UNIVERSITAT DE BARCELONA

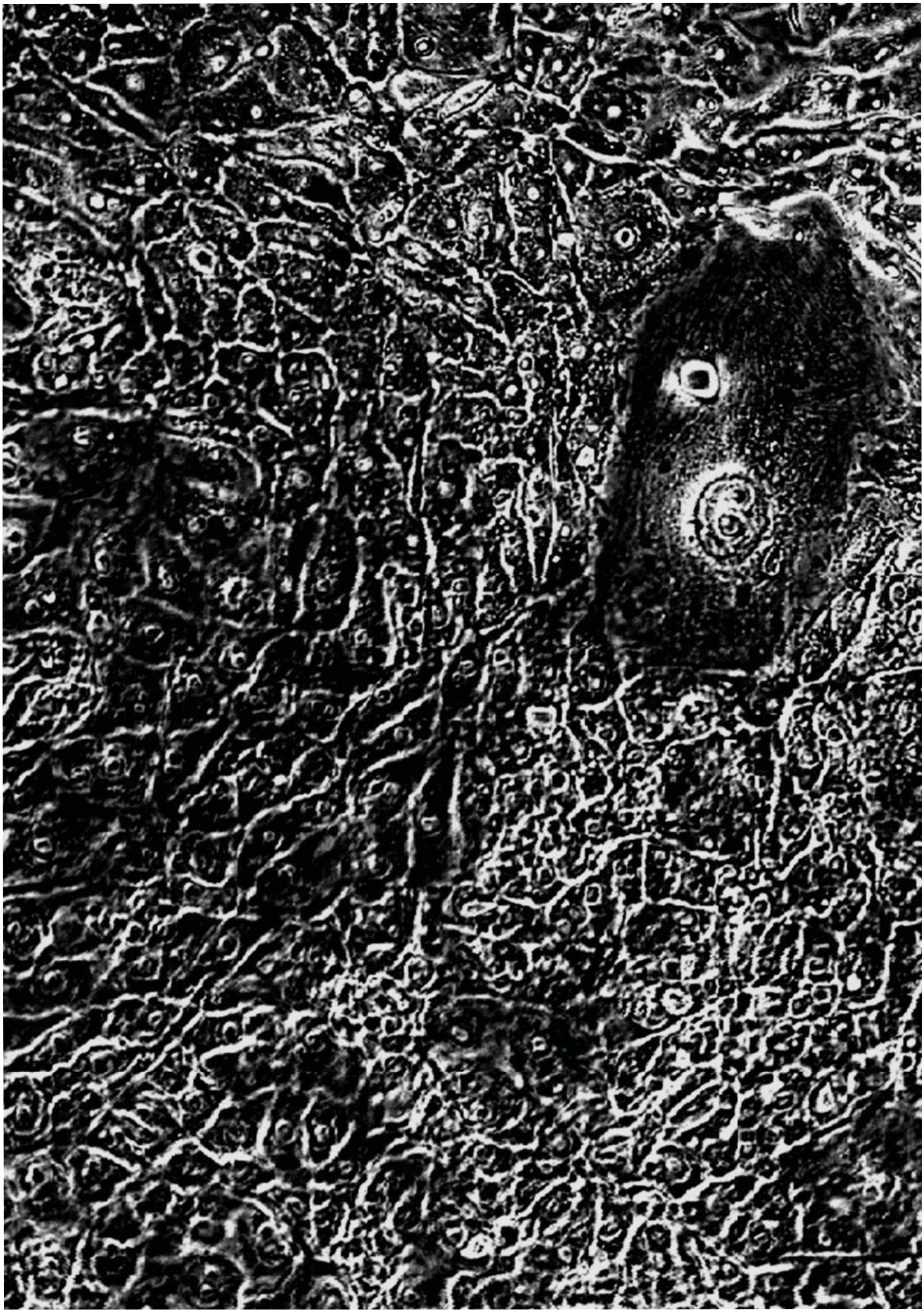
Assessment of the lipidomic effects of environmental pollutants on exposed organisms using chemometric and analytical methods

Eva Gorrochategui Matas

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- CompartIgual 3.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - CompartirIgual 3.0. España de Creative* Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-ShareAlike 3.0. Spain License***.



CHAPTER 2

Novel data analysis approaches for metabolomics/lipidomics

CHAPTER 2

As previously mentioned in the Introduction Section, the analysis of metabolomic (and lipidomic) LC-MS data sets is one of the bottlenecks of omic studies due to the complexity of the nature of the data sets to be analysed. Several data analysis packages and software have been developed in the last years for that purpose. However, the automatic characteristics of these data analysis tools are not always adequate for a proper understanding of the nature of the LC-MS metabolomic (and lipidomic) data sets and of the procedure required to extract meaningful information from them. Moreover, most of these approaches present some drawbacks (*e.g.*, requirement of chromatographic alignment and peak shape modelling). Therefore, one of the main objectives of this Thesis was the development of a data analysis strategy that allowed researchers to analyse their own data without the requirement of any external data analysis package, which also properly addressed the issues found in the already existing data analysis packages.

This Chapter of the Thesis is structured in the following manner: an introduction section, a scientific research section including the scientific articles related with *omic data analysis procedures*, a discussion section and some specific conclusions.

2.1. INTRODUCTION

2.1.1. SCIENTIFIC ARTICLE I

Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow

E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler

Trends in Analytical Chemistry (2016) **82**, 425-442

Supplementary liveslides at: http://audioslides.elsevier.com/ViewerSmall.aspx?doi=10.1016/j.trac.2016.07.004&Source=0&resumeTime=0&resumeSlideIndex=1&width=800&height=639

Trends in Analytical Chemistry 82 (2016) 425–442

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow

Eva Gorrochategui, Joaquim Jaumot, Sílvia Lacorte *, Romà Tauler **

Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Catalonia 08034, Spain

ARTICLE INFO

Keywords:
Metabolomics
Data analysis
Mass spectrometry
Liquid chromatography
Target
Untarget
Chemometric tools

ABSTRACT

Data analysis is a very challenging task in LC-MS metabolomic studies. The use of powerful analytical techniques (e.g., high-resolution mass spectrometry) provides high-dimensional data, often with noisy and collinear structures. Such amount of information-rich mass spectrometry data requires extensive processing in order to handle metabolomic data sets appropriately and to further assess sample classification/discrimination and biomarker discovery.

This review shows the steps involved in the data analysis workflow for both targeted and untargeted metabolomic studies. Especial attention is focused on the distinct methodologies that have been developed in the last decade for the untargeted case. Furthermore, some powerful and recent alternatives based on the use of chemometric tools will also be discussed. In general terms, this review helps researchers to critically explore the distinct alternatives for LC-MS metabolomic data analysis to better choose the most appropriate for their case study.

© 2016 Elsevier B.V. All rights reserved.

Contents

Abbreviations: ABF, Analysis services backup file; ASCA, ANOVA-simultaneous component analysis; CART, Classification and regression trees; CAWG, Chemical analysis working group; CCSWA, Common components and specific weights analysis; CE-MS, Capillary electrophoresis-mass spectrometry; CMTF, Coupled matrix and tensor factorization; CWT, Continuous wavelet transform; DISCO-SCA, Distinctive and common components with simultaneous-component analysis; DNA, Deoxyribonucleic acid; DTW, Dynamic time warping; FT-ICR, Fourier transform ion cyclotron resonance; GC, Gas chromatography; GC-MS, Gas chromatography coupled to mass spectrometry; GC-MS/MS, Gas chromatography tandem mass spectrometry; GSVD, Generalized singular value decomposition; HMDB, Human metabolome database; ^{1}H-NMR, Proton nuclear magnetic resonance; HPLC, High-performance liquid chromatography; HRMS, High-resolution mass spectrometry; HRMS/MS, High-resolution tandem mass spectrometry; ICA, Independent component analysis; IPA, Ingenuity pathway analysis; IS, Internal standard; IT, Ion trap; JIVE, Joint and individual variation explained; KEGG, Kyoto encyclopedia of genes and genomes; LC-MS, Liquid chromatography coupled to mass spectrometry; LC-QTOF-MS, Liquid chromatography coupled to quadrupole time-of-flight mass spectrometry; LLR, Linear logistic regression; LOESS, Locally estimated scatter plot smoothing; LRMS/MS, Low-resolution tandem mass spectrometry; MCR-ALS, Multivariate curve resolution-alternating least squares; MFICA, Mean-field independent component analysis; MMSAT, Metabolite mass spectrometry analysis tool; MS, Mass spectrometry; MS^{E}, Mass spectrometry$^{\text{Elevated energy}}$; MSI, Mass standards initiative; m/z, Mass-to-charge; NAC, N-acetylcysteine; NMR, Nuclear magnetic resonance; NOMIS, Normalization using optimal selection of multiple internal standards; OBI-warp, Ordered bijective interpolated warping; OPLS, Orthogonal projections to latent structures; O2PLS, Two-way orthogonal projections to latent structures; OnPLS, Multiblock orthogonal projections to latent structures; PARAFAC, Parallel factor analysis; PARAFAC2, Parallel factor analysis2; PBL, Peripheral blood lymphocytes; PCA, Principal component analysis; PCDA, Principal component discriminant analysis; PLS, Partial least squares; PLS-DA, Partial least squares-discriminant analysis; PPP, Pentose phosphate pathway; PQN, Probabilistic quotient normalization; QCs, Quality control sample; QLIT, Quadrupole linear ion trap; QqQ, Triple quadrupole; Q-TOF, Hybrid quadrupole orthogonal time-of-flight; RANSAC, Random sample consensus; RNA, Ribonucleic acid; ROI, Region of interest; SIM, Selected ion monitoring; SLE, Systemic lupus erythematosus; SNR_{Thr}, Signal-to-noise ratio threshold; SR, Selectivity ratio; SRM, Selected reaction monitoring; TLD, Trilinear decomposition; TOF, Time-of-flight; TPP, Trans-proteomic pipeline; UHPLC, Ultra high-performance liquid chromatography; UPLC-TOF, Ultra performance liquid chromatography coupled to time-of-flight mass spectrometry; VAST, Variable stability scaling; VIP, Variable importance on projection; XCMS, Various forms (X) of chromatography mass spectrometry.

* Corresponding author. Tel.: +34 934006133; fax: +34932045904.
E-mail address: slbqam@cid.csic.es (S. Lacorte).
** Corresponding author. Tel.: +34 934006140; fax: +34932045904.
E-mail address: roma.tauler@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.trac.2016.07.004
0165-9936/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Metabolomics [1–3] is one of the categorical platforms that constitute omics [4] (see Fig. 1). Omics is a field that aims at the study of the abundance and (or) structural characterization of a broad range of molecules in organisms under distinct scenarios. In the clinical field, high-throughput omic technologies are used for the characterization of diseases to better predict the clinical course of organisms and to evaluate the efficacy of existing or under-development therapies [5]. In food science, omics plays a significant role in the light of an improvement of human nutrition [6]. In the environmental field, omic studies aim at the evaluation of the alterations that organisms might suffer after exposure to environmental stressors [7,8].

In all cases, the expressed molecules are involved in most crucial biological processes, and principally comprehend deoxyribonucleic acid (DNA) (genomics [9], epigenomics [10]), ribonucleic acid (RNA) (transcriptomics [11]), proteins (proteomics [12]), and other small molecules (metabolomics [1–3]). In more recent years, another categorical omic platform named fluxomics [13,14], which aims at the study of the fluxome, or the total set of fluxes in the metabolic network of the biological specimen, has gained relevance. Apart from these categorical omic platforms, a variety of omic subdisciplines

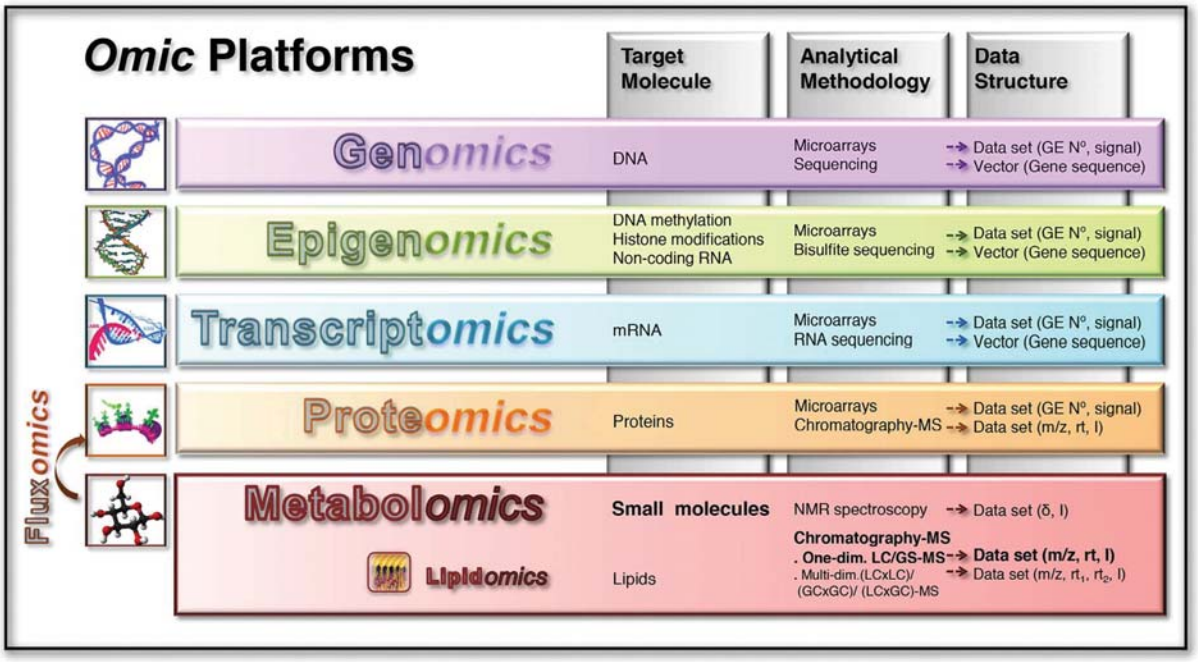


Fig. 1. Overview of OMIC platforms: target molecules, analytical methodologies used and structure of the generated data (GE N°: number of genes, δ: chemical shift, m/z: mass-to-charge ratio, rt: retention time, I: intensity). *Data structure shown when considering only one sample.*

have also emerged (e.g., lipidomics [15], glycomics [16], foodomics [6,17], interactomics [18], and metallomics [19]), showing that omics is a constantly evolving discipline. Among all these omic platforms, metabolomics is becoming increasingly popular and is used to detect the perturbations that disease, drugs or toxins might cause on concentrations and fluxes of metabolites involved in key biochemical pathways [20]. Due to its importance and relevance, the current study concentrates on metabolomic data.

