Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Toward automated chromatographic fingerprinting: A non-alignment approach to gas chromatography mass spectrometry data[☆]

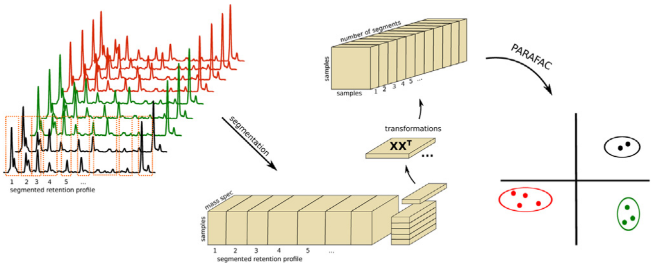
Jochen Vestner[a, b, c, *], Gilles de Revel[a, b], Sibylle Krieger-Weber[d], Doris Rauhut[c], Maret du Toit[e], André de Villiers[f]

[a] *Université de Bordeaux, ISVV, EA 4577, Unité de recherche Œnologie, 33882 Villenave d'Ornon, France*
[b] *INRA, ISVV, USC 1366 Œnologie, 33882 Villenave d'Ornon, France*
[c] *Department of Microbiology and Biochemistry, Hochschule Geisenheim University, Von-Lade-Straße 1, 65366 Geisenheim, Germany*
[d] *Lallemand, In den Seiten 53, 70825 Korntal-Münchingen, Germany*
[e] *Institute of Wine Biotechnology, Department of Viticulture and Oenology, Stellenbosch University, Private Bag X1, Matieland (Stellenbosch) 7602, South Africa*
[f] *Department of Chemistry and Polymer Science, Stellenbosch University, Private Bag X1, Matieland (Stellenbosch) 7602, South Africa*

HIGHLIGHTS

- A novel data processing procedure for non-targeted gas chromatography mass spectrometry (GC–MS) data is proposed.
- Basic matrix manipulation of segmented GC–MS chromatograms and PARAFAC multi-way modelling is used.
- Retention time shifts and peak shape deformations between samples are taken into account.
- The procedure is demonstrated on an artificial and an experimental full-scan GC–MS data set.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form
14 January 2016
Accepted 19 January 2016
Available online 22 January 2016

Keywords:
Non-targeted analysis
Gas chromatography
Fingerprinting
Multi-way analysis
Metabolomics
Non-alignment

ABSTRACT

In contrast to targeted analysis of volatile compounds, non-targeted approaches take information of known and unknown compounds into account, are inherently more comprehensive and give a more holistic representation of the sample composition. Although several non-targeted approaches have been developed, there's still a demand for automated data processing tools, especially for complex multi-way data such as chromatographic data obtained from multichannel detectors. This work was therefore aimed at developing a data processing procedure for gas chromatography mass spectrometry (GC–MS) data obtained from non-targeted analysis of volatile compounds. The developed approach uses basic matrix manipulation of segmented GC–MS chromatograms and PARAFAC multi-way modelling. The approach takes retention time shifts and peak shape deformations between samples into account and can be done with the freely available N-way toolbox for MATLAB. A demonstration of the new

[☆] Selected paper from the 9th Edition of the In Vino Analytica Scientia Symposium in Trentino Alto Adige (Italy) in 14–17 July 2015.

[*] Corresponding author. Université de Bordeaux, ISVV, EA 4577, Unité de recherche Œnologie, 33882 Villenave d'Ornon, France.

E-mail address: jochen.vestner@gmx.de (J. Vestner).

http://dx.doi.org/10.1016/j.aca.2016.01.020

fingerprinting approach is presented using an artificial GC–MS data set and an experimental full-scan GC–MS data set obtained for a set of experimental wines.

1. Introduction

Non-targeted analysis has increasingly gained importance in numerous domains of analytical chemistry such as life science, food science and especially the '-omics' related sciences. In contrast to conventional targeted analysis, non-targeted analysis aims to gather qualitative and quantitative information on as many compounds as possible in the analysed samples in a short period of time, and thus to provide the researcher with a more holistic view of the composition of samples [1]. Holistic strategies benefit from the vast amount of information obtained from modern analytical instrumentation. However, the main challenges are data handling and full exploitation of dimensionality of the acquired data.

