Journal of Chromatography A, xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Short communication

Avoiding hard chromatographic segmentation: A moving window approach for the automated resolution of gas chromatography–mass spectrometry-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

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 24 October 2016
Accepted 25 October 2016
Available online xxx

Keywords:
Gas chromatography
Orthogonal signal deconvolution
Multivariate curve resolution
Moving window
Independent component analysis
Metabolomics

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.

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 (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 [7–9]. 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

* Corresponding author at: Metabolomics Platform – IISPV, Department of Electrical and Automation Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain.
E-mail address: xavier.domingo@urv.cat (X. Domingo-Almenara).

http://dx.doi.org/10.1016/j.chroma.2016.10.066
0021-9673/

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 for GC–MS data processing 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 [11–15]. 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 [16–19]. 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 among 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 the best of 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. This approach avoids hard segmentation or windowing of the chromatogram. We propose a duplicity filter based on the minimization of the residual sum of squares 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.

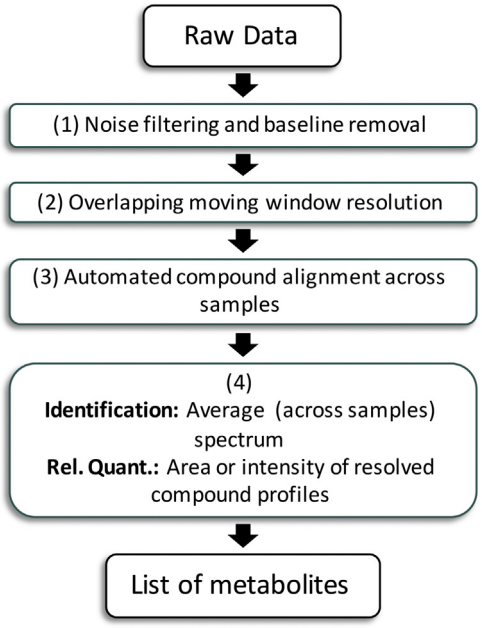

Fig. 1. Scheme showing the data processing pipeline. After a preprocessing step (1), compounds are resolved through the moving window-based applications (2), and resolved compounds are posteriorly aligned across samples (3). Finally, identification and quantification of aligned compounds (4) allow outputting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

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] for GC–MS data processing, 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 × 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 35–700 with an acquisition rate of 5 spectra per second.

2.2. Data processing workflow

The data processing 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 window lengths, 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

Please cite this article in press as: X. Domingo-Almenara, et al., Avoiding hard chromatographic segmentation: A moving window approach for the automated resolution of gas chromatography–mass spectrometry-based metabolomics signals by multivariate methods, J. Chromatogr. A (2016), http://dx.doi.org/10.1016/j.chroma.2016.10.066