Avoiding hard chromatographic segmentation: a moving window approach for the automated resolution of GC–MS-based metabolomics signals by multivariate methods.

Xavier Domingo-Almenara[a,b,*], Alexandre Perera[c,d], Jesus Brezmes[a,b]

[a]*Metabolomics Platform - IISPV, Department of Electrical and Automation Engineering (DEEEA). Universitat Rovira i Virgili, Tarragona, Catalonia, Spain.*
[b]*Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain.*
[c]*B2SLAB, Department of ESAII – Center for Biomedical Engineering Research (CREB), Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain.*
[d]*Biomedical Research Centre in Bioengineering, Biomaterials and Nanomedicine (CIBERBBN), Madrid, Spain.*

Abstract

Gas chromatography – mass spectrometry (GC–MS) produces large and complex datasets characterized by co-eluted compounds and at trace levels, and with a distinct compound ion-redundancy as a result of the high fragmentation by the electron impact ionization. Compounds in GC–MS can be resolved by taking advantage of the multivariate nature of GC–MS data by applying multivariate resolution methods. However, multivariate methods have to be applied in small regions of the chromatogram, and therefore chromatograms are segmented prior to the application of the algorithms. The automation of this segmentation process is a challenging task as it implies separating between informative data and noise from the chromatogram. This study demonstrates the capabilities of independent component analysis – orthogonal signal deconvolution (ICA–OSD) and multivariate curve resolution – alternating least squares (MCR–ALS) with an overlapping moving window implementation to avoid the typical hard chromatographic segmentation. Also, after being resolved, compounds are aligned across samples by an automated alignment algorithm. We evaluated the proposed methods through a quantitative analysis of GC-qTOF MS data from 25 serum samples. The quantitative performance of both moving window ICA–OSD and MCR–ALS-based implementations was compared with the quantification of 33 compounds by the XCMS package. Results shown that most of the R^2 coefficients of determination exhibited a high correlation (R^2>0.90) in both ICA–OSD and MCR–ALS moving window-based approaches.

Keywords: gas chromatography, orthogonal signal deconvolution, multivariate curve resolution, moving window, independent component analysis, metabolomics.

1. Introduction

Gas chromatography – mass spectrometry (GC–MS) has been extensively applied for compound profiling in metabolomics experiments due to the highly reproducible electron impact ionization process [1]. Electron impact

*Corresponding author: xavier.domingo@urv.cat. Phone +34 977559619. Fax: +34 977559605

(EI) is a high fragmentation ionization method which leads to an extensive fragmentation. Therefore, the richness of GC–MS data relies on an inherent correlation – or ion-redundancy – between fragments or ions from the same compound, i.e., different peak fragments appear at the same retention time and with the same elution profile [2]. However, compounds in GC–MS might appear co-eluted - chromatographycally not completely separated or resolved - and/or at trace levels. Due to the multivariate nature of GC–MS data, some approaches for its processing have been focused on the implementation of multivariate methods.

The most reported multivariate methods applied for the resolution of GC–MS signals are those based on multivariate curve resolution – alternating least squares (MCR–ALS) [3, 4], or parallel factor analysis (PARAFAC) [5], including one of its most frequently used variants, PARAFAC2 [6]. Algorithms based on independent component analysis (ICA) have also been applied for GC–MS signal resolution [7, 8, 9]. More recently, we introduced an alternative application of ICA, called independent component analysis – orthogonal signal deconvolution (ICA–OSD) [2, 10], for the resolution of GC–MS chromatograms, where the concept of independence was twisted: whereas the aforementioned ICA-based methods consider the spectra as the independent source in the chromatograms, ICA–OSD considers the elution profile as the independent source, as opposite to the spectra [10]. In that sense, in ICA–OSD, ICA is employed to extract the elution profiles and then determine the spectra by means of OSD. Orthogonal signal deconvolution (OSD) is a method that uses principal component analysis (PCA) as an alternative to the typical use of least squares (LS) used for example in MCR–ALS. When applying LS, no correlation or covariance information is taken into account, and this might introduce a bias into the LS regressors specially in situations of co-elution or under undue biological matrix interference [2, 10]. OSD allows the extraction of more pure spectra in comparison with least squares-based algorithms.

