Journal of Chromatography A, 1503 (2017) 57–64

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Gas chromatography – mass spectrometry data processing made easy

Lea G. Johnsen[a,*], Peter B. Skou[b], Bekzod Khakimov[b], Rasmus Bro[b]

[a] MS-Omics, Birkehegnet 40, Ålsgårde, Denmark
[b] Copenhagen University, Thorvaldsensvej 40, Frederiksberg, Denmark

ARTICLE INFO

Article history:
Received 22 January 2017
Received in revised form 24 April 2017
Accepted 25 April 2017
Available online 27 April 2017

Keywords:
PARAFAC2
Chromatography
Data processing
Deconvolution
GC–MS

ABSTRACT

Evaluation of GC–MS data may be challenging due to the high complexity of data including overlapped, embedded, retention time shifted and low S/N ratio peaks. In this work, we demonstrate a new approach, PARAFAC2 based Deconvolution and Identification System (PARADISe), for processing raw GC–MS data. PARADISe is a computer platform independent freely available software incorporating a number of newly developed algorithms in a coherent framework. It offers a solution for analysts dealing with complex chromatographic data. It allows extraction of chemical/metabolite information directly from the raw data. Using PARADISe requires only few inputs from the analyst to process GC–MS data and subsequently converts raw netCDF data files into a compiled peak table. Furthermore, the method is generally robust towards minor variations in the input parameters. The method automatically performs peak identification based on deconvoluted mass spectra using integrated NIST search engine and generates an identification report. In this paper, we compare PARADISe with AMDIS and ChromaTOF in terms of peak quantification and show that PARADISe is more robust to user-defined settings and that these are easier (and much fewer) to set. PARADISe is based on non-proprietary scientifically evaluated approaches and we here show that PARADISe can handle more overlapping signals, lower signal-to-noise peaks and do so in a manner that requires only about an hours worth of work regardless of the number of samples. We also show that there are no non-detects in PARADISe, meaning that all compounds are detected in all samples.

1. Introduction

In chromatographic methods, such as gas or liquid chromatography coupled with mass spectrometry detectors, the goal is to identify compounds and compare their concentrations across and within samples. To achieve this goal, data processing must fulfil two criteria: (I) it must correctly determine the mass spectrum of the individual compounds for identification and; (II) it must accurately calculate the abundance of chromatographic peaks corresponding to those compounds in each sample. These two tasks are often challenging and time consuming mainly due to the co-elution of chromatographic peaks within a single chromatogram, as well as retention time (RT) shift of peaks across samples. These two challenges lead to mixed mass spectra and complicates compound identification and quantification. For these reasons processing of GC–MS data is challenging using currently available techniques that may perform inadequately both with respect to identification and quantification leading to compounds being wrongly interpreted or simply left undetected.

Most traditional vendor software quantifies compounds based on peak area or height using total ion count (TIC), base peak chromatogram (BPC) or from the extracted ion chromatogram (EIC) by selecting m/z value(s) typical for the given compound. These approaches are susceptible to co-eluting compounds since a contribution to the signal from other compounds is not adequately handled and may significantly affect both quantitative and qualitative results. Furthermore, it is challenging to estimate baseline contributions and this may also lead to errors in quantification. Most of currently applied approaches use simple subtraction of background from nearby baseline or a shoulder of a given peak of interest. Often this is not sufficient to handle overlapping and/or co-eluting peaks.

A more recent approach dealing with overlapping signals is to model the signals using e.g. Gaussian curves [1]. However, these models are not unique [2], instead, a number (actually infinitely many) of completely different sets of Gaussian peaks can model the data equally well. Hence, the solution becomes arbitrary. The development of the software package Automatic Mass spectral Deconvolution and Identification System (AMDIS) [3] was a big step towards resolving complex data. AMDIS automatically calculates the area of the deconvoluted component in terms of the area of the reconstructed total ion current (TIC) chromatogram. AMDIS is freely available standalone software, and is also implemented in

* Corresponding author.
E-mail addresses: lgj@msomics.com (L.G. Johnsen), peter.b.skou@food.ku.dk (P.B. Skou), bzo@food.ku.dk (B. Khakimov), rb@life.ku.dk (R. Bro).

http://dx.doi.org/10.1016/j.chroma.2017.04.052

commercial software like Masshunter (Agilent Technologies, USA). Another commercial software is ChromaTOF (LECO Inc., USA) that became a common tool to process GC–MS data based on a Time-Of-Flight (TOF) mass analyser. Like in AMDIS, ChromaTOF performs automatic deconvolution of peaks from each sample separately and compares the deconvoluted spectra against integrated libraries. Estimation of the peak area in ChromaTOF can either be based on the TIC, BPC, deconvoluted mass spectra or any *m*/*z* ion(s) that are defined by the user. ChromaTOF utilises a proprietary deconvolution technique, but it requires several input parameters, concerning noise level, peak width, retention time shift allowance and more, to be set by the user depending on the sample type and data quality. After peak detection, ChromaTOF can generate the final metabolite table by aligning peaks across samples based on user defined parameters such as RT shift window, noise level, spectral similarity and how often peaks are detected among investigated samples. Both AMDIS and ChromaTOF perform calculations on each sample independently of the other samples.