Several analytical techniques have been developed for each of the omic platforms (see Fig. 1), including DNA microarray-based and RNA-sequencing techniques [21], nuclear magnetic resonance (NMR) spectroscopy [22,23] and mass spectrometry (MS) methods [24,25]. In the field of metabolomics, both NMR and MS techniques are the most popular. High-resolution proton NMR spectroscopy (^{1}H-NMR) has proved to be one of the most powerful technologies for examining biofluids and studying intact tissues, producing a comprehensive profile of metabolite signals without separation, derivatization, and preselected measurement parameters [26,27]. On the other hand, MS methods, both by direct injection [28] or coupled to chromatographic techniques [29], have also evolved into a powerful technology for metabolomics due to their ability in the analysis of low molecular weight compounds in biological systems. These two approaches (i.e., NMR and MS) are complementary, and the integration of both technologies to provide more comprehensive information is now pursued in the metabolomics field. Nevertheless, this study concentrates on MS-based metabolomic data.

Concerning MS instrumentation, high-resolution mass spectrometers are the most powerful analysers due to their ability to improve accurate mass determination. In fact, spectrometers such as time-of-flight (TOF) [30], quadrupole time-of-flight (Q-TOF) [31], and Fourier transform ion cyclotron resonance (FT-ICR) [32] spectrometers and orbital ion traps [33], have substituted in many cases the conventional low-resolution quadrupoles and linear ion traps (IT), due to their ability to resolve isomeric and isobaric species and elucidate elemental composition [34]. Regarding chromatographic techniques, early metabolomic studies were commonly based on gas chromatography (GC), since it is a highly efficient, sensitive and reproducible technique [35]. However, GC has the drawback that only volatile compounds or compounds that are made volatile after derivatisation can be analysed, and extensive sample preparation is often required. In contrast, high-performance liquid chromatography (HPLC) and ultra high-performance liquid chromatography (UHPLC) are considered to be more comprehensive than GC since they allow the analysis of a wider range of metabolites without the requirement of derivatisation [36–39]. Hence, liquid chromatography coupled to mass spectrometry (LC-MS) has lately gained popularity in the metabolomics field in detriment of gas chromatography coupled to mass spectrometry (GC-MS), this being the reason why this study is focused on the former technique.

The improvement of analytical techniques has gradually caused metabolomic data sets to become larger with more intricate inner structures [40]. Mass spectrometric based techniques generate highly complex data, due to the vast number of measurements (i.e., MS spectrum at each retention time) related to the number of observations (i.e., samples). In the case of LC-MS analysis (see Fig. 1), data generated from each chromatogram are arranged in data sets containing information of mass-to-charge (m/z), retention times and intensities. Hence, massive amounts of information-rich MS data are generated in the analysis of every sample, thus requiring specific standard approaches for its study and interpretation [41].

In general terms, data analysis strategies are classified in two groups: data analysis strategies for targeted (Fig. 2) and untargeted (Fig. 3) metabolomic studies. The reason for such differentiation is due to the different types of data generated in these two approaches, which require being handled accordingly. Targeted studies [42] focus the research on a set of known metabolites whereas untargeted studies [43] allow a more comprehensive evaluation of metabolomic profiles. Most of the methodologies used in early targeted studies just allowed the identification of a few number of metabolites [44]. Nevertheless, recent targeted methodologies enable large-scale metabolic profiling, including hundreds of compounds [45–47]. However, the number of compounds analysed in untargeted studies is even larger. This is so because one must process entire data sets including thousands of metabolite signals, and among these, few are finally identified as candidate biomarkers [48]. Therefore, data analysis strategies for untargeted studies require highly-extensive processing of LC-MS chromatograms. A large number of data analysis strategies are found in the literature but none of them can be singled out as the optimal choice in all cases, which makes data analysis an open task in the bioinformatics research. In fact, the field of MS-based metabolomics is rather young, and new methods, software and platforms are being regularly published or updated [49,50].

A recent review of Yi et al. [51] summarizes recent and potential advances in chemometric methods in relation to data processing in untargeted metabolomic studies. Various aspects, including raw data pre-processing, metabolite identification, and variable selection and modeling are accurately discussed and presented there. The present review complements the previous one with some data analysis steps not covered or partially covered by the former (e.g., data acquisition, data storage and conversion, data import, data compression and feature detection or peak resolution), presents novel and little known chemometric tools for data analysis and includes an overview of the data analysis strategies for targeted studies. Moreover, it is intended to contribute to the state-of-art by providing comprehensive information on bioanalytical and data processing tools rather than describing the principles of the chemometric methods that can be used in LC-MS metabolomic data analysis.

2. General overview of the data analysis approaches

LC-MS metabolomic data analysis strategies are primarily designed for targeted and untargeted studies. However, future advances in LC-MS metabolomics may lead to a merging of targeted and untargeted analyses; with the targeted approach providing more sensitive and accurate detection of predetermined metabolites, and the untargeted approach being able to detect and identify unknown metabolites [52]. Indeed, first steps in this direction were made by Savolainen et al. [53], who collected for the first time targeted and untargeted metabolomic data from human plasma using gas chromatography coupled to tandem mass spectrometry (GC-MS/MS). Next, a brief introduction to both approaches is presented.

Data analysis in targeted metabolomics [42] aims to process data sets coming from a subset of the metabolome: a predefined group of chemically characterized and biochemically annotated metabolites contained in referential databases. The advantages of performing a targeted search are mainly attributed to two factors: first, analytical artifacts are not carried through to downstream analysis, and second, just a selected group of metabolites is studied. Even though this fact facilitates data analysis, the process becomes quite time-consuming and tedious if one wishes to study a large number of metabolites. In those cases, in order to reduce the effort and time required for the data analysis, some alternative automated methodologies have been developed [54–59] (see Section 3.1.5.).

The untargeted approach [43] attempts the comprehensive analysis of all measurable analytes in a sample, including uncharacterized metabolites. No previous knowledge of the sample is required, and

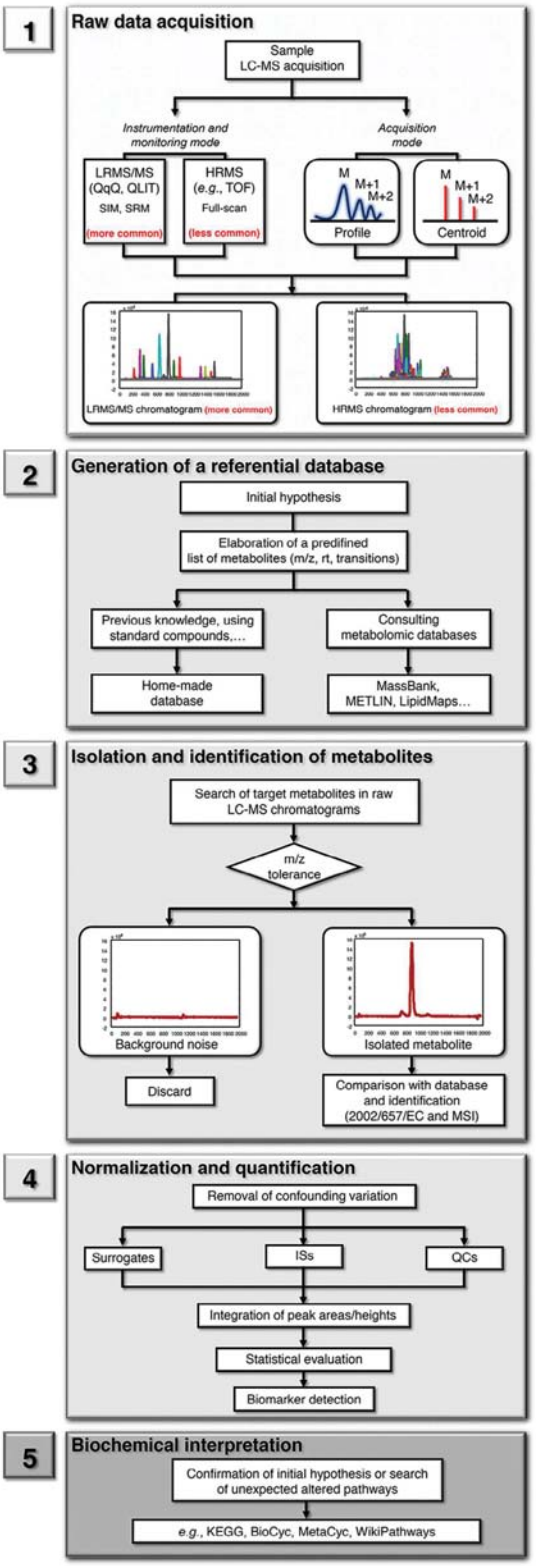


Fig. 2. Overview flowchart listing the five steps (grey shaded areas) involved in the data analysis approach for targeted studies: raw data acquisition, generation of a referential database, isolation and identification of metabolites, normalization and quantification, and biochemical interpretation. These steps are grouped in three major areas: data aquisition (light-grey), data processing and feature detection (medium-grey) and interpretation (dark-grey). In this figure rectangles indicate processing steps, diamonds indicate key contributional choices and in rounded rectangles are included illustrative representations of MS data and LC-MS chromatograms. Note that this flowchart does not consider the possibility of using automated data analysis tools such as MRMPROBS, MMSAT or OpenChrom, which have their own specific workflow (see Section 3.1.5.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

no referential database is necessary. However, its comprehensive nature requires the analysis of whole data sets, which include gigabytes of information. This is not possible without a previous reduction of their dimensions into more computationally manageable formats, but this compression must be carried out without significantly compromising the experimental information contained within. Moreover, the compressed data need further and extended analysis in order to finally detect most discriminant metabolites (i.e., potential biomarkers).

In Figs. 2 and 3 is shown a detailed scheme of the steps involved in data analysis strategies for targeted and untargeted studies, respectively. As shown in the former, the targeted approach can be broken down into five different parts (grey shaded areas): raw data acquisition, generation of a referential database, isolation and identification of metabolites, normalization and quantification, and biochemical interpretation. These parts can be grouped in three major areas: data acquisition (light-grey), data processing and feature detection (medium-grey) and interpretation (dark-grey). On the other hand, in Fig. 3 the untargeted approach is divided in nine parts, regrouped using the same criterion as in Fig. 2: raw data acquisition (light-grey area), data storage and conversion, import, compression, normalization, scaling and transformation, feature detection or peak resolution, biomarker screening and identification (medium-grey area) and biochemical interpretation (dark-grey area). Note that some steps are common in the targeted and the untargeted schemes. See Section 3 for a detailed explanation of both approaches.

3. The data analysis workflow for targeted and untargeted metabolomic studies

This section provides details of the steps involved in data analysis workflows for targeted and untargeted studies (highlighting common aspects), and finishes with a common explanation of the biochemical interpretation for both approaches.

3.1. Data processing steps for targeted studies

3.1.1. Raw data acquisition

Targeted analyses require collecting metabolite specific information typically using low-resolution tandem mass spectrometry (LRMS/MS) instrumentation such as triple quadrupole (QqQ) and quadrupole/linear ion trap (QLIT), which allow proper quantification. Both QqQ and QLIT are routinely operated via selected ion monitoring (SIM) and selected reaction monitoring (SRM). In addition, QLIT permits advanced MS^3 functionality together with QqQ fragmentation patterns, thus, providing more useful information needed for structural knowledge [52]. Although the use of LRMS/MS instrumentation is the most popular practice in targeted metabolomics, high-resolution mass spectrometry (HRMS) [60,61] can also be used in targeted analyses, operating in full-scan.

Acquisition mode of LC-MS data (i.e., centroid or profile, Figs. 2 and 3) is influential on the final identification of metabolites. Acquisition in centroid mode was introduced in the early days of MS

instrument development, when the amount of data and the data collection rate overwhelmed the state-of-art data system and data storage [62]. Consequently, early mass spectrometers (e.g., low-resolution quadrupoles and IT) were designed to reduce the acquired raw MS data to a stick spectrum, or centroid data, in a process known as *centroiding*. Centroiding processes each mass spectrum and combines multiple data points representing the same peak into a single data point with one *m/z* and intensity value. Nowadays, acquisition in centroid mode is no longer mandatory since data communication rate and storage capacity are not obstacles in most data systems anymore. In fact, acquisition in profile mode occurs by default in many HRMS instrumentation.

Centroiding has the obvious advantage of generating lighter data files (up to 100-fold smaller). However, centroid data are obtained at the expense of significant information loss, including noise characteristics, linearity of the ion signal, mass spectrally interfering ions and isotope fine features that can be obtained with HRMS when acquiring in profile or continuum mode. Such information is highly desirable since it facilitates the differentiation of formula candidates hard to distinguish [62].

For instance, a feature identification software named *MassWorks* (*Cerno Bioscience,* http://www.cernobioscience.com) takes advantage of the information gained under profile mode to reduce the number of possible formula candidates and achieve better results in the identification step [63,64].

3.1.2. Generation of a referential database

As previously stated, targeted metabolomics aims to search for a specified list of metabolites, typically focusing on one or more related pathways of interest [65]. In order to search for the metabolites of interest, the first step required is the elaboration of a referential database containing information of their nominal and exact mass, chemical formula, retention time and precursor and product m/z values. As observed in Fig. 2, such referential database can be constructed in two ways. One would be to take benefit from previous biochemical knowledge or from previous studies performed on the same type of organisms or groups of compounds, with the help of standard compounds (home-made database). The other approach consists of consulting retrospectively online metabolomic databases [e.g., human metabolome database (HMDB), METLIN, MassBank, LipidMaps & LipidBlast, NIST and mzCloud]. The readers interested in mass spectral databases for LC-MS metabolomic data sets are advised to consult the recent work of Vinaixa et al. [66].

3.1.3. Isolation and identification of metabolites

Following the generation of a referential database, next step is the isolation and identification of the target metabolites. Most targeted metabolomic studies use LC-MS vendor software [e.g., *Masslynx* (Waters), *Xcalibur* (Thermo Fischer), *Analyst* (AB Sciex), *Compass* (Bruker), *MassHunter* and *Chemstation* (Agilent)] for both isolation and identification of compounds, with the support of the referential database. Only in few cases, data are analysed out of the vendor software (see Section 3.1.5.).

Identification of metabolites is still evolving within the metabolomics community, with active discussion on how to define which features constitute valid metabolite identification [67]. Discussing all the identification strategies is out of the scope of this review, and only basic guidance is given. According to the criteria proposed by the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI: http://msi-workgroups.sourceforge.net), four levels of identification can be defined [68]. Level 1 refers to definitive identification, possible when having, at least, two orthogonal molecular properties of the putative metabolite confirmed with an authentic chemical standard analysed under identical analytical methodology (not necessarily in the researcher's laboratory). Levels 2 and 3 refer to putative or tentative identification so that comparison against literature and data sets is sufficient. Putative identification can provide metabolite-specific (level 2) or class-specific (level 3) identification. Level 4 refers to unknown compounds. Moreover, in the European Directive 2002/657/EC, the criteria for unequivocal identification of compounds according to the analytical platform used are presented [69].

As explained in Section 3.1.1., in targeted studies, two platforms can be used to enable proper identification of metabolites: LRMS/MS, which is the most common approach, and HRMS. When working with LRMS/MS, the standard procedures are SIM and SRM [70], as they enable high sensitivity, reproducibility and a broad dynamic range. Significant advances have been made to perform SRM experiments, and routine methods are now available for analysing most of the metabolites in central carbon metabolism, as well as amino acids and nucleotides at their naturally occurring physiological concentrations [71–73]. Moreover, most of the currently existing LRMS/MS targeted methods have been developed to enable large-scale metabolic profiling, including hundreds of compounds. Sawada et al. [45], optimized the SRM conditions of 497 plant metabolites and finally quantified 100 of them in each of 14 plant accessions from *Brassicaceae*, *Gramineae* and *Fabaceae*. Also, Gu et al. [47], optimized 595 precursor ions and 1890 SRM transitions for the analysis of serum metabolites. In most cases, the utilmate objective of these LRMS/MS methods is the screening of targeted lists of metabolites as potential metabolic signatures for diseases. Indeed, targeted screening on human plasma was used to reveal citric acid metabolites and a small group of essential amino acids as metabolic signatures of myocardial ischaemia and diabetes, respectively [74,75]. The little percentage of studies that use HRMS instrumentation operating in full-scan mode for targeted metabolomics utilize the mass deviation as the principal criteria for formula identification. In those cases, a deviation of 5 ppm is generally established as the admissible mass error [76–78]. Garanto et al. [60] characterized the mouse retinal sphingolipidome by ultra performance liquid chromatography coupled to time-of-flight mass spectrometry (UPLC-TOF), operating in full-scan mode, in a targeted lipidomic study. In that study, quantification was carried out using the ion chromatogram obtained for each compound using 50 mDa windows and positive identification of compounds was based on the accurate mass measurement with an error <5 ppm and its LC retention time, compared to that of standards.

Regardless the instrumentation used for targeted metabolomics (i.e., LRMS/MS or HRMS), identification of metabolites can be enhanced when acquiring data in profile mode, as explained in Section 3.1.1. For instance, Erve et al. [63] and Amorisco et al. [64] used the advantages of acquiring in profile mode to ensure precise identification of compounds.

3.1.4. Data normalization and quantification

The aim of normalization is to remove confounding variations attributed to experimental sources (e.g. analytical noise or experimental bias) in ion intensities among measurements while preserving the relevant variation (due to biological source). Chemical heterogeneity of metabolites, leading, for example, to distinct recoveries during extraction or responses during ionization in the mass spectrometer, makes separation between interesting biological variation and unwanted systematic bias a necessary labor [79]. In order to minimize undesired variations, some considerations must be taken, which are discussed below.

First, sample analysis for a particular study should be conducted in a randomized sample order, and the data should be acquired in the same batch on the same day, minimizing internal variation within a particular study set. Second, single or multiple surrogates (added to sample prior to extraction), internal standards (IS) (added to sample after extraction), and quality control

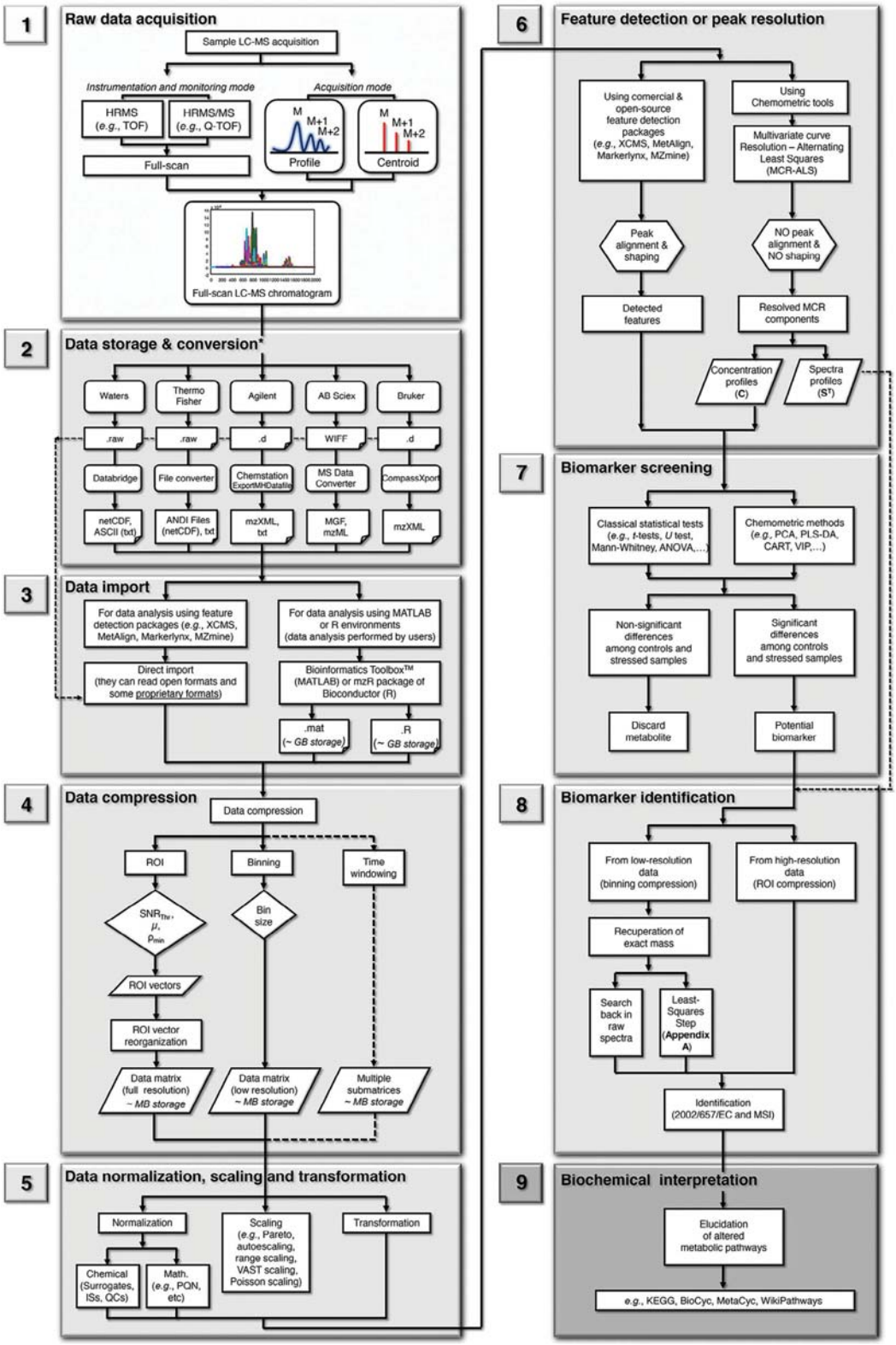

1
Raw data acquisition
Sample LC-MS acquisition
Instrumentation and monitoring mode
Acquisition mode
HRMS (e.g., TOF)
HRMS/MS (e.g., Q-TOF)
Full-scan
M
M+1
M+2
Profile
Centroid
Full-scan LC-MS chromatogram
2
Data storage & conversion*
Waters
Thermo Fisher
Agilent
AB Sciex
Bruker
.raw
.raw
.d
WIFF
.d
Databridge
File converter
Chemstation
MS Data Converter
CompassXport
netCDF, ASCII (txt)
ANDI Files (netCDF), txt
mzXML, txt
MGF, mzML
mzXML
3
Data import
For data analysis using feature detection packages (e.g., XCMS, MetAlign, Markerlynx, MZmine)
For data analysis using MATLAB or R environments (data analysis performed by users)
Direct import (they can read open formats and some proprietary formats)
Bioinformatics Toolbox™ (MATLAB) or mzR package of Bioconductor (R)
.mat (~ GB storage)
.R (~ GB storage)
4
Data compression
Data compression
ROI
Binning
Time windowing
SNRTH, μ, Pmin
Bin size
ROI vectors
ROI vector reorganization
Data matrix (full resolution) ~ MB storage
Data matrix (low resolution) ~ MB storage
Multiple submatrices ~ MB storage
5
Data normalization, scaling and transformation
Normalization
Scaling (e.g., Pareto, autoescaling, range scaling, VAST scaling, Poisson scaling)
Transformation
Chemical (Surrogates, ISs, QCs)
Math. (e.g., PQN, etc)
6
Feature detection or peak resolution
Using comercial & open-source feature detection packages (e.g., XCMS, MetAlign, Markerlynx, MZmine)
Using Chemometric tools
Multivariate curve Resolution – Alternating Least Squares (MCR-ALS)
Peak alignment & shaping
NO peak alignment & NO shaping
Detected features
Resolved MCR components
Concentration profiles (C)
Spectra profiles (ST)
7
Biomarker screening
Classical statistical tests (e.g., t-tests, U test, Mann-Whitney, ANOVA,…)
Chemometric methods (e.g., PCA, PLS-DA, CART, VIP,…)
Non-significant differences among controls and stressed samples
Significant differences among controls and stressed samples
Discard metabolite
Potential biomarker
8
Biomarker identification
From low-resolution data (binning compression)
From high-resolution data (ROI compression)
Recuperation of exact mass
Search back in raw spectra
Least-Squares Step (Appendix A)
Identification (2002/657/EC and MSI)
9
Biochemical interpretation
Elucidation of altered metabolic pathways
e.g., KEGG, BioCyc, MetaCyc, WikiPathways

Fig. 3. Overview flowchart listing the nine steps (grey shaded areas) involved in the data analysis approach for untargeted studies grouped in three areas: raw data acquisition (light-grey area), data processing and feature detection (medium-grey area) and biochemical interpretation (dark-grey area). In this figure parallelograms indicate data matrices or vectors, rectangles indicate processing steps, diamonds indicate key contributional choices, corner bend figures indicate file extension formats, in rounded rectangles are LC-MS vendors and their corresponding software as well as illustrative representations of MS data and LC-MS chromatograms and other explicative information is contained in hexagons. For data conversion, other external software (*Sashimi Project* and *ProteoWizard*) can be used (see Section 3.2.2. for more information). Note that in this flowchart only MCR-ALS is presented as the peak resolution method, but other chemometric methods such as PARAFAC, PARAFAC2, ICA, can also be used (see Section 3.2.8.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

samples (QCs) (i.e., pools of several individuals having comparable characteristics that are injected all along the analytical run) [80] should be used to normalize concentrations of metabolites among sample sets and batches.

Quantitative analytical methods have generally relied on the utilization of isotope-labeled internal standards, which can be obtained following the method of Mashego et al. [81], for each metabolite analysed. This normalization strategy has been used to investigate metabolites including glycolytic and tricarboxylic acid cycle intermediates, amino acids, nucleotides and folates from cells including *Escherichia coli*, *Salmonella enterica*, yeast and human fibroblasts [81–86]. Recently, Arrivault et al. [87] have presented the criteria for the selection of most suitable isotope-labeled internal standards according to the case of study.