Despite the availability of multivariate methods for GC–MS signal resolution, the correct answering to biological hypothesis or the discovery of new biological insights is one of the main challenges in untargeted GC–MS-based metabolomics. In that sense, all the implementations of multivariate methods should be fully automated, and this automatization should not be limited to the deconvolution process but it should include the posterior alignment of the resolved metabolites. There is a need for high-throughput application of these multivariate methods. Several automated methods based on the aforementioned algorithms have been reported [11, 12, 13, 14, 15]. However, as curve resolution techniques work in small and regional intervals [14], the application of multivariate methods in high-throughput GC–MS resolution is usually conducted by a hard chromatographic segmentation, i.e., windowing or dividing the chromatogram by selecting those regions with putative information – compounds – to be resolved. The automation of this segmentation process is a challenging task as it implies separating what is informative data and what is noise from the chromatogram and thus, selecting regions of the chromatogram without splitting compounds on window borders or loosing useful information, i.e., considering compounds at trace levels as noise.

Moving windows have been used in GC–MS for factor analysis [16, 17, 18, 19]. In these studies, factor analysis techniques were applied trough a moving window with the aim of detecting components or spectral features. Those spectral features can be later resolved for a posterior resolution and comparison between samples. More recently,

the concept of sliding window multivariate curve resolution (SW–MCR) [20] was introduced for the resolution of ion-mobility gas chromatography data. When using a moving - or sliding - window to resolve the chromatogram, a same compound might be split (resolved) in consecutive windows, leading to duplicated and partial information. In SW-MCR, they tackle this issue by grouping compounds through consecutive windows based on the similarity of their spectra. Grouping compounds across windows based on spectral similarity is a challenging task, as due to noise, the spectra of the same compounds deconvolved from two consecutive windows might change. To our knowledge, the performance and suitability of moving window MCR–ALS and ICA–OSD-based approaches for the automated resolution of GC–MS metabolomics samples has not yet been studied.

In this study we propose an automated application of multivariate methods for the resolution of GC–MS signals in biological samples through an overlapping moving window approach. This approach avoids hard segmentation or windowing of the chromatogram. We propose a duplicity filter based on the minimization of the residual error to filter duplicated compounds resolved across windows, and thus selecting the best models. Also, to increase the automated reproducibility of the results, we used an existing automated method for aligning compounds across samples. To demonstrate the capabilities of the proposed overlapping moving window approach, we chose ICA–OSD and MCR–ALS as resolution methods. We evaluated the proposed methods through a quantitative analysis of GC–qTOF/MS data from serum samples and the quantitative results were compared with XCMS [21, 22], an automated workflow for GC–MS data processing.

2. Materials and methods

2.1. Materials

The methods were compared by the quantification of 33 metabolites across 25 serum samples, analyzed through GC-qTOF MS. This same dataset was previously used to demonstrate the capabilities of the *eRah* R package [23], and raw GC–MS files are available at MetaboLights with accession number MTBLS321. More details on the dataset, sample preparation and methods can be found in the original study. Briefly, analysis was carried out on a qTOF MS 7200 (Agilent, Santa Clara, CA, USA) coupled to an Agilent 7890A gas chromatograph (GC). Derivatized samples (1 μL each) were injected in the gas chromatograph system with a split inlet equipped with a J&W Scientific DB5–MS+DG stationary phase column (30 mm $\times$ 0.25 mm i.d., 0.1 μm film, Agilent Technologies). Helium was used as a carrier gas at a flow rate of 1 mL/min in constant flow mode. The injector split ratio was adjusted to 1:5 and oven temperature was programmed at 70 °C for 1 min and increased at 10 °C/min to 325 °C. The MS was operated in the electron impact ionization mode at 70 eV. Mass spectral data were acquired in full scan mode from m/z 35 to 700 with an acquisition rate of 5 spectra per second.

