the highest peak capacity and therefore it allowed for a better distribution of metabolites over the whole available chromatographic space.

Moreover, the best instrumental chromatographic configuration was also investigated by comparison of the peaks areas of metabolites obtained in the analysis of different D. magna samples using synchronized chromatographic configurations (Table 1) and instrumental setups. Fig. 5(b) compares the peak area of one of the metabolites (lactic acid) when analyzed by mentioned chromatographic setups. According to Fig. 5(b), the highest peak area of lactic acid has been obtained when the sample analyzed by configuration S6. Therefore, configuration S6 was additionally confirmed to be the best GC×GC configuration for metabolomic study of D. magna, and configuration S6 is proposed to be used in future GC×GC-TOFMS metabolomics studies.

4.2. Testing GC×GC-TOFMS data structure and selection of the best chemometric approach

To explore the structure of the GC×GC-MS datasets, modulations 94–100 of sample S1 were first selected and analyzed in detail. Comparison of SVD analysis of row- and column-wise augmented data matrices showed some differences on how the size of the singular values decrease with the number of components. Sizes of singular values of column-wise augmented matrix (sharing the same MS spectral vector space) decrease much faster than those of row-wise augmented matrix, especially for the first ones. This is a preliminary indication of possible deviations from the trilinear model.

Fig. 6 shows the 2D elution profiles of one of the resolved components in modulations 94–100 by different modeling methods. 2D elution profile resolved by PARAFAC2 (in blue) was physically meaningless (with negative values and more than one peak). In case of bilinear MCR-ALS, 2D elution profiles had small changes in the shape and position in successive modulations. In non-synchronized trilinear MCR-ALS, shapes of elution profiles are forced to be the same in successive modulations, but, due to shape changes of elution profiles from modulation-to-modulation in raw data, the fitting was worse in comparison with bilinear MCR-ALS [49,50]. In synchronized trilinear MCR-ALS, shapes of elution profiles and their retention times are forced to be the same in successive modulations, but, due to within run retention time shifts from modulation to modulation in raw data, the fitting was worse in comparison with non-synchronized trilinear MCR-ALS. The elution profile of PARAFAC is rather acceptable, but because of retention time shifts and trilinearity deviation of the data, the LOF and R² values are worse in comparison with bilinear MCR-ALS. From results in Table 2, rank analysis of row and column wise augmented data matrices and results shown in Fig. 6, bilinear MCR-ALS was concluded to be the best chemometric approach for the investigated data. The remaining models include artificial constraints which are not applicable at all, at least in principle, to chromatographic data. An additional advantage of bilinear MCR-ALS modeling is that data

<table>
<thead>
<tr>
<th>Method</th>
<th>LOF (%)</th>
<th>R² (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilinear MCR-ALS</td>
<td>6.22</td>
<td>99.61</td>
</tr>
<tr>
<td>Non-synchronized trilinear MCR</td>
<td>11.76</td>
<td>98.61</td>
</tr>
<tr>
<td>Synchronized trilinear MCR</td>
<td>24.98</td>
<td>93.76</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>11.50</td>
<td>98.68</td>
</tr>
<tr>
<td>PARAFAC2</td>
<td>78.36</td>
<td>38.58</td>
</tr>
</tbody>
</table>
matrices from different samples or modulations could have different number of rows (retention times).

MCR-ALS results in the analysis of one Daphnia sample from a single GC×GC–MS chromatographic run (Xaug).

Fig. 7 shows MCR-ALS results in the analysis of sample analyzed by configuration S1, using the compressed and uncompressed column-wise augmented data matrices. Fig. 7 (a) shows the contour plot of the total ion chromatogram (TIC) of this sample. Chromatographic signals appearing around scan number at 150 in second column and at all tR were mostly only due to the derivatization reagents. In GC×GC analysis, derivatization of metabolites produced more than one reactant and hence elution profile for the same metabolite, which made the analysis of the full scan GC×GC-TOFMS data more challenging. The inset of Fig. 7 (a) shows the small part of the chromatogram which was analyzed (modulations 94–100) in more detail. This chromatographic region has a peak

---

**Fig. 4.** Three-dimensional total ion chromatograms (TIC) of the D. magna sample extracts analyzed by GC×GC-TOFMS using different chromatographic configurations. TIC plot of: (a) S1, (b) S2, (c) S3, (d) S4, (e) S5, and (f) S6.