A completely different approach for handling co-elution and retention time shifts, is to use the so-called PARAllel FACtor analysis2 (PARAFAC2) model [2,4]. PARAFAC2 is able to deconvolute co-eluted, retention time shifted and low signal-to-noise (S/N) ratio chromatographic peaks for all investigated samples in a given retention time region simultaneously [2]. In contrast to other methods, the PARAFAC2 approach only requires a single parameter to be set by the user prior to achieving sufficient data processing for the given retention time region of the chromatogram. This parameter is the number of factors (or real chemical compounds) in the investigated region of the chromatogram. There are simple methods for determining this number as will be explained later. PARAFAC2 modelling allows extraction of the pure spectra of co-eluting compounds as well as it simultaneously computes their peak areas (relative concentrations). The compounds are quantified using the entire pure spectrum and retention time region corresponding to a specific peak. It has previously been shown that PARAFAC2 is superior to commercial solutions [5,6]. However, current implementations of PARAFAC2 are not accessible for non-mathematical users and requires extensive coding for efficient use. Here, we develop an integrated approach called PARAFAC2 based Deconvolution and Identification System (PARADISe), which combines workflow from raw data inspection to metabolite (relative) quantification and identification in a graphical user interface (GUI). Within the PARADISe approach, we included tools required in all steps of the GC–MS data processing; 1) data visualization, 2) division of data into retention time intervals, 3) PARAFAC2 based deconvolution of peaks, 4) validation and extraction of deconvoluted peaks, 5) identification of compounds from raw as well as deconvoluted mass spectra using NIST search engine and NIST mass spectra library and/or any other libraries in NIST format, 6) generation of the final metabolite table. In the following sections, several examples are provided illustrating the power and limits of PARADISe.

2. Materials and methods

2.1. Preparation of a standard mixture sample

Ten chemical compounds including valine, alanine, serine, threonine, *gamma*-aminobutyric acid (GABA), ascorbic acid, fumaric acid, citric acid, gallic acid and *p*-hydroxyphenylacetic acid were used to prepare a standard mixture sample. Compounds were purchased from Sigma-Aldrich (Sigma-Aldrich Denmark A/S, DK) at the highest available purity. The standard mixture sample was prepared by mixing equal volumes of 20.0 mM solutions of compounds in milliQ water. Thus, in the final standard mixture sample the concentration of each compound was 2.0 mM, which was used for preparation of ten different dilution series samples where concentration of each compound ranged from 0.05 to 0.6 mM.

2.2. GC–MS analysis of standard mixture samples

Prior to GC–MS analysis 30 μL of each dilution series samples were dried using ScanVac (Labogene, DK) at 40 °C inside 150 μL glass inserts, sealed with air tight magnetic lids into GC–MS vials and derivatized by addition of 30 μL trimethylsilyl cyanide (TMSCN) [7]. All steps involving sample derivatization and injection were automated using a Dual-Rail MultiPurpose Sampler (MPS) (Gerstel, GmbH & Co. KG, DE). Following reagent addition, the sample was transferred into the agitator of the MPS and incubated at 40 °C for 40 min at 750 rpm. This procedure ensures precise derivatization time and reproducible sample injection. Immediately after derivatization, 1 μL of the derivatized sample was injected into a cooled injection system (CIS4, Gerstel, GmbH & Co. KG, DE) port in splitless mode. The septum purge flow and purge flow to split vent at 2.5 min after injection were set to 25 and 15 mL min^{-1}, respectively. Initial temperature of the CIS port was 40 °C, and heated at 12 °C s^{-1} to 320 °C (after 30 s of equilibrium time), where it was kept for 5 min. After heating, the CIS port was gradually cooled to 250 °C at 5 °C s^{-1}, and this temperature was kept constant during the run. A GC–MS consisted of an Agilent 7890 B gas chromatograph (GC) and a high-throughput Pegasus GC-TOF-MS mass spectrometer (LECO Inc. USA). More details of GC oven and cooled injection system (CIS4) condition were the same as previously described [7]. Mass spectra were recorded in the *m*/*z* range of 45–600 with a scanning frequency of ten scans sec^{-1}, and the MS detector and ion source were switched off during the first 4.5 min of solvent delay time. The transfer line and ion source temperature were set to 280 °C and 250 °C, respectively. The mass spectrometer was tuned according to manufacturer's recommendation using perfluorotributylamine (PFTBA). The MPS and GC–MS was controlled using vendor software Maestro (Gerstel, GmbH & Co. KG, DE) and ChromaTOF (LECO Inc., USA). Samples were randomised prior to derivatization and GC–MS analysis, and a blank sample containing only derivatization reagent, and an alkane mixture standard (all even C10-C40 alkanes at 50 mg L^{-1} in hexane) were analysed at least between five real samples prior to monitor GC–MS performance.

2.3. Analysis of complex samples

The dataset investigated in this study consisted of 69 samples including blank samples and pooled quality control samples. The complex samples are media samples obtained from fermentation of CHO cells in complex media, the cells are removed by filtration and the spent media is kept on −20 °C until the time of derivatization. Prior to the analysis, the samples were derivatized using a procedure based on the protocol described by Smart et al. [8]. All samples were analysed in a randomised order. A 6890N GC in conjunction with a 5975 B quadrupole mass spectrometer (Agilent Technologies, USA) were used to analyse the samples. The system was controlled by ChemStation (Agilent Technologies, USA).

3. Theory

PARADISe is based on PARAFAC2 modelling, which allows simultaneous deconvolution of pure mass spectra of peaks and integration of areas of deconvoluted peaks for all samples. Resolved peaks are identified using their deconvoluted pure mass spectra and the final peak table is generated. Thus, PARADISe is based on five major steps: