



Gas chromatography – mass spectrometry data processing made easy

Johnsen, Lea G.; Skou, Peter Bæk; Khakimov, Bekzod; Bro, Rasmus

Published in:
Journal of Chromatography A

DOI:
[10.1016/j.chroma.2017.04.052](https://doi.org/10.1016/j.chroma.2017.04.052)

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license:
[CC BY-NC-ND](#)

Citation for published version (APA):
Johnsen, L. G., Skou, P. B., Khakimov, B., & Bro, R. (2017). Gas chromatography – mass spectrometry data processing made easy. *Journal of Chromatography A*, 1503, 57-64.
<https://doi.org/10.1016/j.chroma.2017.04.052>



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, Ålgårde, 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.

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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).

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 μ L of each dilution series samples were dried using ScanVac (Labogene, DK) at 40 °C inside 150 μ L glass inserts, sealed with air tight magnetic lids into GC–MS vials and derivatized by addition of 30 μ L 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 μ L 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:

1. Define intervals
2. Resolve compounds
3. Validate models
4. Identify compounds
5. Create peak table

PARADISE, integrates all these as outlined below.

Intervals are selected manually through an interactive TIC plot in such a way that approximate baseline-resolved intervals, with preferably less than six peaks, are obtained. As will be illustrated later, the specific definition of the intervals is not critical (within reason). Having defined each interval, the PARAFAC2 model can resolve the underlying and possibly overlapping compounds in each of these intervals. For each interval, a separate PARAFAC2 model is built. To do so, the number of chemical compounds (including baseline) must be defined for the specific PARAFAC2 model. PARADISE will by default calculate models with one to eight components, and it is the user that must decide which of the models to use. Automated methods exist for determining the number of components [5] but in PARADISE, the user has to do this. Normally, the number of components is set to the highest number that still maintains a sufficiently high core consistency (above 50%). Visualizations of the models can be used for intervals that may pose special problems to further guide the user but this is mostly not critical. Once the model for a given interval is determined, compounds of interest can be tagged (e.g. compounds that are not baseline or tails from peaks surrounding the interval) and only these compounds will be included in the final report. For a more thorough description of the theory behind PARAFAC2 the reader is referred to the supplementary material.

The PARAFAC2 model of each compound provides the relative concentration (peak area) directly and users can evaluate elution profiles of deconvoluted peaks. Identification is also a crucial part of the chromatographic analysis, and PARADISE enables the user to make library lookups of both mass spectra from raw data and PARAFAC2 deconvoluted mass spectra (pure compound spectra). The lookup is performed by exporting relevant spectra to the NIST MSsearch, which therefore must be installed prior to use the library lookup function. The user can then perform the evaluation of any library hits directly in the MSsearch software.

PARADISE is built around two main interfaces; one, which is used for inspection of raw data and creation of intervals, and one, which is used to visualize and validate models prior to select deconvoluted peaks and to create a final report. The software is compiled via Matlab and is thus platform independent and can work without NIST software. However, using the PARADISE without the NIST software eliminates the possibility of performing library searches of mass spectra. An overview of the full workflow is illustrated in Fig. 1.

Two formats of raw data can currently be imported; either data in the cdf format for mass spectrometry, or for users who are familiar with Matlab, data can be imported from the Matlab format.

4. Results

In the following we will illustrate the capabilities of PARADISE through a number of small examples, each aimed at different typical challenges encountered in chromatographic data analysis.

4.1. Quantification

Quantification is an important part of data analysis. To illustrate the capabilities of PARADISE concerning quantitative determination of compounds, a dilution series of the standard mixture sample were analysed. The obtained data was processed using ChromaTOF,

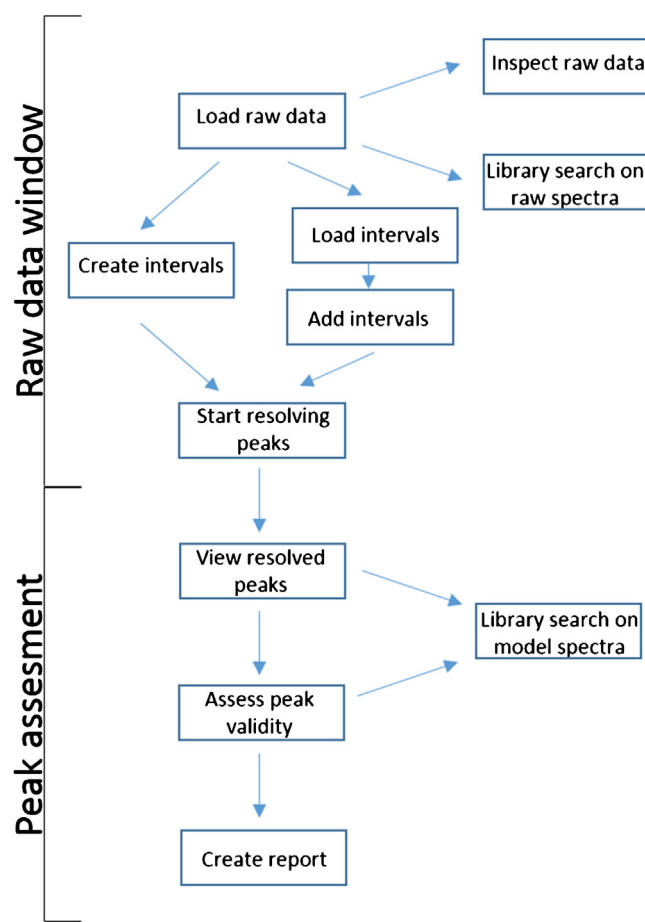


Fig. 1. Flowchart illustrating the workflow in PARADISE; from loading of raw data to generation of the final report with relative concentrations of detected compounds.

AMDIS and PARADISE (Fig. 2 and Fig. S2). All three software packages performed equally well when the S/N ratio of peaks was high. However, for the lower S/N ratio peaks, AMDIS and ChromaTOF results were sensitive to the settings of the user-defined parameters, while PARADISE performance was more consistent regardless of S/N ratio of peaks.

4.2. Co-elution

To demonstrate application of PARADISE to complex GC–MS profiles, a data set obtained from GC–MS analysis of spent media from cell cultures grown in complex media was investigated. One of the huge advantages of using PARADISE is its ability to deconvolute overlapping peaks. An example of the deconvolution power is illustrated in Fig. 3. The TIC of this data interval shows one peak, one baseline and one tail from a neighbouring peak. Upon inspection of the data using PARADISE, it becomes apparent that the interval is covering not one but three peaks and the interval is therefore best described with a five-component PARAFAC2 model: one component describing baseline, one the tail and one for each of the three peaks, respectively (see Fig. 3). Inspection of characteristic m/z ions (m/z 127, 216, and 130) of the deconvoluted peaks shows that the three peaks can be recognised from the corresponding extracted ion chromatograms (bottom plots in Fig. 3). It is worth to mention here that PARADISE allows such a deconvolution and provides pure spectra of deconvoluted peaks for even more complex chromatographic data intervals, without any user defined settings, besides the number of components.

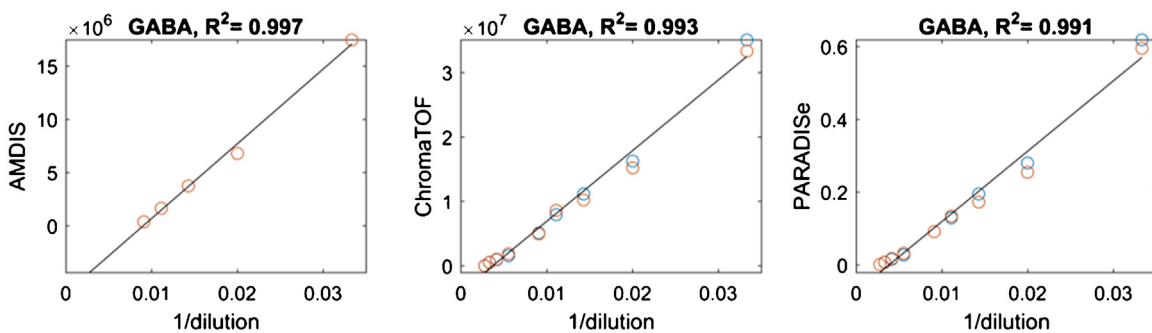


Fig. 2. Illustration of obtained relative concentrations from AMDIS, ChromaTOF and PARADISE from dilution series analysis of GABA (red and blue correspond to replicates). GABA was not detected by AMDIS in the most diluted samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

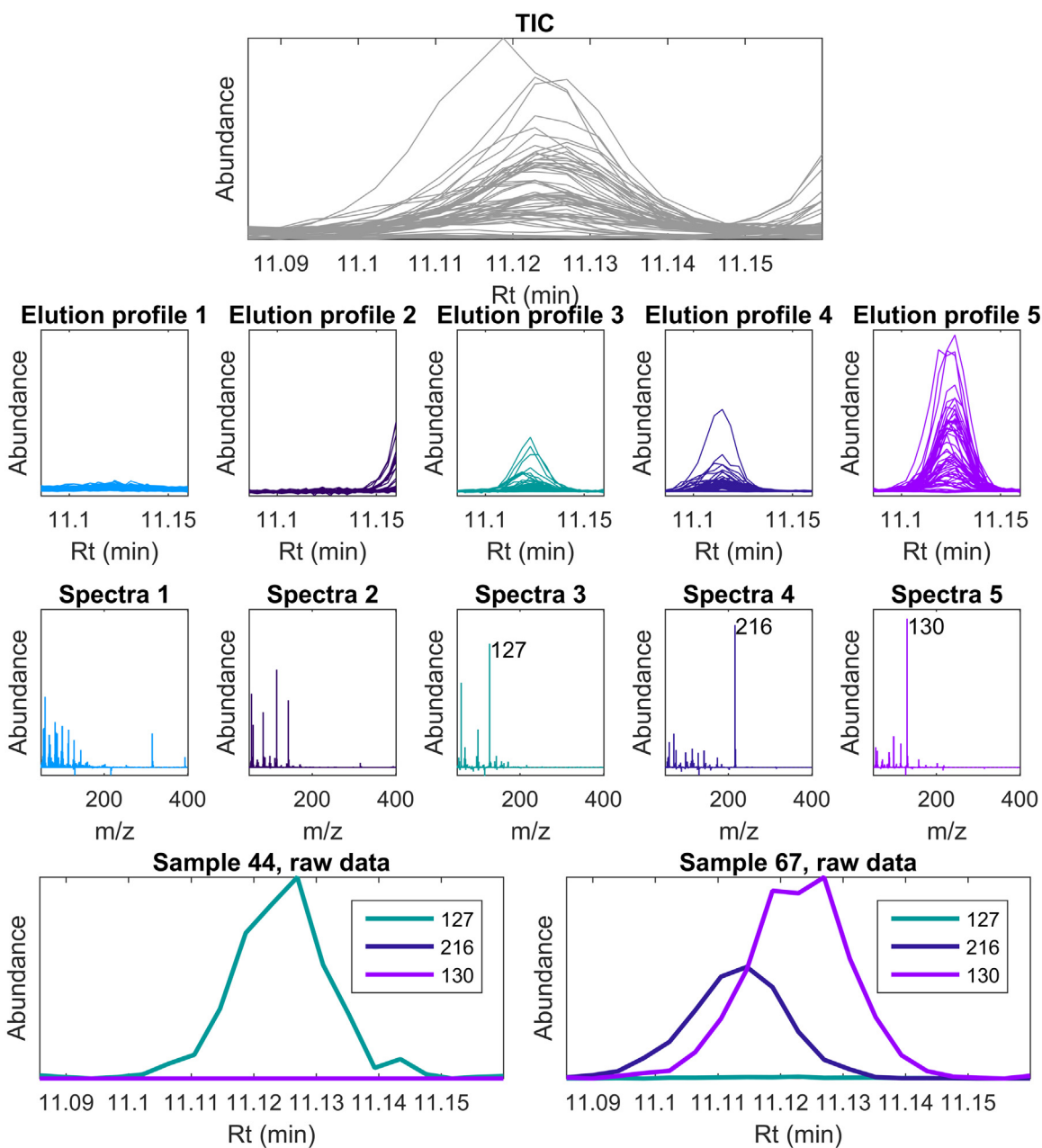

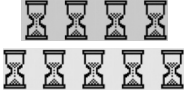
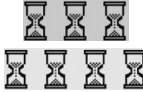







Fig. 3. Top: TIC of the interval, row 2: obtained elution profiles from a five-component model, row 3: model spectra obtained from the five-component model. Row 4: EIC of characteristic masses from the model (extracted from raw data).