The data generated by hyphenated chromatographic techniques such as GC–MS or LC–MS are especially information rich. Feature extraction such as peak integration in single ion chromatograms, total ion chromatograms or deconvoluted signals are the most common approaches to extract information from chromatographic data and result in relatively small data tables which are straightforward to analyse [2–8]. Although various peak integration algorithms and software packages have been developed [9–12], automated peak integration remains troublesome due to coelution and potential erroneous peak integration and/or assignment. Time consuming manual correction of the results is often necessary. Moreover, relevant information from the raw data can be lost due to such feature extraction before modelling [13,14]. Deconvoluting chromatographic signals can also be time-consuming in terms of model construction and evaluation of results [15,16,2,17].

An alternative, more comprehensive approach aiming at the extraction of more information and underlying patterns in the data involves the usage of the two dimensional raw data signal of each sample in entirety as a chromatographic fingerprint for modelling. Examples for holistic non-targeted analyses can be found in numerous reports [14,18–25], some of which also include the application of multi-way analysis methods such as TUCKER3, PARAFAC and N-PLS to hyphenated chromatographic data. When factor models are used on chromatographic data, challenges are associated with the increased size of data and the handling of shifts and peak shape deformation, which result in distortion of the bilinear/trilinear structure of the data. Several algorithms and software programmes have been developed for peak alignment [26–30]. Depending on the data, shift correction can, however, be difficult and time-consuming.

The above described problems of conventional data analysis approaches to non-targeted GC–MS analysis, in particular challenges with automated peak integration and retention time alignment of chromatograms, were the main motivation for the development of an alternative data analysis approach. The major consideration to overcome the peak integration issue was the direct modelling of the chromatographic raw data (without feature selection), including a reduction of the data. The main idea to master the distortion of bilinear/trilinear structure of the data due to shifting peaks was the consideration of a mathematical transformation of pieces (segments) of the chromatograms using sums of squares and cross product (SSCP) matrices. SSCP matrices are positive, squared and symmetric, similar to variance-covariance matrix [31], which are utilised for instance in PARAFAC2, STATIS and the calculation of R_V-coefficients [32,19,33–35]. Particularly the indirect fitting algorithm for PARAFAC2 [36] served as major inspiration for the development of the new approach. Moreover, for the sake of simplicity another aim was to use a single model for the entire set of chromatograms of all samples to find systematic differences among samples and to identify important regions of the chromatograms which, if desired, can be further deconvoluted and investigated using e.g. PARAFAC2. A method using multiple PARAFAC2 models on segmented chromatograms has been reported recently [37]. This approach gives very detailed information on fully decomposed mass spectra and peak profiles, which are finally summarized using PCA. The here described new approach can be considered as a 'segment pre-selection tool' for subsequent deconvolution of only important chromatogram segments. By this means a significant amount of time used for the construction and evaluation of PARAFAC2 models can so be saved.

This paper gives an overview on the algorithm of the new data analysis approach, including the theoretical background such as the calculation of SSCP matrices and all other mathematical transformations used. The approach is explained and tested on an artificial, well defined GC–MS data set with and without peak shifts. After the theoretical discussion, the approach is tested on a real GC–MS dataset of experimental wines and results are confirmed using a reference method for data analysis approaches including PARAFAC2 deconvolution and peak integration of deconvoluted peak profiles of the entire segmented chromatograms with subsequent PCA on the obtained peak table.

2. Theory

2.1. Defined, artificial GC–MS data set

To demonstrate and verify the developed algorithm a defined, artificial GC–MS data set was created using an in-house developed MATLAB script. The data set consists of 20 chromatograms, each containing 9 to 10 Gaussian peaks with different mass spectra (mz 35 to mz 318) and different degrees of overlapping. The whole chromatogram can be divided into five segments. Segment one contains two peaks which perfectly overlap. Peaks three and four partially coelute in segment two, which is also the case for the peaks five, six and seven in segment three. Peak eight is in segment four and the last segment contains the last two peaks nine and ten, which also partially coelute (Fig. 14 in Supporting Information). Peak sizes vary among chromatograms as indicated in Table 1, consequently samples can be divided into four groups. Moreover, a small random variation was added to all peak sizes to simulate a natural deviation of measurements. To simulate baseline noise a random normal distributed noise was added to the whole data set. Each chromatogram can be considered as a matrix of dimensions

Table 1
Differing peaks among samples in the defined, artificial GC–MS data set.

Segment	Peak no.	Size difference	Sample no.
1	2	only present in	14 & 15
2	4	0.7× higher in	1 to 5
5	9	3× higher in	1 to 10