2.2. Data processing workflow

The data analysis pipeline is shown in Fig. 1. First, chromatographic signals were filtered by noise and baseline removal as described in [2]. Second, both moving window-based ICA–OSD and MCR–ALS implementations were

used to automatically extract and deconvolve the compounds concentration profiles and spectra. The methods were compared using different lengths of window, concretely, we used 50, 75 and 100 scans length corresponding to 10, 15 and 20 s respectively. We used an overlap of 50 % for all the implementations. The number of factors or components for both ICA and MCR was determined by a singular value decomposition (SVD), as described in [24]. MCR–ALS was initialized by means of a principal component analysis (PCA). MCR–ALS was constrained with a non-negative least squares regression, and both ICA–OSD and MCR–ALS resolved elution profiles were constrained with unimodality. Both MCR–ALS and ICA–OSD algorithms employed were those included in the R package *osd* [2]. Due to its improved run-time performance, the ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [25] implemented in the R package *JADE* [26]. Fragments at m/z 73, 74, 75, 147, 148, and 149 were excluded before any processing, since they are widespread mass fragments typically generated from compounds carrying a trimethylsilyl-moiety [27].

Third, and once compounds were resolved, they were posteriorly aligned across samples with the *eRah* [23] alignment algorithm. Finally, after alignment, an average spectrum for each compound – determined by the mean of the compounds spectra across samples – was compared using the cosine dot product to reference spectra. Reference spectra were provided by the Golm Metabolme Database (GMD) [28]. This comparison allowed a putative identification of the resolved and aligned compounds.

Additionaly, GC–MS chromatograms were processed using XCMS in order to detect and align features. A feature is defined as an ion entity with a unique m/z and a specific retention time (mzRT). XCMS analysis provided a list containing the retention time, m/z value, and peak intensity (or area) of each feature for each serum sample. XCMS parameters used to process the data are described elsewhere [23].

2.3. Moving window resolution of chromatographic signals

The aim of the method is to achieve the resolution of an entire chromatogram. Then, a moving window is proposed where, in each iteration, the window is displaced with a determined overlap along the retention time (Fig. 2 (a)). Each chromatographic window is resolved into pure chromatographic profiles and spectra (Fig. 2 (b)).

We employed two methods for the resolution of mixtures, one is the widely used multivariate curve resolution – alternating least squares (MCR–ALS), and the other is independent component analysis - orthogonal signal deconvolution (ICA–OSD). Both algorithms share the same objective based on the assumption of the Lambert-Beer's law, which can be mathematically described as follows:

$$D = CS^T + E \tag{1}$$

where D ($I \times J$) is the chromatographic window to be resolved, C ($I \times n$) is the resolved concentration profile matrix, S ($J \times n$) is the resolved spectra matrix and E ($I \times J$) is the error matrix. In this notation, I is the number of mass spectra scans (retention time), J is the acquisition range of the mass-charge ratio (m/z), and n is the number of components or compounds in the model. MCR–ALS uses an iterative least squares algorithm (ALS) to determine both

C and S matrices by minimizing the error matrix E. A detailed explanation of MCR–ALS, together with pseudocode, is given elsewhere [29]. In ICA–OSD, independent component analysis is used to extract the chromatographic profile matrix C by considering the elution profiles as the independent sources in the chromatogram. After that, orthogonal signal deconvolution (OSD) is applied to determine S. OSD purpose is to extract and deconvolve the spectrum of a given compound only with the information relative to the compound elution profile - which is previously determined by ICA -. A detailed explanation of ICA–OSD is given elsewhere [2, 10].

Both methods obtain a local C and S matrices, corresponding to the resolution of each window into pure chromatographic profiles and spectra, respectively. Each local C and S matrices are appended to a general C_g and S_g matrices containing the resolution of all the chromatogram. As mentioned before, when using a moving window to resolve the chromatogram, the consecutive windows have to be overlapped to ensure that one compound that could be split on window borders is fully covered by the next window. Then, compounds - or partial eluting profiles when they are split by the window border - are expected to be resolved in more than one window. This leads to multiple duplicates that difficulties the selection of the quantitative - correctly resolved - compound. To ensure only one chromatographic profile and spectrum per compound, a duplicity filter based on minimizing the residual sum of squares (RSS) is proposed. First, a correlation matrix for C_g is determined, and those groups of chromatographic profiles that correlate in more than an user-defined threshold - typically 75 % - are considered that might be duplicated. These groups might be composed of two or more chromatographic profiles. After that, all the possible combinations are considered. As an illustrative example, let us consider that three (N=3) chromatographic profiles C_1, C_2 and C_3 correlate between them. Then, 8 ($2^{N=3}$) possible scenarios are considered. For each scenario, first, a chromatographic matrix D is determined comprising the retention time of the all the considered chromatographic profiles (C_1 – C_3), and after that, a putative D^* matrix is determined by:

$$D^*(k) = C_j S_j^T \tag{2}$$

where $D^*(k)$ is the reconstructed matrix and the subindex j denotes the compounds considered in each k=1,2,...,N case. Then, a RSS for each scenario is determined as follows:

$$RSS(k) = \sum_{i=1}^{N} (D - D^*(k))^2 \tag{3}$$

The scenario with the least RSS is considered to be the combination that best describes the data, and the chromatographic profiles that are not included in this combination are removed from C_g and S_g.