Table 1

An overview of the data processing steps that require user defined parameters in three different GC–MS data processing software, AMDIS, ChromaTOF, and PARADISE. The number of hourglass indicates how many parameters must be set by the user in the given step of data processing, (–) indicates that this step is not performed by the software, and empty cells illustrate steps that do not require any parameters to be set by users for the given software.

Data processing steps that require parameters to be set by the user	Software		
	AMDIS	ChromaTOF	PARADISE
1) Define RT intervals for processing	–	–	
2) Deconvolution			
3) Peak filtering and removing baseline			
4) Mass spectrometer dependent parameters			
5) Processed data validation			
6) Alignment of peaks across samples	–		

4.3. Low signal-to-noise

In contrast to other approaches, PARADISE is not so sensitive to the S/N ratio of peaks and is able to deconvolute extremely small peaks directly from the raw data (Figs. 4 and 5). In Fig. 4, the PARADISE results reveal that the investigated noisy interval actually contains two overlapping peaks with a very low S/N ratio. Inspection of characteristic m/z values in the raw data confirms that, within the given interval, two compounds are eluting with different mass spectra. Subsequently, a four-component PARAFAC2 model deconvoluted two peaks corresponding to two chemicals plus two components reflecting the background.

The second example (Fig. 5) shows how well the mass spectra from a low S/N ratio peak is modelled using PARADISE. Despite extreme low S/N ratio of this peak, its deconvoluted mass spectrum allowed identification using the NIST mass spectral library, found as dimethyl malonic acid. The identity of this compound was validated with an authentic standard, which was found to have the same retention time and mass spectrum.

4.4. Baseline

Baseline contributions present in a raw GC–MS data heavily influence both peak identification and quantification, thus it is important that data processing techniques can remove baseline contributions. In the model illustrated in Fig. 4 two different baselines are present and shows that it is possible to automatically remove these artefacts using PARADISE. It is often seen that the baseline is modelled using more than one PARAFAC2 factor, because the background is often a mixture of several contributions (e.g. column bleed, derivatization reagent, mobile phase, or electronic noise) All models presented in this paper illustrate how the PARADISE approach removes baseline contributions as separate PARAFAC2 components from eluting compounds eliminating any need for raw data pre-treatment.

4.5. Retention time drift

In the examples illustrated throughout this paper, different degrees of shift in RT are present (see Figs. 3–5). In all cases, PARADISE handles the drift without any prior assumptions

about maximum allowed shift. PARADISE is also able to correctly determine peaks that have severe RT shifts across samples that sometimes result in complete cross RT shifts with nearby eluting peaks as well as with co-eluting peaks. This is only possible due to the unique mass spectrum of each compound and flexibility of deconvolution engine, PARAFAC2. However, in order to correctly determine all peaks present in a given chromatographic data interval, the width of the interval must be wide enough to cover RT shifts.

4.6. Limitations

There are two major cases when PARADISE fails to deconvolute GC–MS peaks: 1) when a GC–MS data interval contains two or more peaks with identical mass spectra, 2) when a GC–MS data interval contains two or more peaks that co-vary completely in their concentrations. In both cases PARADISE will find those co-varying peaks as a single compound. In the example illustrated in Fig. 6, two of the four peaks are lumped into one common component (Elution profile 4). This happens regardless of how many PARAFAC2 components are included in the model. Inspection of the raw data reveals that the two peaks have identical mass spectra (Top two rows, right in Fig. 6). It is a premise of PARAFAC2 that each chemical compound in a given interval must have at least slightly different spectral signature. Hence, when two compounds have identical spectra as here, they cannot be separated in a PARAFAC2 model. The only alternatives then are either 1) to split the data in between the two peaks, or 2) try to separate the peaks by other means (chemically or mathematically).

One can also choose either to exclude the compound from the final data set, or to use it, bearing in mind, that the reported concentration profile/spectra will be a combination of both peaks. Working within smaller retention time intervals minimizes the risk of modelling problems if different peaks co-vary across samples.

