Journal of Chromatography A, 1409 (2015) 226–233

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation

Xavier Domingo-Almenara[a,c,*], Alexandre Perera[b], Noelia Ramírez[a,c], Nicolau Cañellas[a,c], Xavier Correig[a,c], Jesus Brezmes[a,c]

[a] Metabolomics Platform – IISPV, Department of Electrical and Automation Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[b] B2SLAB, Department d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, CIBER-BBN, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain
[c] CIBERDEM, Biomedical Research Networking Center in Diabetes and Associated Metabolic Disorders, Madrid, Spain

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 9 July 2015
Accepted 10 July 2015
Available online 15 July 2015

Keywords:
Compound deconvolution
Independent component analysis
Multivariate curve resolution
Orthogonal signal deconvolution
Mass spectrometry
Metabolomics

ABSTRACT

Metabolomics GC–MS samples involve high complexity data that must be effectively resolved to produce chemically meaningful results. Multivariate curve resolution-alternating least squares (MCR-ALS) is the most frequently reported technique for that purpose. More recently, independent component analysis (ICA) has been reported as an alternative to MCR. Those algorithms attempt to infer a model describing the observed data and, therefore, the least squares regression used in MCR assumes that the data is a linear combination of that model. However, due to the high complexity of real data, the construction of a model to describe optimally the observed data is a critical step and these algorithms should prevent the influence from outlier data. This study proves independent component regression (ICR) as an alternative for GC–MS compound identification. Both ICR and MCR though require least squares regression to correctly resolve the mixtures. In this paper, a novel orthogonal signal deconvolution (OSD) approach is introduced, which uses principal component analysis to determine the compound spectra. The study includes a compound identification comparison between the results by ICA-OSD, MCR-OSD, ICR and MCR-ALS using pure standards and human serum samples. Results shows that ICR may be used as an alternative to multivariate curve methods, as ICR efficiency is comparable to MCR-ALS. Also, the study demonstrates that the proposed OSD approach achieves greater spectral resolution accuracy than the traditional least squares approach when compounds elute under undue interference of biological matrices.

1. Introduction

The analysis of samples from a metabolomics perspective allows the phenotyping of organisms at a molecular level [1]. At the same time, metabolomics provides a means of detecting early biochemical changes in organisms before the appearance of a disease and thus, a means of finding predictive biomarkers [2]. Among the analytical techniques used in metabolomics, gas chromatography–mass spectrometry (GC–MS) is a well stablished platform due to its robustness and its applicability to a wide range of matrices and metabolites through silylation of the polar groups.

Because of the high complexity of biological fluids, the complete chromatographic resolution of all the metabolites in a sample cannot be easily achieved as the co-elution of two or more of them usually occurs. The correct identification of co-eluted compounds depends mostly on the degree of the chromatographic separation and their spectral dissimilarity. Likewise, the metabolites in the samples usually occur at low concentrations and the background signal, inherent in the instrument and the sample biological matrix, interferes in their correct identification and quantification. The use of resolution algorithms, which can help extract the purest compound elution profile and spectra, is mandatory for GC–MS data processing.

One of the best-established algorithms for application to chromatographic data to resolve co-eluted compounds is multivariate curve resolution-alternating least squares (MCR-ALS) [3,4]. MCR-ALS can resolve a mixture of compounds into a pure concentration profile matrix and a pure spectra matrix [5]. In recent years, a blind source separation (BSS) technique known as independent

* Corresponding author at: Metabolomics Platform – IISPV, Department of Electrical and Automation Engineering (DEEEA), Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain. Tel.: +34 977559619; fax: +34 977559605.
E-mail address: xavier.domingo@urv.cat (X. Domingo-Almenara).

http://dx.doi.org/10.1016/j.chroma.2015.07.044
0021-9673/

component analysis (ICA) [6], already widely applied for the resolution of spectroscopic mixtures [7–11], has also been applied for the resolution of GC–MS samples [12]. In a GC–MS chromatogram, the compounds elution profiles appear mixed with their respective spectra. In these cases, ICA-based approaches are able to recover the different independent sources contained in data and, eventually, resolve GC–MS data. MCR-ALS approaches this problem by minimizing the residual error between the data and the predicted model, whereas ICA focuses on estimating the original sources – or components – by maximizing their statistical independence. Actual ICA-based methods to resolve chromatographic data include mean-field ICA (MF-ICA) [13], post-modification based on chemical knowledge (PBCK) [14], window ICA (WICA) [15] and non-negative ICA [16]. Artificial immune system algorithms involving the use of ICA have also been proposed [17]. The first step of the resolution procedure in these methods is the use of ICA to resolve the mass spectrum for each compound in the mixture. The above-mentioned algorithms use different approaches to determine the elution profile of each compound, since the elution profiles determined by ICA tend to be inaccurate or affected by various ICA ambiguities such as negativity or variance (energy) indetermination [18]. Recently, these ICA-based methods were compared with MCR for the resolution of GC–MS data by Parastar and co-workers [19] who showed that the ICA-based resolutions methods show the same performance than MCR. A natural extension of ICA to recover co-eluted profiles might be independent component regression (ICR), which was first used to resolve mixtures in near infrared (NIR) spectra by Shao et al. [20], but whose efficiency on GC–MS data treatment has not yet been studied.