<table>
<thead>
<tr>
<th>Setup</th>
<th>CP1</th>
<th>CP2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>275</td>
<td>20</td>
<td>5500</td>
</tr>
<tr>
<td>S2</td>
<td>280</td>
<td>20</td>
<td>5600</td>
</tr>
<tr>
<td>S3</td>
<td>275</td>
<td>20</td>
<td>5500</td>
</tr>
<tr>
<td>S4</td>
<td>280</td>
<td>25</td>
<td>7000</td>
</tr>
<tr>
<td>S5</td>
<td>310</td>
<td>30</td>
<td>9300</td>
</tr>
<tr>
<td>S6</td>
<td>310</td>
<td>30</td>
<td>9300</td>
</tr>
</tbody>
</table>

**Fig. 5.** (a) Calculated peak capacities of the different chromatographic configurations used for D. Magna samples analysis, PC1: peak capacity in first column, PC2: peak capacity in second column, Total: total peak capacity. (b) Peak area bar plots of metabolite 1 (lactic acid) in the six D. Magna samples analyzed by the different chromatographic configurations. The peak areas are obtained from Csuperaug (Section 3.3).
cluster of potential metabolites that are highly overlapped, and their identification would be rather difficult without using chemometric methods. Raw chromatographic and MS spectral data from this region were arranged into a column-wise augmented data matrix with 2800 rows and 250 m/z values in its columns. This data matrix needed 5.6 megabytes of memory storage. Fig. 7(b) shows the elution profiles of the metabolites resolved directly by MCR-ALS analysis of this single data matrix. The compressed data matrix using level-4 wavelets (see Section 3.2) had a reduced number of 175 rows, and the same number of columns (m/z). It only required computer storage of 0.35 megabytes (reduction in storage of 16 times). In Fig. 7(b) and (c), elution profiles resolved by MCR-ALS of these two augmented data matrices (raw- and compressed) are compared. Since data compression was performed only in the retention time direction, shapes of elution profiles in Fig. 7(b) and (c) were different, but resolved mass spectra profiles were practically the same. Fig. 7(d) shows the MCR-ALS resolved mass spectra obtained using the compressed augmented data matrix, in which the relevant chemical metabolites are already identified in the legend. These chemical compounds were tri methyl silyl (volatilization agent) derivatives including, 2, 2, 2-trifluryl-N-methyl acetamide, trimethyl silanol and dimethyl silyl fluoride which were fragments of the derivatization agent. Other identified compound was pyridine which was used as derivatization media and it was present in all modulations. The two remaining components were the 1-Nannonol, and the MeOx and TMS derivatives of ribulose. Fig. 7(e) shows the final first column (1D) elution profiles and Fig. 7(f) shows the final second column (2D) elution profiles of the identified components (Section 3.3). Fig. 7(e) confirms that pyridine and other fragments of the derivatization agent were constantly present in all modulations, while the two metabolites of interest were modulated for a limited number of times (typically within 2–4 modulations in the case understudy) from 1D to 2D columns. These results demonstrate the reliability of the proposed MCR-ALS analysis and its results.

LOF and R² (Eqs. (4) and (5)) values obtained in the MCR-analysis of the uncompressed data were equal to 6.05% and 99.63% respectively. These results were reasonable considering the experimental error level of the data. LOF and R² values of MCR-ALS analysis of the same data once were compressed were 2.11% and 99.95% respectively. This improvement is due to the noise filtering of the wavelet decomposition compression procedure, which removes low and high frequency noisy parts from data.

4.3. Elution profiles recovery after simultaneous analysis of two samples

Fig. 8(a) shows 1D elution profiles and Fig. 8(b) shows 2D elution profiles of modulations 94–100 when samples analyzed by S1 and S2 condition sets were simultaneously modeled by MCR-ALS using the column-wise super-augmented data matrix, Xsuperaug (Fig. 3). In order to check for the differences between elution profiles obtained from compressed and uncompressed data, only scans 201–400 of modulations 94–100 are given in Fig. 8. Original raw chromatographic data were 16 times compressed (level-4 compression), with a change of computer storage from 2.8 megabytes to 0.175 megabytes. Compression was done for each sample columns-wise augmented data matrix separately (as in previous Section 4.5) before building the two samples column-wise super-augmented data matrix. Four components were selected to model the systematic data variance observed for this particular GC×GC-TOFMS data section. LOF (Eq. (4)) and R² (Eq. (5)) values were equal to 6.04%
and 99.53 for the raw uncompressed data and 6.19% and 99.61% for compressed data. In this case, LOF values for wavelets compressed data were practically equal to uncompressed data. Compared to the previous case, where only one augmented data matrix was analyzed, the LOF values of uncompressed data are higher, because, between run differences for the two samples were also affecting the fit. Since the two chromatographic runs were wavelet compressed separately along their retention times, differences between sam-
samples raised and LOF values increased a little bit although wavelet compression reduced data noise for every run separately [51–53]. The two effects were compensating each other, and final fit resulted to be similar in both cases.

As shown in Fig. 8, MCR-ALS elution profiles obtained from raw uncompressed and wavelets compressed data, resulted to be very similar, which proved again the reliability of bilinear MCR-ALS results in the simultaneous analysis of the two samples analyzed by chromatographic setups S1 and S2. This shows that wavelet compression doesn’t result in the loss of useful information and it only stores the same information in much lower digital space, because the elution profiles recovered form uncompressed and compressed data are nearly the same. Fig. 8 does show that peak heights/areas/shapes are well preserved after compression with little changes. Accepting these small changes might be a compromise that is necessary to reap the benefits of applying MCR-ALS to such a large dataset. This result indicated that, it could be possible to extend the proposed procedure of wavelet compression to the simultaneous analysis of multiple (more than two) chromatographic GC×GC-TOFMS runs (multiple samples) by MCR-ALS [54,55]. This will be especially helpful when multiple samples are compared and analyzed simultaneously (like in metabolomic studies), which due to the huge data storage requirements in these cases, the size reduction by wavelet procedure proposed here is necessary for fast metabolomics profiling.

![Graph](image)

**Fig. 8.** elution profiles of (a) first column and (b) second column obtained by MCR-ALS simultaneous analysis of the column-wise super-augmented GC×GC-TOFMS (Eq. (3)). X_{superaug}, compressed and uncompressed data matrices, for modulations 94–100 of two samples, analyzed using setups S1 and S2. Different colors are used for elution profiles of the different components resolved by MCR-ALS, and different symbols are used for elution profiles resolved from compressed and uncompressed chromatograms (see text in the figure).

### Table 3

Computer storage and modeling time requirements of samples datasets analyzed by different chromatographic configurations (in megabytes, Mb) before and after compression.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Uncompressed (Mb)</th>
<th>Time (min)</th>
<th>Compressed (Mb)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>419</td>
<td>480</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>S2</td>
<td>419</td>
<td>480</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>S3</td>
<td>708</td>
<td>760</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>S4</td>
<td>459</td>
<td>550</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>S5</td>
<td>641</td>
<td>680</td>
<td>32</td>
<td>42</td>
</tr>
<tr>
<td>S6</td>
<td>489</td>
<td>690</td>
<td>30</td>
<td>43</td>
</tr>
<tr>
<td>all samples</td>
<td>3135</td>
<td>3700</td>
<td>187</td>
<td>230</td>
</tr>
</tbody>
</table>

* Time (minutes) required for data modeling by MCR-ALS.

### 4.4. MCR-ALS simultaneous analysis of multiple GCGC-TOFMS chromatographic runs (X_{superaug})

Datasets of *D. magna* samples analyzed by different chromatographic conditions listed in Table 1, were individually submitted to level-4 wavelet compression along the retention time direction, which computer storage reduced 16 times. Then, compressed data were column-wise augmented in a single super-augmented data matrix. Computer storages of every chromatographic run and time required for MCR-ALS modeling of every dataset before and after level-4 wavelet compression are given in Table 3. The computer storage for one single chromatographic run (sample) was around 550 megabytes and for the sum of the six chromatographic runs of uncompressed raw data matrices this resulted to be more than three gigabytes. This was too big for its direct simultaneous analysis by MCR-ALS. Without any compression this would need several hours to complete MCR-ALS modeling. Once data were compressed, much less time was needed (around one hour). After data compression, the computer storage required to store the same six chromatographic runs decreased up to 187 megabytes, which could be easily analyzed by MCR-ALS using current laboratory PC computers with only some few Gb of memory.