3. Results and discussion

The moving window-based ICA–OSD and MCR–ALS implementations were used to automatically extract and deconvolve the compounds concentration profiles and spectra from all the 25 serum samples. Three different window

lengths were employed: 10, 15 and 20 s, all with an overlap of 50 %. After being resolved, compounds were aligned across samples by an automated alignment algorithm (see Materials and methods section). We will now refer the two methods as ICA–OSD and MCR–ALS-based approaches, and each approach implies its moving window-based application and the subsequent automated alignment. After alignment, average – across samples – compound spectra were compared to the GMD database, providing a putative identification. From among all the identified compounds by the ICA–OSD and MCR–ALS-based approaches, we focused on the 33 compounds known to appear in the samples by name and RT matching with reference values. For each compound, we manually selected a quantitative mzRT feature – from the XCMS output – corresponding to a selective/quantitative ion. Table 1 shows the list of the 33 metabolites with their retention time, the quantification m/z and a linear regression coefficient of determination (R^2) between the proposed methods (ICA–OSD and MCR–ALS-based approaches) and the selective ion area or intensity (reference model) from the XCMS output. In order to demonstrate the ICA–OSD/MCR–ALS-based approaches quantification capability along a wide dynamic range of metabolite concentration, we determined the relative compound concentration (Rel. C.) which is the quotient between the mean concentration of each compound and the mean concentration of all the compounds listed in the table. Those compounds noted with NF (*Not Found*) have not been found by the algorithm: we considered one compound as NF when it was resolved and aligned in less than 9 samples (9 samples correspond to the 80 % of the size of one clinical condition in the original study of this dataset [23]).

From the table, both ICA–OSD and MCR–ALS moving window-based approaches with the subsequent automated alignment shown a comparable quantitative performance. This can also be observed in the R^2 values box plots for each method and window length (Fig. 3). Overall, results shown excellent linear relations (R^2>0.90) for most compounds and methods, including low concentrated and co-eluted compounds, as shown in the resolution examples in Fig. 4. Differences were found, for example, for the case of leucine at 8.75 min, where the ICA–OSD-based approach failed in detecting the compound whereas the MCR–ALS-based approach successfully detected it. Similarly, the ICA–OSD-based approach successfully quantified glycerol at 8.8 min, whereas the MCR–ALS-based approach shown a poor performance.

Also, differences between window lenght were observed. For example, hexanoic acid was not found for the 20 s window size. This can be attributed to the fact that its low concentration (9 %) affected its detection when more compounds were included in a window. Contrarily, in small windows, the compound had relatively more importance - variance - respect the whole window, which allowed its correct detection in the 10 and 15 s windows cases. Similarly, isoleucine and proline eluting at 9.06 and 9.10 min, with relative concentrations of 2 and 9 % respectively, shown no linear relation in the 20 s window length, since its low relative variance when using a longer window lead to their incorrect resolution.

R^2 values varied when comparing area and intensity. Generally, intensity R^2 values shown higher linear relations (Fig. 3). Intensity is expected to be a more robust analysis variable, as the fragments shape can be easily affected by noise or co-elution, whereas in those cases the peak intensity remains more stable. This is because the peak apex is relatively more difficult to be found co-eluted. An example of this is found in lysine eluting at 16.44 min, where

its low concentration hampered its resolution by both methods, leading to low area R^2 values but higher intensity R^2 values.