5. Discussion

PARADISE excels in simplicity because only little input is needed from the user to obtain valid models of the compounds and the inputs typically have a feasible range of settings so that the exact choice is not critical. The data must be split into retention time

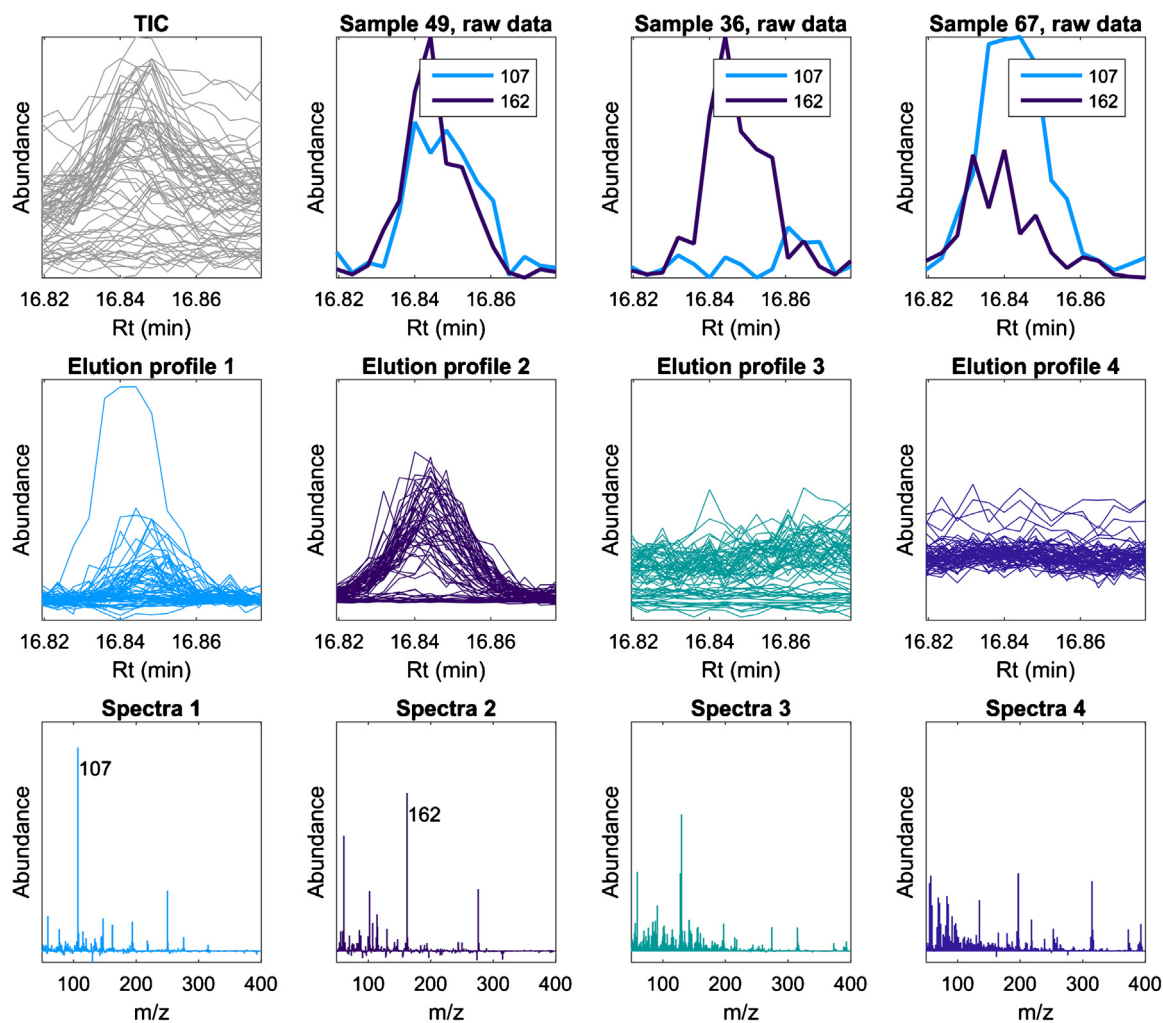


Fig. 4. Top left: TIC from raw data. Top right: EIC from raw data of selected m/z . Middle: elution profiles obtained from a four-component model. Elution profile 3 and 4 represent baseline. Bottom: spectra obtained from a four-component model.

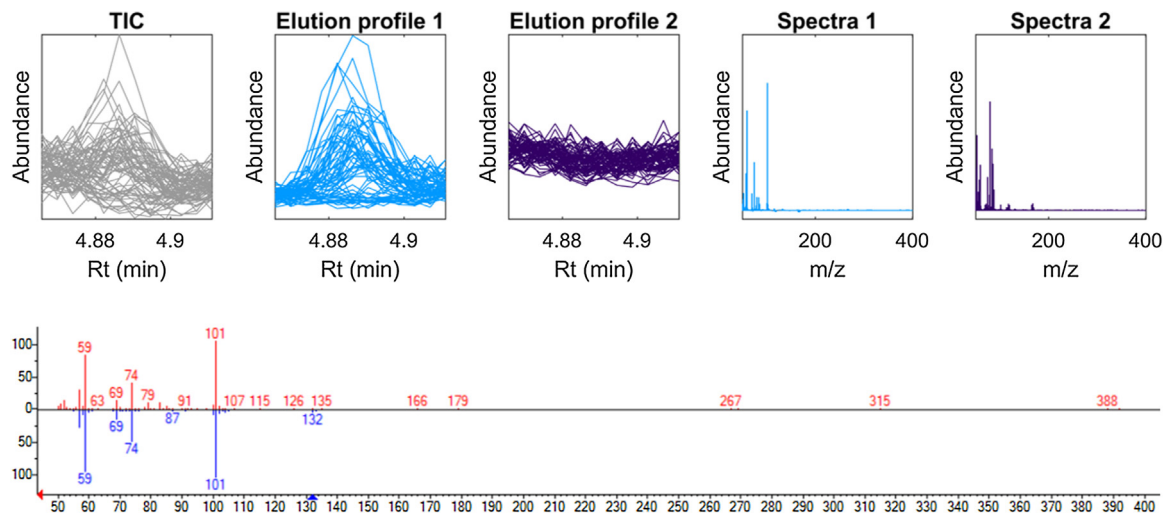


Fig. 5. Top: TIC from raw data, Elution profiles and spectra obtained from a two-component model. Bottom: comparison between the model spectra 1 and the NIST library spectra of dimethyl malonic acid. Profile 2 is representing baseline.

intervals with approximate baseline separation. The interval borders should be determined in a reasonable manner, meaning that the peaks of interest should be included in the interval without

cutting off any tailing or fronting. Even tails from peaks adjacent to the intervals, as shown in Fig. S1, does not pose a problem. Further, as few compounds as possible should be included when selecting

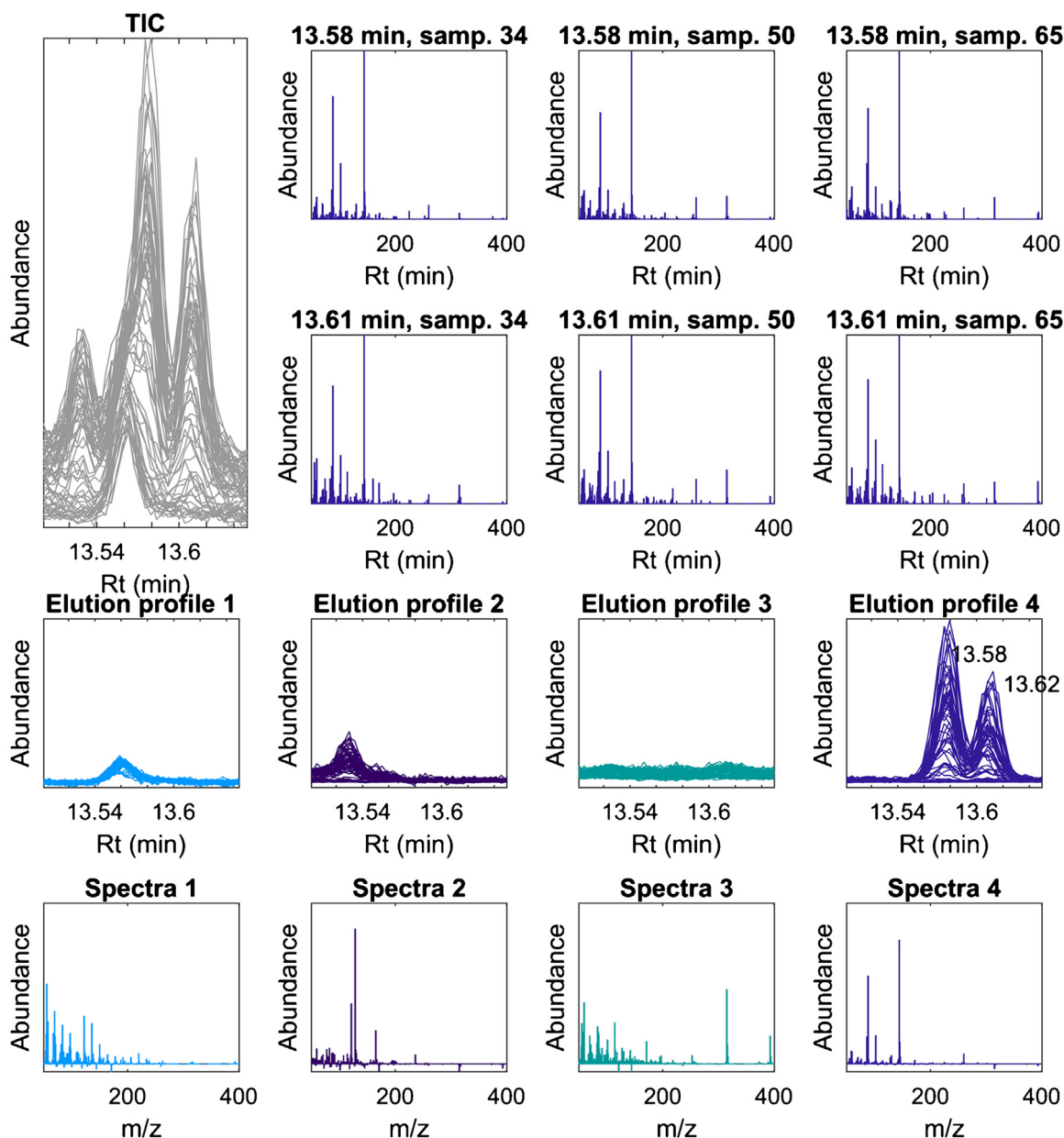


Fig. 6. Top left: TIC from raw data. Top 2 rows, right: spectra obtained from three different samples of the two peaks eluting at 13.58 and 13.62 min, respectively. Row 3: elution profiles obtained from a four-component model. Bottom row: spectra obtained from a four-component model.

intervals. Selecting a simpler (fewer compounds) interval reduces computation time and prevents small errors accumulating in more complicated models with many compounds.

Unlike some tools for processing of GC–MS data, the same model describes all samples when using PARADISE. This means that if a model is accepted as valid, all samples are well described in that particular interval and the developed method can routinely be applied to new samples without any user interaction.

An added benefit from using PARADISE is that there will not be any non-detects. In many methods, the user must specify parameters used to define a peak (e.g. peak width, signal to noise levels etc.). This means that if a peak does not match these criteria they will appear as “not detected”. In most cases this will be due to a peak being lower than the limit of detection. These missing values will cause problems if the data is to be used in either classical statistics [9] or multivariate statistics. In more severe cases, a peak may actually be present but not fulfilling the initially set parameters. If the

user does not recognize this, it will most likely be wrongly interpreted. In PARADISE there are no assumptions made about peak shape, signal-to-noise ratio or expected retention time shifts. When peaks are deconvoluted there will always be an estimate of the concentration (also in cases with signal being lower than the limit of detection), and the problems with missing values are therefore not an issue. In essence, the problem of non-detects is moved to the subsequent data analysis. All peaks are quantified and the possible decision of where to set the limit of detection can be decided after the quantification has been performed.

PARADISE cannot process one sample at a time but requires several samples prior to processing any dataset. It is not enough to analyse the same sample several times or to make dilutions of the samples and analyse these. If one wants to use PARADISE at least five samples with independent variations must be included in the sample set and preferably more.