The use of least squares (LS) regression, common to most algorithms in GC–MS data resolution, has a major drawback, induced by the inherent correlation between ions related to the same compound. This correlation yields an ion-redundancy which means that, for each compound, different ions, also called fragments or m/z, elute at the same retention time and with the same elution profile. When fitting the elution profiles to data, no correlation information between the ions is taken into account, so the LS regression does not distinguish between noise and the compound ions that are being regressed; this may introduce a bias into the LS regressors. This effect includes instrumental or experimental noise as baseline, peak-tailing, or compound co-elution. The performance of the resolution of mixtures with least squares may, therefore, depend on the correct estimation of the underlying model from the data.

This study proposes the use of ICR for GC–MS compound identification. In this approach, we integrate ICA and MCR with a novel orthogonal spectra deconvolution (OSD) as an alternative to least squares regression with a view to improve the determination of the compound spectra when compounds elute under the interference of a biological matrix.

2. Materials and methods

This section describes MCR-ALS, ICR and their variants integrated with the OSD algorithm (ICA-OSD and MCR-OSD). The proposed methods were evaluated by comparing the resolution of the spectra of 38 compounds in a pure standards sample and 25 compounds in a human serum sample. A match score between the resolved and the reference spectra was determined for each compound and method. The samples were processed by MCR, the proposed ICR, both ICA and MCR using the OSD approach (ICA-OSD and MCR-OSD). The goal was to use the different methods compared in this study to extract the most pure spectra for each compound. The spectra extracted were matched against a reference MS spectra database. For this study, the Golm Metabolome Database (GMD) [21] was used as a reference database.

2.1. Materials

A set of four pure standards samples – four sample repetitions – and a total of eight biological samples – four sample repetitions of a human serum sample, and two repetitions of two human urine samples from healthy volunteers – were used for evaluation. The standard mixture was composed of 26 metabolites (see Table S1 of the Supplementary Material) previously found in the human serum and urine metabolome [22]. First, all samples were characterized by a curated identification of the reference compounds (standards). The pure standards samples were taken as a reference to later identify the same compounds in the human serum and urine samples. Two compounds identified in the biological samples that are not included in the pure standards set were validated also analyzing their corresponding standard references.

The metabolites of the human serum and urine samples were extracted and derivatized following a standard protocol [23] with slight modifications to optimize the process. Extracts were analyzed using a 7890 gas chromatograph from Agilent (Palo Alto, CA, USA) coupled to a Pegasus IV TOF/MS from Leco (St. Joseph, MI, USA) using a DB5-MS capillary column (30 m × 0.25 mm × 0.25 μm, 5% diphenyl, 95% dimethylpolysiloxane) from Agilent. Analyses were performed by injecting 1 μL of the extracts into a split/splitless inlet at 250 °C with a split flow of 5 mL min^{-1} and a helium constant flow of 1 mL min^{-1} (99.999%, Abelló Linde, Barcelona). The oven temperature of the GC was initially held at 50 °C for 1 min, then raised to 285 °C at a rate of 20 °C min^{-1} and held at that temperature for 5 min. The GC-TOF/MS interface was set at 280 °C and the ion source at 250 °C. The mass spectrometer acquired m/z ratios between 35 and 600 amu at 10 Hz and an electron impact energy of 70 eV.

2.2. Data pre-processing and analysis

In order to analyze an entire dataset using the MCR or ICA-based approaches, each chromatogram was divided in chromatographic peak features (CPFs) using the same criteria as in [24]. The different CPFs contained several compounds, so the algorithm had to deconvolve them in case of co-elution. The number of factors or components used to initialize both MCR and ICA was determined by cross-validation (described in Section 2.6). A unimodality constraint [25] was applied to the resolved profiles and the same non-negative least squares algorithm was applied for both MCR and ICR. The simple mean spectra determined either by ICA-OSD, MCR-OSD, ICR or MCR in the different samples for each compound were compared using the dot product [26] against the GMD MS spectra database.

The masses 73, 74, 75, 147, 148, and 149 m/z were excluded before processing the sample, since they are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety [21]. They were also excluded in the identification. Only the fragments from m/z 70 to 600 were taken into account when comparing reference and empirical spectra, since this is the m/z range included in the downloadable GOLM database. Also, the human serum and urine samples signal was filtered using a Savitzky–Golay filter [27] and the baseline was removed using a semi-supervised spline interpolation to reduce the interaction of the biological matrix (described in Section 3.2). The ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [28].