Using a set of selected surrogates and internal standards is a good alternative when a full set of isotope-labeled standards is not available and a single calibration curve for each metabolite cannot be applied. Actually, these methods fall in the middle between targeted and untargeted approaches and are classified as semi-targeted methods. For instance, Bijlsma and colleagues [88] utilized three internal standard references for lipid profiling representing most abundant lipid classes in their respective region of retention time. Also, Sysi-Aho et al. [89], developed the NOMIS (normalization using optimal selection of multiple internal standards) method using the variability information from multiple IS compounds to find the optimal normalization factor for each individual molecular species. On the other hand, the use of QCs enables the evaluation of the analytical platform stability and allows the correction of the intensity deviation.

Next step following normalization is metabolite quantification, performed by integrating the signals (i.e., peak height or area) of the target metabolites and building analytical calibration curves (different analytical strategies such as external calibration curves with standards, standard addition and internal standard are possible depending on the case, sample matrix effects, and detector reproducibility). As occurred in the previous step, most of targeted studies use LC-MS vendor software for metabolite quantification, whereas few of them utilize external tools for automated processing (Section 3.1.5.). Following quantification, some statistical tests may be applied in order to evaluate the significance of variations in peak areas/heights among controls and stressed samples and find most discriminant metabolites (i.e., potential biomarkers). In general, for targeted metabolomics, basic statistical tests such as Student's *t*-test, analysis of variance, and non-parametric tests like Kruskal-Wallis test may provide adequate statistical means to assess the presence of a signal and its association with a trait of interest. However, many metabolomic signals are highly correlated and thus violate fundamental assumptions of independence for these tests. In those cases, multivariate methods provide an attractive choice and also allow for other purposes such as sample classification or discrimination (see Section 3.2.9. where some of these methods are described). For instance, Bajoub et al. [90] used principal component analysis (PCA) combined with partial least squares-discriminant analysis (PLS-DA) to classify 25 olive oil samples belonging to five different varieties and to build predictive models for varietal classification. In this targeted metabolomic study Bajoub and colleagues could identify the varietal markers for extra-virgin olive oil obtained from *Arbequina*, *Picual*, *Cornicabra*, *Hojiblanca* and *Frantoio cv*. After quantification and assessment of statistical relevance, it is possible to make a biological interpretation of the data. This final step is described together for both targeted and untargeted approaches in Section 3.3.

3.1.5. Data analysis steps all-in-one: tools for automated processing

Some software tools for the analysis of metabolomic data obtained in targeted studies have been developed. Some of the most recent are MRMPROBS [55,56], metabolite mass spectrometry analysis tool (MMSAT) [57] and OpenChrom [59]. MRMPROBS allows metabolome analysis of large-scale SRM experiments. This program provides a process pipeline from the raw-format import to high-dimensional statistical analysis. To convert SRM raw data files to ABF (analysis services backup file) format, MRMPROBS uses an independent and freely available converter at http://www.reifycs.com/english/AbfConverter/, which supports four vendor formats: Agilent Technologies (.d), Shimadzu (.LCD), AB Sciex (.WIFF) and Thermo Fisher Scientific (.raw). In addition, this software also supports the mzML data format, provided by open-source file translators such as *ProteoWizard* (described in more detail in Section 3.2.2.), which also allows Waters (.raw) files to be imported. In order to identify the metabolites, an SRM standard library of 301 metabolites with 775 transitions is available. Such library containing SRM transitions with information of precursor and product m/z values can also be prepared by users and imported as a txt file. The output files of this software (e.g., data tables, statistical analyses such as PCA) can be exported in tab-separated text and image formats (JPEG, PNG, BMP, TIFF and GIF) for PCA. On the other hand, MMSAT is a software platform for automated quantification of metabolites from SRM experiments. This software can be used independent of any MS instrument and is compatible with mzXML converted data (obtained using open source-file translators such as *Proteowizard*) from major mass spectrometer vendors. It allows automatically detection and quantification of metabolites present across all SRM transitions, such that no prior knowledge of metabolites is required. The output quantitative data can be exported in tab delimited format to facilitate downstream statistical analysis and visualization using packages such as *Excel* or *R*. Finally, OpenChrom is an extensible cross-platform open source software for the analysis of LC-MS data, available free of charge at http://www.openchrom.net. This approach supports Agilent data formats as well as XML, mzXML and netCDF open formats and provides tools to correct baselines, to detect, integrate and identify peaks and to compare mass spectra.

The three automated platforms hereby described, together with other existing tools such as MRMer [58], appear as an alternative procedure for researchers who want to analyse LC-MS data out of vendor software. The readers interested on these types of tools are advised to consult OMICtools (http://omictools.com) and ms-utils (www.ms-utils.org) platforms. OMICtools is an online platform for genomic, transcriptomic, proteomic, and metabolomic data analysis that contains 11130 tools classified by omic technologies, applications and analytical steps. The other platform, ms-utils, provides comprehensive lists of tools, some of them designed for data visualization and analysis, format conversion, peak picking and deconvolution, calibration and alignment and retention time prediction.

3.2. Data processing steps for untargeted studies

3.2.1. Raw data acquisition

Untargeted analysis of LC-MS data is performed using high-resolution mass spectrometers such as TOF and orbital ion trap and hybrid instruments such as quadrupole/Q-TOF and quadrupole/orbital ion trap [52], operating in full-scan. Only when using GC-MS, low-resolution single quadrupoles also permit identification of metabolites in untargeted studies due to the specific fragmentation pattern of the compounds analysed [91].

Moreover, as previously stated, the acquisition mode of LC-MS data (i.e., centroid or profile, Figs. 2 and 3) is influential on the final identification of metabolites, which is enhanced with profile data, since profile acquisition allows the determination of fine isotopic distributions. See Section 3.1.1. for a detailed explanation.

3.2.2. Data storage and conversion

Once the full-scan LC-MS chromatograms are acquired, the first step required previous to their analysis involves the conversion of their original proprietary formats, which are difficult to analyse outside the vendor software, into open data formats that are readable in most standard statistical environments (e.g., MATLAB or R). Among the existing open data formats, the most popular are XML-based formats (mzXML, mzData [92] and mzML [93]), netCDF [94] (also known as ANDI-MS) and classical text files (e.g., JCAMP-DX [95] or txt). Most software packages of LC-MS manufacturers have tools that enable the conversion of proprietary data formats into open data formats (see Fig. 3). Waters and Thermo Fisher provide vendor software (*Masslynx* and *Xcalibur*, respectively) with specific tools for data conversion (*Databridge* and *File converter*, respectively). *Databridge* tool allows conversion of Waters raw data into netCDF or ASCII (txt) files whereas *File converter* enables the conversion of Thermo Fischer raw data into ANDI Files (netCDF format) or txt files (please refer to a detailed LC-MS data conversion protocol [96]). Also, Bruker and AB Sciex vendors have developed freely available external software (*CompassXport* and *MS Data Converter*, respectively), which allow the conversion of raw files (.d and .WIFF format, respectively) into mzXML for Bruker Corporation and into MGF peak lists or mzML files for AB Sciex. Finally, data acquired using Agilent instruments (.d files) can be directly converted using *Chemstation* but *MassHunter* files need the use of the *ExportMHDatafile* tool, which allows the conversion to mzXML format.

In all those cases, some external software (or projects) for data conversion can be used. On the one hand, the *Sashimi Project,* included in the trans-proteomic pipeline (TPP) [97] and, founded by the proteomics group of the Institute for Systems Biology in Seattle, contains converters that read different vendor-specific data and convert them into mzXML format. Another popular software, *ProteoWizard*, contains a set of open-source, cross-platform tools and libraries for proteomics data analysis, specifically suitable for reading and conversion of a large variety of vendor-specific formats into open data formats [98]. In particular, *ProteoWizard* uses a command line tool named *msconvert* (available with a graphical user interface as well), also included in the *Sashimi Project,* which allows the conversion of vendor formats into several open data formats, including mzML, mzXML and txt. In Fig. 3, raw data extension formats and final data extension formats of most important LC-MS manufacturers are shown, together with the software options that enable such conversions. Only when using feature detection packages that can read proprietary formats [e.g., various forms (X) of chromatography mass spectrometry (XCMS) [99]], data conversion is no necessary (dashed line in steps 2–3 of Fig. 3).

3.2.3. Data import

Once files have been converted into open data formats, next step is their import into the data analysis platforms. As observed in Fig. 3, when using feature detection packages [e.g., XCMS [99], MetAlign [100], Markerlynx, MZmine [101,102]], such import is direct since they contain specific tools for that purpose. Several feature detection packages have been developed for untargeted MS-based metabolomic data analysis. The readers interested in these tools are advised to consult OMICtools (http://omictools.com) and ms-utils (www.ms-utils.org) platforms. For data analysis performed by researchers, either in MATLAB or R environments, such import is possible using distinct strategies.

When working in MATLAB environment, the quickest and easiest method for LC-MS data import is the use of the routines included in the *Bioinformatics Toolbox™*. A step-by-step example providing details of these routines is shown by Gorrochategui et al. [96]. When working in R environment, LC-MS data are usually imported by means of the mzR package available at Bioconductor [103,104]. mzR provides a unified interface for most of the open data formats described above such as mzXML, mzML, mzData and netCDF. The key function of this package is *openMSfile* which allows exporting the information from the MS open formats to a format-specific mzR object with all the MS raw data and metadata contained in the original files. Afterwards, *peaks* function can be used to extract all MS spectral data into a matrix to be further analysed. In addition to this possibility for accessing to MS raw data for the experienced researchers, the mzR package is also used in the most popular R-based feature detection packages (i.e. XCMS [99] and MSnbase [104]) for data import.

3.2.4. Data compression and matrix construction

Handling LC-MS data in its raw form is difficult because of their large size. Thus, data compression is usually necessary to reduce them into more computationally manageable formats and avoid issues associated with the limited memory capacity of the computers, but preventing a loss of experimental information during the process. In addition to compression, the initial LC-MS data sets containing scans of unequally spaced masses must be mapped onto matrices with rows representing each of the scans (i.e., retention times) and columns representing the same mass values in all samples.

Different methodologies enable data compression as well as their processing or visualization in its native two-dimensional form. Among them, the procedures of "binning" and the "search of regions of interest (ROI)" are the most adequate to the nature of LC-MS data sets. Apart from these methodologies, in this section we also shortly describe another strategy that is commonly used together with the binning compression in order to further reduce data dimensions: time windowing.

Binning. Binning is one of the most used procedures for raw LC-MS data compression. The application of binning involves the transformation of raw data into a matrix representation (x,y), with retention times in the x-dimension and m/z values in the y-dimension. Conversion of high-resolution raw mass spectra into a matrix representation requires the division of the m/z axis into equidistant sections with a specific bin size. Thus, the compression of the data and their mapping to a matrix are carried out at the same time. However, as a consequence, a relevant drawback of the binning procedure is the difficulty associated with the proper selection of the bin size for a particular data set, being this parameter strongly related to the chromatographic profile. If the bin size selected is too small, chromatographic peaks might alternate among bins and thus not be detected due to the loss of the chromatographic peak shape. On the contrary, if the bin size is too large, multiple coelutions between peaks can exist, and small peaks may disappear by the increased noise level. Another disadvantage of the binning procedure is the loss of spectral resolution derived from the data compression performed in the m/z-mode dimension [37].

Fig. 4 shows an example of the binning procedure applied to a region of an LC-HRMS chromatogram, with a bin size of 0.1 ppm. The

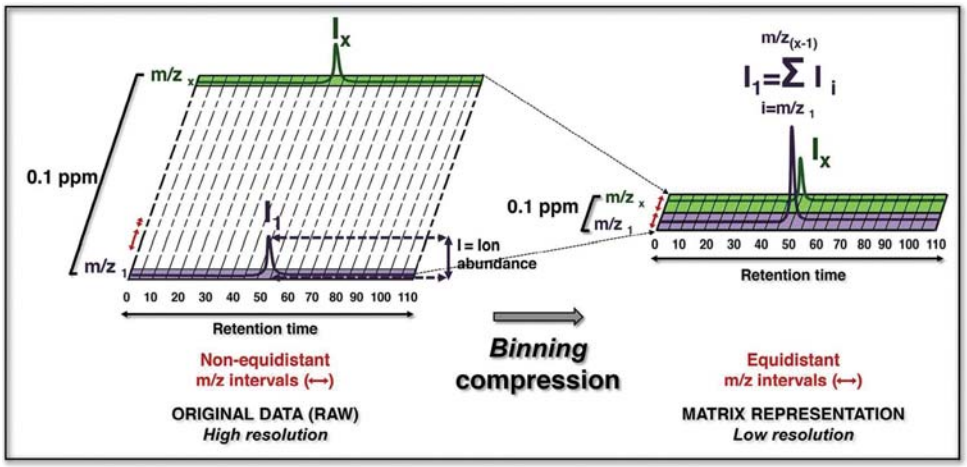


Fig. 4. Scheme of the steps involved in the compression of data when using binning. Example shown for a particular region of an LC-HRMS chromatogram, using a bin size of 0.1 ppm.

intensities corresponding to all m/z values comprised between the lower limit (m/z_1) and the upper limit ($m/z_1 + 0.1$ ppm) are added up and attributed to m/z_1, thus decreasing file size but also the spectral resolution.

Regions of interest (ROI). Data compression based on the search of ROI is an alternative technique to the binning procedure. This method, first presented by Stolt et al. [105], is based on the concept of considering analytes as a region of data points with a high density ranked by a specific "data void". These ROI contain data from interesting mass traces, which means values with a significant intensity higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}). Moreover, ROI must contain a minimum number of consecutive data points (ρ_{min}) compressed within a particular mass deviation (μ), typically set to a generous multiple of the mass accuracy of the mass spectrometer. This condition prevents ionic signals or noise to be considered as an ROI. In Fig. 5a an example of a mass trace for a particular region of the chromatogram obeying these criteria and thus, considered as an ROI (ROI_i), is represented. As shown in this figure, ROI_i can be clearly distinguished from low-intensity signals that are subsequently filtered out. As shown in Fig. 5b, ROI are searched among all the chromatogram and vectors of distinct length (depending on the number of ROI found at each retention time) are obtained. Finally, these vectors are reorganized into a matrix. To do that, common ROI among all the retention times are grouped and final m/z of each ROI (mzmean) is calculated as the mean of all the m/z values from the series of data points grouped within the same ROI. The obtained matrix contains the retention times in the x-dimension and the final mzmean values of ROI in the y-dimension (Fig. 5c).

With the ROI compression, no loss of spectral accuracy occurs, as opposed to the binning strategy. ROI strategy was introduced in the *centWave* algorithm of XCMS software [99] and it is increasingly used in feature detection packages as a substitute to the classical binning [37].

Time windowing. This strategy is based on the partition of the LC-MS chromatograms into distinct regions of time (i.e., time windows) to be analysed separately [106–108]. It is an additional step used to further reduce sample size if data compression using binning is not sufficient. The level of compression achieved with the ROI strategy is generally high enough so that entire chromatograms can be analysed at a time.

3.2.5. Data intensity normalization, scaling and transformation

In untargeted approaches, three strategies can be used for removing the unwanted systematic bias in the measurements: sample normalization, data scaling and data transformation. Sample normalization is necessary to adjust the differences among samples whereas data scaling and transformation allow the comparison among metabolites of distinct samples. Thus, normalization refers to row-wise corrections (i.e., within chromatograms) whereas scaling and transformation refer to column-wise corrections (i.e., between chromatograms).

Sample normalization strategies can be chemical or mathematical. The first ones, which are based on the use of a single or multiple surrogates, internal standards, and quality controls, have been already described in the targeted approach (see Section 3.1.4). On the other hand, mathematical normalization strategies use computation models to achieve the same purpose. A numerical normalization method based on the use of QCs proposed by Dunn et al. [109] is the locally estimated scatterplot smoothing (LOESS). In this method, each variable in each sample is individually corrected according to the evolution of its value in the neighbouring QCs. Also, van der Kloet et al. [110] proposed in 2009 a correction based on the average or on the median of the QC replicates analysed in different batches. A novel and alternative method for correction of analytical bias is common components and specific weights analysis (CCSWA), originally developed by Qannari et al. [111] and recently used by Dubin et al. [112] for correction of analytical bias. This method is reported as a good alternative to LOESS signal correction when samples and QCs do not behave in the same way. Other mathematical normalization strategies are based on the assumption that the signal of the majority of metabolites is stable. Under this assumption, normalization can be efficiently achieved by calculating the relative ratio of abundance of metabolites respect to all other peaks (e.g., unit norm [113] and median intensities normalization [114]). However, these strategies fail when changes in concentration of metabolites occur due to laboratory system errors and (or) differences among large scale biological experiments. In these cases, normalization based on the total chromatogram is not appropriate and can cause serious data distortions. Another normalization method widely used is the probabilistic quotient normalization (PQN) [115]. This method scales all the intensities in a spectrum using the most probable multiplicative factor calculated as the median of the quotients of the

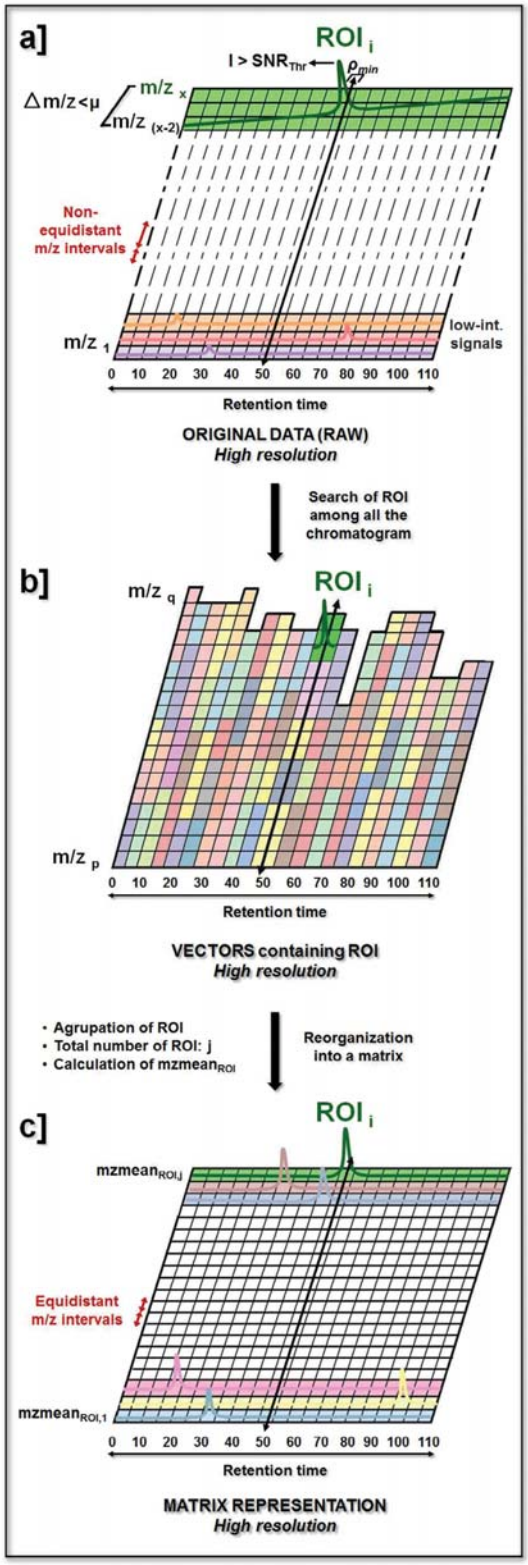


Fig. 5. Scheme of the steps involved in the compression of data by the search of ROI: a] original data with non-equidistant m/z intervals where a significant mass trace is represented as ROI_i (green) and distinguished from low-intensity signals (orange, pink and violet), b] vectors containing the distinct ROI (represented by sequences of squares of the same colour) obtained at different regions of the chromatogram, including the previous ROI_i (green) and c] matrix constructed from the reorganization of ROI vectors, again containing the same ROI_i (in green). (SNR_{Thr}: signal-to-noise ratio threshold, mzmean: mean of all the m/z values from the series of data points grouped within the same ROI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

amplitudes of each point in a spectrum and a reference spectrum. PQN normalization is highly recommendable for cases were size effects are noticeable, and internal normalization is not suitable since it destroys relative peak information within the chromatogram.

Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the fold differences between the distinct metabolites by converting the data into differences in concentration relative to the scaling factor [116]. Depending on the scaling factor used, scaling methods are divided in two subclasses. The first class uses a measure of the data dispersion (e.g., standard deviation) as a scaling factor, while the second class uses a size measure (e.g., the mean). Scaling methods that use a dispersion measure for scaling include autoscaling [117], Pareto scaling [118], range scaling [119], and variable stability (VAST) scaling [120]. Autoscaling [117], also called unit or unit variance scaling, is the most used in metabolomics and it provides equal variance to each variable (i.e., all metabolites have a standard deviation of one). Pareto scaling [118] is very similar to autoscaling, but instead of the standard deviation, the square root of the standard deviation is used as the scaling factor. Range scaling [119] uses the range (i.e., difference between minimal and maximal value or concentration of a metabolite in a set of experiments) as the scaling factor. VAST scaling [120] is an acronym of variable stability scaling and it is an extension of autoscaling. Scaling methods based on average value include level scaling, which converts the changes in metabolite concentrations into changes relative to the average concentration of the metabolite and Poisson scaling or "square root mean scale", which scales each variable by the square root of the mean of the variable. Examples of Poisson scaling to correct MS data effectively are found in the literature [121,122].

Finally, transformations are nonlinear conversions of the data such as the log and the power transformation [116]. These methods are commonly used to correct for data heteroscedasticity [123], which in the case of metabolomic data refers to non-equal variance uncertainty variations related to some or all metabolites under analysis.

Some of the existing LC-MS feature detection frameworks allow normalization based on the use of internal standards and scaling. For instance, the algorithm of MZmine 2 [102], called *linear normalizer*, divides the height or area of each peak by a normalization factor, such as the average of peak height, the average of the squared peak height, the maximum peak height or the total raw signal within the chromatogram. In contrast, MetaboAnalyst [124,125] performs normalization (to allow comparisons among samples) and scaling (to allow comparisons of magnitude of features) sequentially. Wu et al. [126] have recently provided a summary of the reported sample normalization methods used over the past several years together with their pros and cons. They conclude that for the appropriate selection of a normalization methodology, the biological system of study must be thoroughly evaluated. In this study, Wu and colleagues propose two distinct normalization methodologies, one for urine samples and another for cellular extracts.

3.2.6. Feature detection or peak resolution

Feature detection and peak resolution are two closely-related concepts. Feature detection aims to search for features, using the term

"feature" for a bounded, two-dimensional (m/z and retention time) LC-MS signal [37]. On the other hand, peak resolution[1] aims to identify the pure components responsible for these features, associated with a pure spectrum or elution profile, after solving some chromatographic problems (e.g. coelutions). Generally, feature detection is carried out by different algorithms featured in available software. On the other hand, some chemometric methods have also been developed to resolve second order data such as LC-MS data. Among them, multivariate curve resolution-alternating least squares (MCR-ALS) [127] has proved to be powerful when dealing with LC-MS metabolomic data sets [96,106,107,128–133]. The ultimate goal of feature detection and peak resolution is to distinguish real chemical compounds from false positives (e.g., background noise).

Most of the existing feature detection packages [e.g., XCMS [99], MetAlign [100] and MZmine [101,102]] require preliminary peak alignment and usually peak shaping previous to feature detection. On the other hand, some chemometric methods such as MCR-ALS allow peak resolution without previous peak correction. A detailed explanation of both methodologies is shown below.

3.2.7. Feature detection (and alignment)

In most of feature detection software, peak alignment is necessary in order to search for corresponding peaks across distinct chromatographic runs and compare them between samples. Together with peak alignment, peak shaping is generally applied so that peaks finally have a defined and more symmetrical shape, usually fitting a Gaussian curve to the experimental features.

The search for corresponding peaks is a cumbersome task since matching peaks usually have differences in m/z and retention time values [134]. In fact, when searching for matching peaks, some remarks should be made. First, the differences in the retention time across samples may be non-linear. Second, a feature in a sample may have multiple possible matching features based on m/z and retention time values, potentially leading to false matching. Finally, some peaks may not appear in some samples [49].

Because of the issues mentioned above, different alignment algorithms have been proposed to correct retention time differences among samples. Considering the most popular feature detection packages, some of these algorithms can be highlighted. First, the OBI-warp [135] (ordered bijective interpolated warping) method, used in the XCMS software, which allows aligning matrices along a single axis using dynamic time warping (DTW) together with a bijective (one-to-one) interpolated warp function. Thus, OBI-warp (first used in the proteomics field) produces a smooth warping function able to align multiple chromatographic runs. Alternatively, an alignment method based on the random sample consensus (RANSAC) [136] algorithm is used in the MZmine 2 [102] software. RANSAC is an iterative method that allows the estimation of parameters of a mathematical model by random sampling of the observed data that could contain outliers. Finally, the combination of RANSAC and LOESS regression allows the determination of optimal parameters of the mathematical model for peak alignment. More options for peak alignment can be found in the review works of Katajamaa [79] and Bloemberg [137]. Concerning peak shaping, some feature detection algorithms initially used models of specific peak width to fit features (e.g. *Matched Filter* algorithm of XCMS software [99]). However, those models failed when the selected peak width did not fit all features properly.

In order to overcome this issue, some feature detection packages (e.g., *centWave* algorithm of XCMS) use continuous wavelet transform (CWT) to perform peak shaping. The CWT reliably detects chromatographic peaks of differing width and is widely used in signal processing and pattern recognition [138], and furthermore is able to resolve an additional problem concerning feature detection, as it is the presence of close-by or coeluted peaks. With the CWT analysis, the intensity of every peak is estimated by the maximum value of the centroid peak in the calculated peak boundaries. The same approach can be used to eliminate noise contributions known as "shoulder peaks" (small peaks from residues of the Fourier transform calculated by the MS instrument). These contributions can also be removed by fitting a theoretical model (e.g., Gaussian or Lorentzian).

3.2.8. Peak resolution (without alignment)

Recently, some little explored but highly useful chemometric tools have proved to be powerful methods for LC-MS metabolomic data analysis. Among them, MCR-ALS has emerged as a powerful tool to resolve the profiling problems in LC-MS metabolomic data sets without previous peak correction [127]. MCR-ALS is based on Equation (1):

$$D = CS^T + E \quad (1)$$

It is seen that MCR-ALS methods share the underlying bilinear mathematical model of PCA but under completely different constraints and with a different goal. In the case of LC–MS data, **D** matrix ($I \times J$) contains the MS spectra at all retention times ($i = 1, \ldots I$) in its rows, and the chromatograms at all spectra m/z channels ($j = 1, \ldots J$) in its columns. This data matrix is decomposed in the product of two factor matrices, **C** and $\mathbf{S^T}$. The **C** ($I \times N$) matrix contains column vectors which correspond to the elution profiles of the N ($n = 1, \ldots N$) pure components of matrix **D**. In $\mathbf{S^T}$ ($N \times J$) matrix, row vectors correspond to the spectra of the N pure components. The part of **D** that is not explained by the model forms the residual matrix, **E** ($I \times J$). MCR-ALS methods assume that the variation measured in all samples in the original data set can be described by a combination of a small number of chemically meaningful profiles. In the case of LC–MS data sets, information of the data table can be reproduced by the combination of a small number of pure mass spectra (row profiles in the $\mathbf{S^T}$ matrix) weighted by the concentration of each of them along the elution direction (the related chromatographic elution peaks, column profiles in **C**). As a result from the MCR-ALS analysis, we obtain a set of components, with their corresponding elution and spectra profiles. The equivalence between an MCR-ALS component and a feature is high since both of them correspond to a chemically meaningful profile. However, they differ in the fact that one feature is associated with a unique m/z value whereas one MCR-ALS component can be associated with various m/z values (i.e., distinct m/z values can describe the same elution profile).