150 components were resolved by MCR-ALS analysis of the super-augmented compressed data matrix of six samples. Since the goal was to identify as many metabolites as possible, this large number was initially chosen to model most of the data variance presented in the data. Non-negativity constraints were applied to concentration and spectral profiles, and spectra were also normalized to have equal height. MCR-ALS LOF value was only 2.60% which indicated a very good data fitting, and R² was equal to 99.93%. By proposing larger number of components, no additional interpretable information was obtained.

Individual mass spectra of MCR-ALS resolved components were converted one-by-one into .txt files readable by NIST library. Elution profiles of the corresponding components in the two columns (1D and 2D) were also analyzed (Section 3.3). Retention index (RI) values of components were calculated at the maximum height of 1D elution profiles.

Metabolites were identified by the procedure described in section 0. Identified metabolites were distinguished from derivatizants related components by comparison of the corresponding recovered chromatographic peaks. The derivatization neutralized the polar ends of the metabolites and byproducts of the derivatization were present during the whole chromatographic run giving irregular peak shapes. All components of this type were finally discarded and not used for identification. On the contrary, identified metabolites had regular chromatographic peak profiles and they were only modulated for a limited number of times from 1D column to 2D column. Only those MCR-ALS resolved components showing reasonable chromatographic peak shapes were finally considered for identification and were assumed to be linked to *D. magna* metabolome. A total number of 80 metabolites were finally
identified. Most of them were amino acids, organic acids and carbohydrates, from unstressed D. magna metabolism [28] and, mostly contained the tri methyl silyl masking agents from the derivatization agent. Table M1 in Supplementary material shows metabolites of D. magna that were identified, with their retention times, RI and RMF.

5. Conclusions

In this work, a fast and efficient workflow for untargeted metabolomic profiling of GC×GC-TOFMS data obtained in the analysis of crustacean D. magna samples is proposed. For this purpose, wavelet compression of chromatographic data sets in time direction is proposed to reduce data storage and allow their further high throughput analysis with currently used laboratory computers. Application of chemometric methods helped to overcome the incomplete highly coeluted separations of the metabolites in D. magna samples, background noise correction and multicomponent analysis challenges of GC×GC-TOFMS data sets. GC×GC-TOFMS data did not conform to the trilinear model requirements in general and use of the bilinear MCR-ALS chemometric method was preferred.

The chemometrics workflow proposed in this work was used for metabolomic profiling of crustacean species, D. magna and eighty different metabolites were identified. To find the best GC×GC-TOFMS conditions, metabolic extracts of D. magna were analyzed with six different chromatographic configurations. The best chromatographic configuration was selected by comparing peak capacities and peak areas of the resolved metabolites in each setup. The best column set included a relatively thick non-polar stationary phase in the first column: ZB-5MS (30 m × 0.25 mm ID × 0.25 μm film thickness), and a thick stationary film of intermediate polarity with the same internal diameter in the second column: TRB-50 (2.0 m × 0.25 mm ID × 0.25 μm film thickness). The combinations of these two columns in first and second chromatographic dimension provided the highest peak capacity among the other column sets studied in this work, and it is proposed for future metabolomics work.

Acknowledgements

The authors would like to thank Dr. B. Campos and Dr. C. Barata who supplied the biological samples for this study, and Dr. R. Chaler and D. Fanjul for GC×GC-TOFMS support. Yahya Izadmanesh acknowledges partial funding of his research stay in Institute of Environmental Assessment and Water Research (IDAEA) of CSIC by the Ministry of Science, Research and Technology (MSRT) of Iran. Romà Tauler and Elba Garreta-Lara acknowledge research funding from MINECO Spain grant Nr. CTQ2015-66254-C2-1-P and to CHEMAGEB project, (EU FP/2007-2013/ERC Grant Agreement n. 230737).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jchroma.2017.01.052.

References

31–58.

Chemom. Anal. Acta three-way representation, salicylic thresholds chromatography:
842 alignment Chemom. S. acid–base