To be able to compare the user-friendliness of the software AMDIS, ChromaTOF and PARADISE, we divide the workflow into five parts below for easier comparison:

- 1) Define RT intervals for processing. Division of the chromatographic data into smaller RT intervals is needed for reducing complexity when processing data using PARADISE prior to obtain reliable deconvolution.
- 2) Deconvolution. The deconvolution step in AMDIS and ChromaTOF requires parameters such as peak width, resolution, sensitivity, and shape to be set by users. The number of components must be determined in PARADISE.
- 3) Peak filtering and removing baseline. The peak filtering step requires parameters like S/N ratio, mass threshold, baseline offset, minimum abundance in AMDIS and ChromTOF.
- 4) Mass spectrometer dependent parameters. Mass spectrometer dependent parameters such as m/z range, scan direction, instrument type, file format, threshold are also crucial when using AMDIS.
- 5) Alignment of peaks across samples. Several parameters such as maximum allowed RT shift, spectral similarity, detection frequency (e.g., a peak must be present at least in 50% of samples) are required in ChromaTOF when aligning peaks across samples prior to a final metabolite table.

In Table 1 a summary is given, indicating how many parameters the user needs to set for each step.

PARADISE can be used for targeted analysis, where only the target compounds are processed, as well as untargeted analysis. In cases with routine targeted high-throughput GC–MS methods, interval-files can be predefined and reused. However, it is important to stress that the user should still inspect the raw data before processing the data. This is, in fact, underestimated in many data processing software packages, but we strongly advice data inspection prior to use PARADISE.

6. Conclusions

We have demonstrated a new approach called PARAllel factor analysis 2 based Deconvolution and Identification System (PARADISE), integrating multi-way modelling for processing of raw GC–MS data from several samples simultaneously. PARADISE combines entire workflow from raw data inspection to peak deconvolution and metabolite identification in a graphical user interface. It allows handling very complex situations with severe co-elution even with resolution close to zero. With PARADISE, a single standalone platform is presented covering the entire workflow from

inspecting raw data to identification, including deconvolution of peaks across all samples simultaneously, determination of relative concentrations and compilation of a compound table. The ability to export mass spectra, deconvoluted (pure) as well as raw, to spectral databases can save large amounts of time and will increase the hit quality.

Acknowledgements

We would like to thank Dr. W. Gary Mallard from the National Institute of Standards and Technology (NIST) from sharing insights into how to optimally use the AMDIS software. It was a great help and we are very thankful for your time.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.chroma.2017.04.052>.

References

- [1] M.L. Phillips, R.L. White, Dependence of chromatogram peak areas obtained by curve-fitting on the choice of peak shape function, *J. Chromatogr. Sci.* 35 (1997) 75–81.
- [2] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, *Chem. Rev.* 110 (2010) 4582–4605, <http://dx.doi.org/10.1021/cr900394n>.
- [3] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, *J. Am. Soc. Mass Spectrom.* 10 (1999) 770–781, [http://dx.doi.org/10.1016/S1044-0305\(99\)00047-1](http://dx.doi.org/10.1016/S1044-0305(99)00047-1).
- [4] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—part II. Modeling chromatographic data with retention time shifts, *J. Chemom.* 13 (1999) 295–309, [http://dx.doi.org/10.1002/\(SICI\)1099-128X\(199905/08\)13:3/4<295::AID-CEM547>3.0.CO;2-Y](http://dx.doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y).
- [5] L.G. Johnsen, J.M. Amigo, T. Skov, R. Bro, Automated resolution of overlapping peaks in chromatographic data, *J. Chemom.* 28 (2014) 71–82, <http://dx.doi.org/10.1002/cem.2575>.
- [6] T. Skov, R. Bro, Solving fundamental problems in chromatographic analysis, *Anal. Bioanal. Chem.* 390 (2007) 281–285, <http://dx.doi.org/10.1007/s00216-007-1618-z>.
- [7] B. Khakimov, R.J. Mongi, K.M. Sørensen, B.K. Ndabikunze, B.E. Chove, S.B. Engelsen, A comprehensive and comparative GC–MS metabolomics study of non-volatiles in Tanzanian grown mango, pineapple, jackfruit, baobab and tamarind fruits, *Food Chem.* 213 (2016) 691–699, <http://dx.doi.org/10.1016/j.foodchem.2016.07.005>.
- [8] K.F. Smart, R.B.M. Aggio, J.R. Van Houtte, S.G. Villas-Bôas, Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography–mass spectrometry, *Nat. Protoc.* 5 (2010) 1709–1729, <http://dx.doi.org/10.1038/nprot.2010.108>.
- [9] D.R. Helsel, Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it, *Chemosphere* 65 (2006) 2434–2439.