As previously stated MCR-ALS analysis allows powerful LC-MS data resolution without previous peak alignment or shaping. The reason why peak alignment is not required is attributed to the fact that alignment is produced in the spectral dimension (m/z values), which is common among all samples, and not in the time dimension, which can vary among samples. This is useful with LC-MS data sets, but even more with capillary electrophoresis-mass spectrometry (CE-MS) data sets, which contain analytes showing important retention time peak shifts among samples that in some cases cannot be properly corrected when using feature detection (and alignment) algorithms. The number of MCR-ALS models required to resolve peak signals of one sample depends on the size of the data matrix. Generally, for data compressed using binning strategy, compression is not sufficient, and MCR-ALS has to be applied individually to distinct time windows of the chromatogram (see Section 3.2.4.). On the contrary, when using ROI strategy, the obtained data matrices are small enough so that one MCR-ALS model is generally sufficient to resolve peak signals of the entire chromatographic profile. The readers interested in MCR-ALS analysis are advised to consult http://www.mcrals.info/.

[1] The term "deconvolution" is analogue to "resolution" but is preferred to be used for univariate signals [i.e., first order data (data vector)], whereas resolution is preferred for multivariate signals [i.e., second order data (data matrix)].

There are significant differences between the approaches used by MCR-ALS respect to other feature detection packages, such as XCMS, concerning peak resolution and feature detection strategies. However, a study based on the evaluation of changes induced in rice metabolome by Cd and Cu using LC-MS [132] concluded that both methodologies provided similar results, which suggests that despite the existing differences among these approaches, they are equally valid to analyse LC-MS metabolomic data sets.

Apart from MCR-ALS, other methods for the processing of second-order data are available. Among them, PARAFAC (parallel factor analysis) [139,140], TLD (trilinear decomposition), PARAFAC2 (parallel factor analysis2) [141,142] and independent component analysis (ICA) are some methods proposed for the same goal. PARAFAC and TLD methods require the data to follow the so-called trilinearity model (i.e., all chemical components are defined by a unique elution and spectral profile in all samples, apart from a scale factor). However, LC data do not obey the trilinear model in general, since analyte peaks usually show retention time shifts and peak shape changes from sample to sample, causing trilinearity deviations. In order to restore the trilinearity, PARAFAC and TLD methods should mathematically pre-process each data matrix, so that analyte peaks are properly aligned. Even in this case however, possible run to run peak shape differences compel the fulfillment of the trilinear model in many circumstances. On the other hand, PARAFAC2 employs a more flexible algorithm, which permits a given component to have different time profiles. A study of Khakimov et al. [143] demonstrated the efficiency of PARAFAC2 for exploring complex plant metabolomics LC-MS data. In that study, PARAFAC2 enabled automated resolution and quantification of several elusive chromatographic peaks (e.g., overlapped, elution time shifted and low s/n ratio). However, Bortolato and Olivieri [144] compared the performance of PARAFAC2 and MCR-ALS, arriving at the conclusion that PARAFAC2 produces artificial outputs when elution profile changes are severe, and interferents are present in test samples and therefore, confirmed the higher power and range of applicability of MCR-ALS. Another alternative to PARAFAC, PARAFAC2, TLD methods and MCR-ALS is ICA. The main idea of ICA [145] is to find a mathematical transformation of the data into a linear combination of statistically independent components. However, the condition of independence is generally not fulfilled when using ICA with chromatographic data [146,147]. Among ICA methods, mean-field ICA (MFICA) [148] is the best for multivariate resolution, due to the application of non-negativity constraints in both data modes (i.e., concentration and spectra profiles), and is the only one that can be strictly compared to MCR-ALS. However, the advantage of MCR-ALS is that it is more flexible since it allows the implementation of other constraints (e.g., unimodality, closure, local rank, selectivity or the multi-linear type of constraint) [146]. Recently, Liu et al. [149] have developed a new method named MetICA, inspired from the original *Icasso* algorithm, for the application and validation of ICA on untargeted metabolomic data sets. In that study, the efficacy of MetICA routine was tested on simulated and real MS-based yeast exo-metabolome data.

3.2.9. Biomarker screening or variable selection

Biomarker screening (variable selection) plays an essential role in metabolomics [150,151]. Biomarkers are defined as biological entities that can be used to indicate the status of healthy or diseased cells, tissues, or individuals. Thus, they correspond to molecular markers (i.e., metabolites in the case of metabolomics) that can better discriminate among control and stressed samples, in terms of their concentrations.

However, it is unfortunately quite easy to find markers that, despite being apparently relevant, are in fact spurious. The main sources of error in this aspect, which are not entirely independent of each other, include bias, inadequate sample size (especially relative to the number of metabolite variables and to the required statistical power to prove that a biomarker is discriminant), excessive false discovery rate due to multiple hypothesis testing, inappropriate choice of particular numerical methods, and overfitting (generally caused by the failure to perform adequate validation and cross-validation). Many studies fail to take these problems into account, and thereby fail to find anything significantly true [152]. For instance, classical p-values such as "$p < 0.05$" that are commonly used in biomedicine are far too optimistic when multiple tests are done simultaneously (as occurs in metabolomics) [150]. Indeed, one type of bias, known as "p-hacking", occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant. Head et al. [153], studied the extent and consequences of p-hacking in science arriving at the conclusion that this type of bias probably does not drastically alter scientific consensuses draw from data analyses. However, methods to measure such error and to correct them are highly recommendable.

The classical methods used for biomarker selection were proposed by statisticians and were based on the application of statistical hypothesis testing (e.g., t-tests, Mann-Whitney U test, ANOVA). However, other methods envisaged for biomarker screening have been proposed lately by numerous chemometricians. Some of these methods include PCA [154], ICA [145], PLS-DA [155], linear logistic regression (LLR) [156], classification and regression trees (CART) [157], selectivity ratio (SR) [158,159] and variables importance on projection (VIP) [160]. Another method valid for variable selection is ANOVA-simultaneous component analysis (ASCA) [161,162]. This method can be understood as a direct generalization of ANOVA analysis of variance for univariate data to the multivariate case. ASCA method incorporates the information of the structure of data sets (i.e., underlying factors such as time, dose or combinations thereof), enabling a better understanding of their biological information.

To date, the most popular variable selection method in metabolomics is the VIP [160] method. However, the main drawback of this approach is related to the proper selection of the threshold value. Despite some studies select variables with VIP scores greater than 1 [163,164], such criterion is not always used and the results found in the literature are not always comparable. A study by Gorrochategui et al. [108] compared the number of biomarkers found when using an ANOVA test ($p < 0.05$) followed by a multiple comparison's test and those obtained when using the VIP method fixing distinct threshold values. As it was observed, the number of encountered biomarkers was different in each case, although some of them were common among the strategies. Another method facing the challenge of a proper threshold value selection is SR. Actually, the use of the threshold suggested by the authors Rajalahti et al. [158,159] based on an F-test to define the boundary between variables with high discriminating ability and less interesting regions, is unusually valid for raw large chromatographic data sets, such as LC-MS metabolomic data sets [165]. In those cases, SR can lead to a selection of a reduced number of variables, sometimes not including relevant biomarkers. An alternative strategy to increment the number of selected variables using SR method is the use of ad hoc limits (e.g., average SR over the training set).

Despite the VIP method being the most used in metabolomic studies, there is still some disagreement about which is the best approach for variable selection and a critical evaluation needs to be performed before any of them is selected and, also, once the results have been obtained. Checa et al. [166] concluded that the most crucial step when performing lipidomic data analysis is the proper choice of the chemometric variable selection method according to the crude data. Studies comparing the performance of several of these methods exist in the literature. For instance, Farrés et al. [165] compared SR and VIP variable selection methods observing that in general terms, the VIP method selected a higher number of variables than the SR method. However, they arrived at the conclusion

that final decision about which is the best approach should be performed according to the aim of the study. Also, Andersen et al. [167] concluded that in essence, variable selection should rather be considered as variable elimination where the clearly irrelevant parts are removed and the remaining parts containing potentially useful information are kept for further data analysis.

In order to ensure good performance of the selected discrimination model, further statistical validation of the model is required. Such validation becomes particularly necessary in the case of "undersampling" (i.e., when having a low number of samples compared to the number of variables), since the reduced number of samples becomes insufficient to properly describe the groups and find significant biomarkers. Some of the statistical validation tools that can deal with this problem consist of permutation tests [168], single and double cross-validation [169,170], and the combination of the latter with a new variable selection method, called ranked products [171]. Permutation tests give information about the discrimination performance of the model, which should at the same time be able to properly classify new samples as "stressed" or "control". However, testing the classification ability of the model is impossible when having low number of samples and for this reason, permutation tests are mostly used to evaluate the significance of the discrimination. Double cross-validation takes a better advantage of the data and is the chosen method to estimate the error of the model in classifying unknown samples. Cross-validation procedures generate several models. However, those procedures only give a reliable error rate when the complete modelling step is cross-validated. Cross-validation methods together with bootstrap [172] and jack-knifing methods are classified as resampling methods [173], and are used to determine the optimal number of components in a partial least squares (PLS) regression model [174,175]. Moreover, these methods allow the estimation of the uncertainty of individual variables, in order to find the relevant ones (e.g., relevant VIPs to determine candidate biomarkers). Afanador et al. [176] demonstrated how the use of bootstrapping, in conjunction with permutation tests and the use of 95% lower-bound on the jack-knife confidence interval provide avenues for improvement of the important variable selection process. Finally, the rank products procedure can be described as a natural partner for cross-validation to evaluate the overall importance of a variable. Overall, a combination of these tools for statistical validation of discriminant models is frequently the best option. Smit et al. [171] presented a strategy for the discovery and rigorous statistical validation of candidate biomarkers for proteomics based on the combination of principal component discriminant analysis (PCDA), permutation tests, double cross-validation and variable selection with rank products. A tutorial of validation tools for chemometric models shows how the selection of the level of validation and the method for analyzing data may impact the conclusions and chemical insight gained [173].

3.2.10. Biomarker identification

As stated in Section 3.1.3., the identification of metabolites is a complex task, and it becomes even more complicated in untargeted metabolomic studies. In 2013, Dunn et al. [177] reviewed all the available experimental and computational tools to identify metabolites in untargeted metabolomic studies. In this review, they concluded that the number of identified metabolite features has increased in the last decades due to enhanced mass spectrometry and increased mass resolution, but the proportion of identified metabolites remains still low (ca. 50%). The criteria [68] and directives [69] for the identification of MS data previously presented in the targeted approach are also valid for the untargeted approach. In contrast to targeted studies which can use either LRMS/MS or HRMS instrumentation, untargeted studies are possible with HRMS or high-resolution tandem mass spectrometry (HRMS/MS). Li et al. [178] have recently reported that liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (LC-QTOF-MS) to investigate natural products provides efficient separation and good sensitivity. Also, it allows for the identification of the fragmentation pathways of metabolites [179] and [180], by employing newer mass spectrometry$^{\text{Elevated energy}}$ (MS^E) methods to acquire MS/MS (without specific precursor ion selection) data at both low and high energy from a single injection [181]. Moreover, LC-QTOF-MS^E is proved to be a very versatile technique in metabolomics and it has been shown to be increasingly powerful [182].

However, the high mass accuracy provided by HRMS instrumentation can be partially lost when using binning in the compression step (see Section 3.2.4.). In those cases, HRMS data can be recovered using two approaches.

First, HRMS data can be obtained by looking back in the raw spectra: after the peak resolution step (for instance using MCR-ALS) has been performed on data compressed by binning, those peaks tagged as potential biomarkers are identified by direct comparison with the HRMS spectra. For instance, Bedia et al. [133] identified the lipid species (including phospholipids, sphingolipids, glycosphingolipids and cardiolipin species) altered after long-term exposure of prostate cancer cells to endocrine disruptors using this approach, even though original data were binned with an m/z resolution of 0.05 ppm. The second method consists in a least-squares step which allows HRMS spectra to be obtained from the MCR-ALS elution profiles of binned data and the original HRMS data for a set of LC-MS chromatograms (or the same region of the chromatogram in the case of time windowing). See Appendix A for a detailed explanation of the latter procedure. It should be noted that since the ROI method, used in many of the LC-MS feature detection packages, does not decrease the resolution of the MS data, there is no need for applying these strategies when this compression technique is used. Finally, as stated in Section 3.1.3., another aspect can contribute to an enhanced identification: acquisition in profile mode.

3.3. Final common step: biochemical interpretation

The overall process of LC-MS data analysis ends with the ultimate biological interpretation of the results through the elucidation of the metabolic pathways linked to the identified biomarkers. In targeted metabolomic studies that are driven by an initial biological hypothesis, final interpretation is usually reduced to a confirmation of the predicted alterations. Only in those cases where initial predictions are not fulfilled the unknown altered pathways have to be deciphered. In untargeted metabolomics elucidation is always necessary.

Altered metabolic pathways can be deciphered by consulting online databases such as KEGG (kyoto encyclopedia of genes and genomes) (http://www.genome.jp/kegg/kegg2.html) [183], Biocyc (http://biocyc.org) [184], MetaCyc (http://MetaCyc.org/) [185] or WikiPathways (http://www.wikipathways.org) [186,187]. The representation of these altered pathways in global maps showing an overall picture of metabolism helps to obtain a reliable biological interpretation of the studied system. For instance, Farrés et al. [107] and Ortiz-Villanueva et al. [131] studied the metabolic changes occurring in stressed baker's yeast (*Saccharomyces cerevisiae*) samples. With the help of KEGG database both studies characterized most discriminant metabolites and identified the metabolic pathways with the highest participation in the acclimatization of baker's yeast cells to grow at distinct temperatures (i.e., 42 and 37°C, respectively). Also, Chu et al. [188], studied the therapeutic mechanism of *Rhizoma Alismatis*, a crude herb component in traditional Chinese medicine, on spontaneous hypertensive rats using ingenuity pathway analysis (IPA). With the help of KEGG, HMDB and METLIN databases the authors found the potential biomarkers and potential target pathways of *Rhizoma Alismatis* species. Moreover, Perl et al. [189] studied the mechanism of impact of the amino acid precursor,

N-acetylcysteine (NAC), on the metabolome of systemic lupus erythematosus (SLE) patients by quantitative metabolome profiling of peripheral blood lymphocytes (PBL) using mass spectrometry. The results of this study showed that metabolome changes in lupus PBL affected 27 of 80 KEGG pathways with most prominent impact on the pentose phosphate pathway (PPP), which reflected greater demand for nucleotides and oxidative stress. Overall, their findings contributed to the identification of novel metabolic checkpoints in lupus pathogenesis.

4. LC-MS metabolomic data analysis: an active area in bioinformatics research

The development of tools for data analysis is an active area of bioinformatics research. Recent years have witnessed the development of many software tools for data analysis, but still there is a need for further improvement of the data analysis pipeline. Such improvement should concentrate on two aspects: combination of data analysis strategies and fusion of distinct omic fields.

The combination of various data analysis strategies is necessary to allow a more comprehensive detection of chemical components in LC-MS data for signature discovery. In the last years, some studies have demonstrated the advantages of combining various data analysis strategies. For instance, Coble and Fraga [190] compared the performance of four data analysis tools [i.e., XCMS [99], MetAlign [100], MZmine [101,102], and SpectConnect (this one for GC-MS data)] in terms of their ability to detect components in the chromatography-mass spectrometry data sets, arriving at the conclusion that each of them has its pros and cons. The same study also pointed out that the most pressing improvement needed for all the tested data analysis tools was to reduce the percentage of false peaks, i.e., reported features that are not true peaks, while still detecting the low-intensity peaks. Moreover, some of the existing data analysis methodologies still require a significant level of manual input, which difficults the process and can even make it prohibitive in the case of very large data sets.

The fusion of distinct omic platforms (e.g., transcriptomics, proteomics and metabolomics) is one of the latest objectives pursued by the omics community. Data fusion is a challenging task, in particular, when the goal is to capture underlying factors and use them for interpretation. Numerous strategies have been proposed for integrating data from parallel sources. Among them, some of the most used include GSVD (generalized singular value decomposition) [191], O2PLS (two-way orthogonal projections to latent structures) [192], OnPLS (multiblock orthogonal projections to latent structures) [193], DISCO-SCA (distinctive and common components with simultaneous-component analysis) [194], JIVE (joint and individual variation explained) [195], and CMTF (coupled matrix and tensor factorization) [196]. GSVD provides a comparative mathematical framework for two data sets (e.g., two genome-scale data sets). O2PLS method is build on the basis of orthogonal projections to latent structures (OPLS) [197], which is a supervised multivariate regression method. O2PLS can be used for combining "omics" types of data, separating systematic variation that overlaps across analytical platforms from platform-specific systematic variation. Bouhaddani et al. [198], evaluated the efficacy of O2PLS in the integration of metabolomic and transcriptomic data from a large Finnish cohort (DIGLOM). The results of the simultaneous analysis with O2PLS on metabolome and transcriptome data were in agreement with an earlier study and showed that the lipo-leukocyte module, together with two lipoproteins, were important for the metabolomic and transcriptomic relation. An extension of O2PLS to the multiblock case (involving more than two matrices) was later developed and called OnPLS. OnPLS method is fully symmetric (i.e., it does not depend on the order of analysis when more than two blocks are analysed) and has been used in several multi-omic studies [193,199,200]. DISCO-SCA allows distinguishing common and distinctive information in different data blocks; information that is mixed up when using simultaneous-component and multigroup factor analysis methods. JIVE [195] was created for the integrated unsupervised analysis of metabolomic profiles from multiple data sources. This method separates the shared patterns among data sources (i.e., joint structure) from the individual structure of each data source that is unrelated to the joint structure. CMTF successfully captures the underlying factors by exploiting the low-rank structure of higher order data sets and is particularly useful for joint analysis of heterogeneous data. Apart from these methods, Blanchet and Smolinska [201] have recently proposed a framework which allows the combination of multiple data sets, provided by different analytical platforms. This framework extracts relevant information for each platform in the first step. Then, the obtained latent variables are fused, analysed, and the influence of the original variables is finally calculated back and interpreted. Therefore, new advances in data processing tools should point to opening fields such as data fusion. For instance, in the case of MCR-ALS, data fusion can be easily performed by augmenting data matrices in the row-wise dimension, and some work is now being pursued in this direction.

5. Concluding remarks

From a general point of view, we can conclude that the complexity of LC-MS metabolomic data and the diversity of strategies that are used for their processing makes data analysis an open field in the bioinformatics research. In global terms, targeted strategies allow highly sensitive and accurate detection of predetermined metabolites whereas untargeted strategies are valuable for the detection of unknown metabolites and biochemical pathways. However, both approaches are complementary and can be used simultaneously. Despite recent targeted methodologies enable large-scale metabolic profiling, including hundreds of analytes, the number of compounds to be analysed in untargeted studies is still larger. This is so because entire data sets including thousands of metabolite signals have to be processed in the latter approach. For this reason, later advances in data analysis tools have been focused on the untargeted approach.

In the last years, multiple feature detection software tools for LC-MS data have been developed for untargeted metabolomics. Generally, all of them cover the same steps of data conversion, compression, normalization, feature detection, variable selection and identification. Among them, data compression is one of the most crucial steps, since it must reduce the original dimensions of the data (gigabytes of storage) while avoiding any loss of spectral accuracy. Nowadays, the search of ROI has been reported as a better alternative to the classical binning and it is used in most of these feature detection software during the compression step.

Novel chemometric tools such as MCR-ALS have demonstrated to be powerful tools to analyse LC-MS metabolomic data sets and they are presented in this review as a complement to the existent feature detection packages the use of which can also provide some benefits. The principal advantages of MCR-ALS methodology compared to other feature detection algorithms can be mainly attributed to two aspects. First, MCR-ALS can resolve the coelution chromatographic problems and directly obtain the pure spectra and elution profiles of most of the meaningful metabolites present in the sample. Second, neither peak alignment nor shaping corrections are necessary for this approach, since LC-MS chromatograms are only matched in the mass spectral direction, which is reproducible. Thus, MCR-ALS is considered and proposed as a novel and effective methodology for LC-MS metabolomic data analysis.

Although all data analysis approaches presented in this review have contributed to increasing knowledge in the LC-MS metabolomics field, more recent advances in new areas such as data fusion are still necessary.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) / ERC Grant Agreement n. 320737. First author acknowledges the Spanish Government (Ministerio de Educación, Cultura y Deporte) for a predoctoral FPU scholarship (FPU13/04384).

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.trac.2016.07.004.

References

[1] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R.N. Trethewey, L. Willmitzer, Metabolite profiling for plant functional genomics, Nat. Biotechnol. 18 (2000) 1157–1161, doi:10.1038/81137.
[2] O. Fiehn, Metabolomics – the link between genotypes and phenotypes, Plant Mol. Biol. 48 (2002) 155–171, doi:10.1023/A:1013713905833.
[3] G.J. Patti, O. Yanes, G. Siuzdak, Innovation: metabolomics: the apogee of the omics trilogy, Nat. Rev. Mol. Cell Biol. 13 (2012) 263–269, doi:10.1038/nrm3314.
[4] M. Chadeau-Hyam, G. Campanella, T. Jombart, L. Bottolo, L. Portengen, P. Vineis, et al., Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers, Environ. Mol. Mutagen. 54 (2013) 542–557, doi:10.1002/em.21797.
[5] L.M. McShane, M.M. Cavenagh, T.G. Lively, D.A. Eberhard, W.L. Bigbee, P.M. Williams, et al., Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration, BMC Med. 11 (2013) 220, doi:10.1186/1741-7015-11-220.
[6] F. Capozzi, A. Bordoni, Foodomics: a new comprehensive approach to food and nutrition, Genes Nutr. 8 (2013) 1–4, doi:10.1007/s12263-012-0310-x.
[7] J.G. Bundy, M.P. Davey, M.R. Viant, Environmental metabolomics: a critical review and future perspectives, Metabolomics 5 (2008) 3–21, doi:10.1007/s11306-008-0152-0.
[8] M.R. Viant, U. Sommer, Mass spectrometry based environmental metabolomics: a primer and review, Metabolomics 9 (2012) 144–158, doi:10.1007/s11306-012-0412-x.
[9] M. Adams, J. Kelley, J. Gocayne, M. Dubnick, M. Polymeropoulos, H. Xiao, et al., Complementary DNA sequencing: expressed sequence tags and human genome project, Science 252 (1991) 1651–1656, doi:10.1126/science.2047873.
[10] M.J. Fazzari, J.M. Greally, Epigenomics: beyond CpG islands, Nat. Rev. Genet. 5 (2004) 446–455, doi:10.1038/nrg1349.
[11] A. Abbott, Proteomics, transcriptomics: what's in a name?, Nature 402 (1999) 715–720, doi:10.1038/45354.
[12] N.L. Anderson, N.G. Anderson, Proteome and proteomics: new technologies, new concepts, and new words, Electrophoresis 19 (1998) 1853–1861, doi:10.1002/elps.1150191103.
[13] G. Winter, J.O. Krömer, Fluxomics – connecting 'omics analysis and phenotypes, Environ. Microbiol. 15 (2013) 1901–1916, doi:10.1111/1462-2920.12064.
[14] M. Cascante, S. Marin, Metabolomics and fluxomics approaches, Essays Biochem. 45 (2008) 67–81, doi:10.1042/BSE0450067.
[15] X. Han, R.W. Gross, Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics, J. Lipid Res. 44 (2003) 1071–1079, doi:10.1194/jlr.R300004-JLR200.
[16] J.E. Turnbull, R.A. Field, Emerging glycomics technologies, Nat. Chem. Biol. 3 (2007) 74–77, doi:10.1038/nchembio0207-74.
[17] M. Herrero, C. Simó, V. García-Cañas, E. Ibáñez, A. Cifuentes, Foodomics: MS-based strategies in modern food science and nutrition, Mass Spectrom. Rev. 31 (2012) 49–69, doi:10.1002/mas.20335.
[18] W. Zhang, F. Li, L. Nie, Integrating multiple "omics" analysis for microbial biology: application and methodologies, Microbiology 156 (2010) 287–301, doi:10.1099/mic.0.034793-0.
[19] A.K. Shanker, M. Djanaguiraman, B. Venkateswarlu, Chromium interactions in plants: current status and future strategies, Metallomics 1 (2009) 375–383, doi:10.1039/b904571f.
[20] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, Nat. Rev. Drug Discov. 1 (2002) 153–161, doi:10.1038/nrd728.
[21] B. Campos, N. Garcia-Reyero, C. Rivetti, L. Escalon, T. Habib, R. Tauler, et al., Identification of metabolic pathways in Daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors and 4-nonylphenol using transcriptomic and phenotypic responses, Environ. Sci. Technol. 47 (2013) 9434–9443, doi:10.1021/es4012299.
[22] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based plant metabolomics: where do we stand, where do we go?, Trends Biotechnol. 29 (2011) 267–275, doi:10.1016/j.tibtech.2011.02.001.
[23] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, A quantitative 1H NMR approach for evaluating the metabolic response of Saccharomyces cerevisiae to mild heat stress, Metabolomics 11 (2015) 1612–1625, doi:10.1007/s11306-015-0812-9.
[24] J.M. Halket, Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS, J. Exp. Bot. 56 (2004) 219–243, doi:10.1093/jxb/eri069.
[25] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass spectrometry-based metabolomics, Mass Spectrom. Rev. 26 (2007) 51–78, doi:10.1002/mas.20108.
[26] J.K. Nicholson, I.D. Wilson, High resolution proton magnetic resonance spectroscopy of biological fluids, Prog. Nucl. Magn. Reson. Spectrosc. 21 (1989) 449–501, doi:10.1016/0079-6565(89)80008-1.
[27] J.C. Lindon, E. Holmes, J.K. Nicholson, Peer reviewed: so what's the deal with metabonomics?, Anal. Chem. 75 (2003) 384 A–391 A, doi:10.1021/ac031386.
[28] R.J.M. Weber, A.D. Southam, U. Sommer, M.R. Viant, Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification, Anal. Chem. 83 (2011) 3737–3743, doi:10.1021/ac2001803.
[29] I.D. Wilson, R. Plumb, J. Granger, H. Major, R. Williams, E.M. Lenz, HPLC-MS-based methods for the study of metabonomics, J. Chromatogr. B. Analyt Technol Biomed Life Sci. 817 (2005) 67–76, doi:10.1016/j.jchromb.2004.07.045.
[30] I.D. Wilson, J.K. Nicholson, J. Castro-Perez, J.H. Granger, K.A. Johnson, B.W. Smith, et al., High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies, J. Proteome Res. 4 (2005) 591–598, doi:10.1021/pr049769r.
[31] P.J. Weaver, A.M.-F. Laures, J.-C. Wolff, Investigation of the advanced functionalities of a hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer, Rapid Commun. Mass Spectrom. 21 (2007) 2415–2421, doi:10.1002/rcm.3052.
[32] S.C. Brown, G. Kruppa, J.-L. Dasseux, Metabolomics applications of FT-ICR mass spectrometry, Mass Spectrom. Rev. 24 (2005) 223–231, doi:10.1002/mas.20011.
[33] A. Koulman, G. Woffendin, V.K. Narayana, H. Welchman, C. Crone, D.A. Volmer, High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectrometer, Rapid Commun. Mass Spectrom. 23 (2009) 1411–1418, doi:10.1002/rcm.4015.
[34] E. Rathahao-Paris, S. Alves, C. Junot, J.-C. Tabet, High resolution mass spectrometry for structural identification of metabolites in metabolomics, Metabolomics 12 (2015) 10, doi:10.1007/s11306-015-0882-8.
[35] A. Jiye, J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson, H. Antti, et al., Extraction and GC/MS analysis of the human blood plasma metabolome, Anal. Chem. 77 (2005) 8086–8094, doi:10.1021/ac051211v.
[36] S.G. Villas-Bôas, S. Mas, M. Akesson, J. Smedsgaard, J. Nielsen, Mass spectrometry in metabolome analysis, Mass Spectrom. Rev. 24 (2005) 613–646, doi:10.1002/mas.20032.
[37] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics 9 (2008) 504, doi:10.1186/1471-2105-9-504.
[38] H.K. Kim, R. Verpoorte, Sample preparation for plant metabolomics, Phytochem. Anal. 21 (2010) 4–13, doi:10.1002/pca.1188.
[39] A.H. Wu, R. Gerona, P. Armenian, D. French, M. Petrie, K.L. Lynch, Role of liquid chromatography–high-resolution mass spectrometry (LC-HR/MS) in clinical toxicology, Clin. Toxicol. 50 (2012) 733–742, doi:10.3109/15563650.2012.713108.
[40] J. Boccard, S. Rudaz, Harnessing the complexity of metabolomic data with chemometrics, J. Chemometrics 28 (2014) 1–9, doi:10.1002/cem.2567.
[41] M. Katajamaa, M. Orešič, Processing methods for differential analysis of LC/MS profile data, BMC Bioinformatics 6 (2005) 1, doi:10.1186/1471-2105-6-179.
[42] W. Lu, B.D. Bennett, J.D. Rabinowitz, Analytical strategies for LC-MS-based targeted metabolomics, J. Chromatogr. B. Analyt Technol Biomed Life Sci. 871 (2008) 236–242, doi:10.1016/j.jchromb.2008.04.031.
[43] R.C.H. De Vos, S. Moco, A. Lommen, J.J.B. Keurentjes, R.J. Bino, R.D. Hall, Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry, Nat. Protoc. 2 (2007) 778–791, doi:10.1038/nprot.2007.95.
[44] J.J. Dalluge, S. Smith, F. Sanchez-Riera, C. McGuire, R. Hobson, Potential of fermentation profiling via rapid measurement of amino acid metabolism by liquid chromatography–tandem mass spectrometry, J. Chromatogr. A 1043 (2004) 3–7, doi:10.1016/j.chroma.2004.02.010.
[45] Y. Sawada, K. Akiyama, A. Sakata, A. Kuwahara, H. Otsuki, T. Sakurai, et al., Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants, Plant Cell Physiol. 50 (2009) 37–47, doi:10.1093/pcp/pcn183.
[46] B. Guo, B. Chen, A. Liu, W. Zhu, S. Yao, Liquid chromatography-mass spectrometric multiple reaction monitoring-based strategies for expanding targeted profiling towards quantitative metabolomics, Curr. Drug Metab. 13 (2012) 1226–1243 <http://www.ncbi.nlm.nih.gov/pubmed/22519369> (accessed 20.01.16).
[47] H. Gu, P. Zhang, J. Zhu, D. Raftery, Globally Optimized Targeted Mass Spectrometry (GOT-MS): reliable metabolomics analysis with broad coverage, Anal. Chem. (2015) doi:10.1021/acs.analchem.5b03812.
[48] S. Wang, H. Tu, J. Wan, W. Chen, X. Liu, J. Luo, et al., Spatio-temporal distribution and natural variation of metabolites in citrus fruits, Food Chem. 199 (2016) 8–17, doi:10.1016/j.foodchem.2015.11.113.
[49] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, Algorithms and tools for the preprocessing of LC–MS metabolomics data, Chemometr. Intell. Lab. 108 (2011) 23–32, doi:10.1016/j.chemolab.2011.03.010.