It is worth noting that not all the compounds were found across all the samples. The number of samples for where each compound was automatedly quantified (resolved and aligned) varied between methods (Supplementary Table S1). For example, hexanoic acid eluting at 5.85 min was quantified by the ICA–OSD-based approach 13 and 18 times for the 10 and 15 s windows cases respectively, whereas the MCR–ALS-based approach successfully quantified it in all the 25 samples. Glycerol was detected by the MCR–ALS-based approach only in 4, 17 and 9 samples for the 10, 15 and 20 s windows cases whereas it was detected in almost all the samples by the ICA–OSD-based approach. We attribute these differences between ICA–OSD or MCR–ALS-based approaches to the fact that the automated alignment algorithm clusters the compounds across samples by taking into account both the retention time distance and spectral similarity, and the alignment outcome varies when a resolution method (ICA–OSD or MCR–ALS) outperforms the other in terms of spectral resolution. A successful alignment (detection and quantification) depends on the quality of the prior resolution. In this context, factor analysis, which was conducted by a singular value decomposition, played an important role in the resolution performance. A lower variance threshold when selecting the number of components would probably modify the number of samples in which compounds were detected. However, a lower threshold also leads to more false positive compounds, i.e., some noise can be modeled as a component. In fact, we attribute the quantitative performance differences between methods to the estimation of the number of compounds.

Moving window length and overlap values conditions the method performance. The moving window length has to enclose at least one compound, and at least twice the value of the minimum compound peak width is the necessary minimum window length. More duplicities and more partial elution profiles resolutions are expected for short window lengths, hampering the subsequent duplicity filter and automated alignment performance. Quantitatively, significant performance differences were not observed as the longer the window is, however, low concentrated compounds were not found when increasing the window length. The results of this study suggest that a valid window length is approximately between 10 and 15 times the minimum peak width – 10 and 15 s respectively for the chromatographic method used. The overlap also has to enclose at least one compound peak width. Overlap percentage can be translated to seconds by its product with the window length.

Aside from the inherent limitations of the multivariate methods and their influence in the posterior alignment performance, additional limitations of the proposed method include that the duplicity filter might fail in selecting which is the best combination of compounds that best describe the data, selecting partially resolved profiles – from compounds that were split by the window – to modelize the chromatogram.

Finally, the most significant difference between ICA–OSD and MCR–ALS is their runtime speed. Whereas the approach based on MCR–ALS resolved an entire chromatogram in approximately 3 min (20 s window size), the ICA–OSD-based approach took approximately 1.4 min. A fast runtime speed is and advantageous feature due to the large amount of data that metabolomics experiments generate, and also because when a moving window approach is

employed, the same data is analyzed twice due to the overlap, which means more consumption of time in comparison with the traditional hard segmentation approaches. Processing was conducted with a 2.4 GHz Intel Core i7 with 16 GB of DDR3 memory at 1333 MHz.

4. Conclusions

Different multivariate methods have been reported in literature for the processing of GC–MS data. However, its application in GC–MS data involve segmenting the chromatogram into regions or windows. Hard chromatographic segmentation is a challenging task as it implies separating between informative data and noise from the chromatogram, and it might lead to failure in the detection of compounds. In this study, we proposed the application of an overlapping moving window-based independent component analysis – orthogonal signal deconvolution (ICA–OSD) and multivariate curve resolution – alternating least squares (MCR–ALS) approaches. We evaluated the proposed methods through their quantification capabilities in comparison with the XCMS package. Results shown that the proposed methodology was able to correctly quantify compounds appearing in biological matrices with the advantage that the automation of the method was not limited only to the resolution, but it included the alignment of compounds across samples. Thus, compound resolution is a critical step that allows the posterior alignment of compounds across samples. Alignment implies registering the concentrations changes among samples from different clinical conditions, which allows obtaining new biological insights. In conclusion, this study introduces an automated data processing pipeline based on an overlapping moving window application of MCR–ALS and ICA–OSD, covering both resolution and alignment of metabolites. Altogether, our results strengthen the suitability of the challenged independent component analysis (ICA) technique for multivariate resolution in analytical chemistry [30], and they demonstrate the robustness of ICA–OSD as a complementary and faster method to MCR–ALS for the automated resolution of GC–MS mixtures in metabolomics experiments.

Acknowledgements: The authors also acknowledge Dr. S. Samino from YanesLab (URV) for providing technical assistance. This research was partially funded by MINECO grant TEC2012–31074 and TEC2015-69076-P (to JB) and TEC2013–44666–R and TEC2014-60337-R (to AP). CIBERDEM and CIBER–BBN are initiatives of the Spanish Instituto de Salud Carlos III (ISCIII). XD acknowledges the University Rovira and Virgili Martí and Franquès grants (2012BPURV-43) for financial support.

[1] A. Erban, N. Schauer, A.R. Fernie and J. Kopka, Nonsupervised Construction and Application of Mass Spectral and Retention Time Index Libraries From Time-of-Flight Gas Chromatography-Mass Spectrometry Metabolite Profiles, in: W. Weckwerth (Ed.), *Metabolomics: Methods and Protocols, Humana Press Inc., Totowa, NJ*, 2007, pp. 19-38.

[2] X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig, J. Brezmes, Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *J. Chromatogr. A.* 1409 (2015) 226–33.

[3] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges *Anal. Chim. Acta* 765 (2013) 28–36.

[4] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods* 6 (2014) 4964–4976.

[5] N. M. Faber, R. Bro, P. K. Hopke, Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometr. Intell. Lab.* 65 (2003) 119–137.

[6] J. M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC–MS problems with PARAFAC2 *Trends in Anal. Chem.* 27 (2008) 714–725.

[7] G. Wang, W. Cai, X. Shao, A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometr. Intell. Lab.* 82 (2006) 137–144.

[8] Z. Liu, W. Cai, X. Shao, Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. *J. Chromatogr. A*, 1190 (2008) 358–364.

[9] X. Shao, Z. Liu, W. Cai, Resolving multi-component overlapping GC-MS signals by immune algorithms. *Trends in Anal. Chem.* 28 (2009) 1312–1321.

[10] X. Domingo-Almenara, A. Perera, N. Ramirez, J. Brezmes, Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics. *Comput. Methods Programs Biomed.* 130 (2016) 135–141.

[11] R. H. Jellema, S. Krishnan, M. M. W. B Hendriks, B. Muilwijk, J. T. W. E. Vogels, Deconvolution using signal segmentation. *Chemometr. Intell. Lab.* 104 (2010) 132–139.

[12] J. M. Amigo, M. J. Popielarz, R. M. Callejon, M. L. Morales, A. M. Troncoso, M. A. Petersen, T. B. Toldam-Andersen, Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J. Chromatogr. A* 1217 (2010) 4422–4429.

[13] S. Furbo, J. H. Christensen, Automated peak extraction and quantification in chromatography with multichannel detectors. *Anal. Chem.* 84 (2012) 2211–2218.

[14] L. G. Johnsen, J. M. Amigo, T. Skov, R. Bro, Automated resolution of overlapping peaks in chromatographic data. *J. Chemom.* 28 (2014) 71–82.

[15] V. Jochen, G. de Revel, S. Krieger-Weber, D. Rauhut, M. du Toit, A. de Villiers, Toward automated chromatographic fingerprinting: A non-alignment approach to gas chromatography mass spectrometry data. *Anal. Chim. Acta* 911 (2016) 42–58.

[16] Z. Zeng, C. Xu, Y. Liang, B. Li, Sectional moving window factor analysis for diagnosing elution chromatographic patterns. *Chemometr. Intell. Lab.* 69 (2003) 89–101.

[17] Z-D. Zeng, Y-Z. Liang, Y-L. Wang, X-R. Li, L-M. Liang, Q-S. Xu, C-X. Zhao, B-Y. Li, F-T. Chau, Alternative moving window factor analysis for comparison analysis between complex chromatographic data. *J. Chromatogr. A* 1107 (2006) 273–285.

[18] D-L. Yuan, L-Z. Yi, Z-D. Zeng, Y-Z. Liang, Alternative moving window factor analysis (AMWFA) for resolution of embedded peaks in complex GC–MS dataset of metabonomics/metabolomics study. *Anal. Methods* 2 (2010) 1125–1133.

[19] S. Lopez-Urena, M. Beneito-Cambra, R. M. Donat-Beneito, G. Ramis-Ramos, Overlapped moving windows followed by principal component analysis to extract information from chromatograms and application to classification analysis. *Anal. Methods* 7 (2015) 3080–3088.

[20] S. Oller-Moreno, G. Singla-Buxarrais, J.M. Jimenez-Soto, A. Pardo, R. Garrido-Delgado, L. Arcec, S. Marco, Sliding window multi-curve resolution: Application to gas chromatography–ion mobility spectrometry. *Sens. Actuators B Chem.* 217 (2015) 13–21.

[21] C.A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78 (2006) 779–787.

[22] R. Tautenhahn, G. J. Patti, D. Rinehart, G. Siuzdak, XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* 84 (2012) 5035–5039.

[23] X. Domingo-Almenara, J. Brezmes J, M. Vinaixa, S. Samino, N. Ramirez, M. Ramon-Krauel, C. Lerin, M. Diaz, L. Ibanez, X. Correig, A. Perera-Lluna, O. Yanes, eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC–MS-based metabolomics. *Anal. Chem.* 88 (2016) 9821–9829.

[24] J. Diewok, A. de Juan, M. Maeder, R. Tauler, B. Lendl, Application of a combination of hard and soft modeling for equilibrium systems to the quantitative analysis of pH - modulated mixture samples. *Anal. Chem* 75 (2003) 641–647.

[25] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F* 140 (1993) 362–370.

[26] K. Nordhausen, J-F. Cardoso, J. Miettinen, H. Oja, E. Ollila and S. Taskinen, JADE: Blind Source Separation Methods Based on Joint Diagonalization and Some BSS Performance Criteria *R package* version 1.9-93 (2015). https://CRAN.R-project.org/package=JADE.

[27] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics* 6 (2010) 322–333.

[28] Hummel, J.; Selbig, J.; Walther, D.; Kopka, J, The Golm Metabolome Database: a database for GC–MS based metabolite profiling. *Metabolomics* 18 (2007) 75–95.

[29] I. H. M. van Stokkum, K. M. Mullen, V. V. Mihaleva, Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS. *Chemometr. Intell. Lab.* 95 (2009) 150–163.

[30] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? *Trends Anal. Chem.* 31 (2012) 134–143.

Figures Captions:

Figure 1. Scheme showing the data processing pipeline. After a preprocessing step (1), compounds are resolved through the moving window-based applications (2), and resolved compounds are posteriorly aligned across samples (3). Finally, identification and quantification of aligned compounds (4) allow outputing a compound list with metabolite putative identifications and relative concentration among samples.

Figure 2. In (a), illustration of the moving window approach. A fixed-length window is displaced with a certain overlapp along the chromatogram. The blue lines represent each m/z (extracted ion chromatogram). Each chromatographic window (b) is resolved by ICA–OSD or MCR–ALS into pure chromatographic profiles and spectra. This case shows the resolution of (i) glycerol and (ii) phosphoric acid, which appear strongly co-eluted. For this case, the extracted ion chromatogram is shown in grey, whereas colored solid lines represent the resolved chromatographic profiles. Compound resolved spectra are shown in color red and green along with each GMD reference spectrum negatively rotated in the same axis and shown in black. In this example, the resolved spectra of both phosphoric acid and glycerol seem to be affected by the strong co-elution in which they appear.

Figure 3. R^2 values box plots from Table 1, for both area and intensity, and for the three window lengths and both ICA-OSD and MCR-ALS-based approaches.

Figure 4. Three examples of resolution (one per column) by ICA-OSD for 6 metabolites appearing co-eluted and/or at low concentrations (left, middle and right) for the 15 s window length case. Top row shows the unaligned total ion chromatograms (TIC) per sample, where each line is each TIC in a chromatogram. Bottom row shows the resolved and aligned chromatographic profiles, where each colored line represents a resolved chromatographic profile in a sample, and each color denotes a different compound. Compound profiles were correctly resolved and aligned in all the cases. Of note, isoleucine and proline, although they do not appear co-eluted between them, they appear in co-elution with two other unknown compounds. Also, due to their low concentration, some of isoleucine and proline recovered compound profiles shown a more noisy shape.

Table 1: Retention time (Rt), quantitative fragment ion (m/z) (XCMS), relative concentration (Rel. C) of 33 compounds. Coefficients of determination (R^2) of the regression between the area and intensity of the resolved chromatographic profile (ICA–OSD and MCR–ALS) and the quantitative ion peak (XCMS) is shown. WL and NF stands for *window length* and *not found* respectively. The number of trimethylsilyl (TMS) derivatives groups are not shown, with the exception of those compounds that appear duplicated. For those cases, the number of trimethylsilyl (TMS) groups is shown in brackets.

					R^2											
					ICA–OSD						MCR–ALS					
Cp.	Rt	m/z	Rel.C	Name	WL=10		WL=15		WL=20		WL=10		WL=15		WL=20	
	(min)		(%)		Area	Int	Area	Int	Area	Int	Area	Int	Area	Int	Area	Int
1	5.73	117	1055	Lactic acid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	5.85	173	9	Hexanoic acid	0.41	0.88	0.71	1.00	NF	NF	0.87	1.00	0.93	1.00	NF	NF
3	6.11	72	65	Valine (1)	0.93	1.00	0.98	1.00	0.98	1.00	0.85	1.00	0.95	1.00	0.95	1.00
4	6.53	113	5	Hydroxylamine	0.79	0.74	0.75	0.74	0.74	0.74	0.79	0.70	0.82	0.71	0.80	0.71
5	6.69	131	18	2-hydroxybutyrate	0.80	1.00	0.96	1.00	1.00	1.00	0.97	1.00	0.99	1.00	0.98	1.00
6	7.08	86	33	Leucine (1)	0.97	1.00	0.90	1.00	0.00	0.00	0.93	1.00	0.95	1.00	0.85	1.00
7	7.15	191	19	3-hydroxybutyrate	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	7.97	145	4	Valine (2)	0.32	0.95	0.97	0.97	0.95	0.97	0.97	0.97	0.99	0.97	0.98	0.97
9	8.19	130	206	Urea	0.90	0.91	0.96	0.91	0.92	0.94	0.96	0.89	0.95	0.93	0.95	0.94
10	8.36	179	133	Benzoic acid	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00
11	8.55	116	24	Serine	0.90	0.98	0.98	1.00	0.99	1.00	0.98	1.00	0.98	1.00	0.99	1.00
12	8.75	158	18	Leucine (2)	NF	NF	NF	NF	NF	NF	0.98	1.00	NF	NF	NF	NF
13	8.80	205	209	Glycerol	0.99	1.00	1.00	1.00	0.99	1.00	NF	NF	0.31	0.31	0.85	0.90
14	8.81	299	649	Phosphoric acid	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.75	0.61	0.73	0.80	0.82
15	9.06	218	2	Isoleucine	0.82	0.90	0.79	0.91	0.00	0.00	0.97	0.94	0.92	0.93	0.00	0.00
16	9.10	142	9	Proline	0.99	1.00	1.00	1.00	0.00	0.00	0.99	1.00	0.99	1.00	0.00	0.00
17	9.58	189	5	Glyceric acid	0.97	1.00	0.98	1.00	0.99	0.99	0.98	0.99	0.98	0.99	0.98	0.99
18	9.85	215	14	Nonanoic acid	0.96	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.99	0.99	0.99
19	10.34	291	7	Threonine	0.92	0.96	0.89	0.92	0.98	0.93	0.94	0.97	0.98	0.97	0.98	0.93
20	11.82	230	3	3-hydroxy-trans-proline	0.87	0.99	0.82	1.00	0.95	1.00	0.81	0.99	0.95	0.99	0.95	1.00
21	12.02	156	216	Glutamic Acid (2)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	12.73	142	14	Proline [+CO2]	0.92	0.99	0.97	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
23	13.17	246	6	Glutamic acid (3)	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99	1.00
24	13.27	218	24	Phenylalanine	0.99	1.00	0.98	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
25	13.42	117	101	Dodecanoic acid	0.98	1.00	0.97	1.00	0.98	1.00	0.99	0.99	0.99	0.99	0.99	0.99
26	15.39	142	16	Ornithine	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
27	15.46	273	15	Citric acid	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00
28	15.54	285	22	Tetradecanoic acid	0.96	0.92	0.88	0.85	0.95	0.92	0.97	0.93	0.96	0.87	0.98	0.94
29	16.44	229	<1	Lysine	0.49	0.87	0.50	0.87	0.25	0.79	0.10	0.78	0.11	0.78	0.24	0.78
30	17.48	129	324	Hexadecanoic acid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
31	18.23	305	30	Myo-inositol	0.88	0.99	0.91	1.00	0.93	1.00	0.94	0.99	0.87	0.97	0.70	0.98
32	18.24	441	11	Uric acid	0.99	0.98	0.99	0.99	1.00	0.98	0.91	0.90	0.88	0.91	0.82	0.86
33	19.26	356	16	Octadecanoic acid	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99