[50] M. de Raad, C.R. Fischer, T.R. Northen, High-throughput platforms for metabolomics, Curr. Opin. Chem. Biol. 30 (2015) 7–13, doi:10.1016/j.cbpa.2015.10.012.

[51] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, et al., Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34, doi:10.1016/j.aca.2016.02.001.

[52] T. Cajka, O. Fiehn, Towards merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics, Anal. Chem. 88 (2015) 524–545, doi:10.1021/acs.analchem.5b04491.

[53] O.I. Savolainen, A.-S. Sandberg, A.B. Ross, A simultaneous metabolic profiling and quantitative multimetabolite metabolomic method for human plasma using gas-chromatography tandem mass spectrometry. <http://pubs.acs.org/doi/10.1021/acs.jproteome.5b00790>, 2015 (accessed 20.01.16).

[54] C.D. Broeckling, I.R. Reddy, A.L. Duran, X. Zhao, L.W. Sumner, MET-IDEA: data extraction tool for mass spectrometry-based metabolomics, Anal. Chem. 78 (2006) 4334–4341, doi:10.1021/ac0521596.

[55] H. Tsugawa, M. Arita, M. Kanazawa, A. Ogiwara, T. Bamba, E. Fukusaki, MRMPROBS: a data assessment and metabolite identification tool for large-scale multiple reaction monitoring based widely targeted metabolomics, Anal. Chem. 85 (2013) 5191–5199, doi:10.1021/ac400515s.

[56] H. Tsugawa, M. Kanazawa, A. Ogiwara, M. Arita, MRMPROBS suite for metabolomics using large-scale MRM assays, Bioinformatics 30 (2014) 2379–2380, doi:10.1093/bioinformatics/btu203.

[57] J.W.H. Wong, H.J. Abuhusain, K.L. McDonald, A.S. Don, MMSAT: automated quantification of metabolites in selected reaction monitoring experiments, Anal. Chem. 84 (2012) 470–474, doi:10.1021/ac2026578.

[58] D.B. Martin, T. Holzman, D. May, A. Peterson, A. Eastham, J. Eng, et al., MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments, Mol. Cell. Proteomics 7 (2008) 2270–2278, doi:10.1074/mcp.M700504-MCP200.

[59] P. Wenig, J. Odermatt, OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data, BMC Bioinformatics 11 (2010) 405, doi:10.1186/1471-2105-11-405.

[60] A. Garanto, N.A. Mandal, M. Egido-Gabás, G. Marfany, G. Fabriàs, R.E. Anderson, et al., Specific sphingolipid content decrease in Cerkl knockdown mouse retinas, Exp. Eye Res. 110 (2013) 96–106, doi:10.1016/j.exer.2013.03.003.

[61] E. Gorrochategui, J. Casas, E. Pérez-Albaladejo, O. Jáuregui, C. Porte, S. Lacorte, Characterization of complex lipid mixtures in contaminant exposed JEG-3 cells using liquid chromatography and high-resolution mass spectrometry, Environ. Sci. Pollut. Res. Int. 21 (2014) 11907–11916, doi:10.1007/s11356-014-3172-5.

[62] Y. Wang, M. Gu, The concept of spectral accuracy for MS, Anal. Chem. 82 (2010) 7055–7062, doi:10.1021/ac100888b.

[63] J.C.L. Erve, M. Gu, Y. Wang, W. DeMaio, R.E. Talaat, Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination, J. Am. Soc. Mass Spectrom. 20 (2009) 2058–2069, doi:10.1016/j.jasms.2009.07.014.

[64] A. Amorisco, V. Locaputo, C. Pastore, G. Mascolo, Identification of low molecular weight organic acids by ion chromatography/hybrid quadrupole time-of-flight mass spectrometry during Uniblu-A ozonation, Rapid Commun. Mass Spectrom. 27 (2013) 187–199, doi:10.1002/rcm.6429.

[65] E. Dudley, M. Yousef, Y. Wang, W.J. Griffiths, Targeted metabolomics and mass spectrometry, Adv. Protein Chem. Struct. Biol. 80 (2010) 45–83, doi:10.1016/B978-0-12-381264-3.00002-3.

[66] M. Vinaixa, E.L. Schymanski, S. Neumann, M. Navarro, R.M. Salek, O. Yanes, Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects, TrAC Trends Anal. Chem. 78 (2015) 23–35, doi:10.1016/j.trac.2015.09.005.

[67] D.J. Creek, W.B. Dunn, O. Fiehn, J.L. Griffin, R.D. Hall, Z. Lei, et al., Metabolite identification: are you sure? And how do your peers gauge your confidence?, Metabolomics 10 (2014) 350–353, doi:10.1007/s11306-014-0656-8.

[68] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, et al., Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), Metabolomics 3 (2007) 211–221, doi:10.1007/s11306-007-0082-2.

[69] European Communities (EC), Implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, vol 2002/657/EC, 2002.

[70] M. Gergov, I. Ojanperä, E. Vuori, Simultaneous screening for 238 drugs in blood by liquid chromatography–ionspray tandem mass spectrometry with multiple-reaction monitoring, J. Chromatogr. B 795 (2003) 41–53, doi:10.1016/S1570-0232(03)00498-7.

[71] S.U. Bajad, W. Lu, E.H. Kimball, J. Yuan, C. Peterson, J.D. Rabinowitz, Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry, J. Chromatogr. A 1125 (2006) 76–88, doi:10.1016/j.chroma.2006.05.019.

[72] B.D. Bennett, E.H. Kimball, M. Gao, R. Osterhout, S.J. Van Dien, J.D. Rabinowitz, Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*, Nat. Chem. Biol. 5 (2009) 593–599, doi:10.1038/nchembio.186.

[73] J.M. Buescher, S. Moco, U. Sauer, N. Zamboni, Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites, Anal. Chem. 82 (2010) 4403–4412, doi:10.1021/ac100101d.

[74] M.S. Sabatine, E. Liu, D.A. Morrow, E. Heller, R. McCarroll, R. Wiegand, et al., Metabolomic identification of novel biomarkers of myocardial ischemia, Circulation 112 (2005) 3868–3875, doi:10.1161/CIRCULATIONAHA.105.569137.

[75] T.J. Wang, M.G. Larson, R.S. Vasan, S. Cheng, E.P. Rhee, E. McCabe, et al., Metabolite profiles and the risk of developing diabetes, Nat. Med. 17 (2011) 448–453, doi:10.1038/nm.2307.

[76] O.N. Jensen, A. Podtelejnikov, M. Mann, Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps, Rapid Commun. Mass Spectrom. 10 (1996) 1371–1378, doi:10.1002/(SICI)1097-0231(199608)10:11<1371::AID-RCM682>3.0.CO;2-5.

[77] E. Moskovets, H.-S. Chen, A. Pashkova, T. Rejtar, V. Andreev, B.L. Karger, Closely spaced external standard: a universal method of achieving 5 ppm mass accuracy over the entire MALDI plate in axial matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 17 (2003) 2177–2187, doi:10.1002/rcm.1158.

[78] A.M. Starrett, G.C. DiDonato, High resolution accurate mass measurement of product ions formed in an electrospray source on a sector instrument, Rapid Commun. Mass Spectrom. 7 (1993) 12–15, doi:10.1002/rcm.1290070104.

[79] M. Katajamaa, M. Oresic, Data processing for mass spectrometry-based metabolomics, J. Chromatogr. A 1158 (2007) 318–328, doi:10.1016/j.chroma.2007.04.021.

[80] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, et al., Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (2011) 1060–1083, doi:10.1038/nprot.2011.335.

[81] M.R. Mashego, L. Wu, J.C. Van Dam, C. Ras, J.L. Vinke, W.A. Van Winden, et al., MIRACLE: mass isotopomer ratio analysis of U-13C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites, Biotechnol. Bioeng. 85 (2004) 620–628, doi:10.1002/bit.10907.

[82] L. Wu, M.R. Mashego, J.C. van Dam, A.M. Proell, J.L. Vinke, C. Ras, et al., Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards, Anal. Biochem. 336 (2005) 164–171, doi:10.1016/j.ab.2004.09.001.

[83] W. Lu, E. Kimball, J.D. Rabinowitz, A high-performance liquid chromatography-tandem mass spectrometry method for quantitation of nitrogen-containing intracellular metabolites, J. Am. Soc. Mass Spectrom. 17 (2006) 37–50, doi:10.1016/j.jasms.2005.09.001.

[84] J. Yuan, W.U. Fowler, E. Kimball, W. Lu, J.D. Rabinowitz, Kinetic flux profiling of nitrogen assimilation in *Escherichia coli*, Nat. Chem. Biol. 2 (2006) 529–530, doi:10.1038/nchembio816.

[85] W. Lu, Y.K. Kwon, J.D. Rabinowitz, Isotope ratio-based profiling of microbial folates, J. Am. Soc. Mass Spectrom. 18 (2007) 898–909, doi:10.1016/j.jasms.2007.01.017.

[86] J.D. Rabinowitz, E. Kimball, Acidic acetonitrile for cellular metabolome extraction from *Escherichia coli*, Anal. Chem. 79 (2007) 6167–6173, doi:10.1021/ac070470c.

[87] S. Arrivault, M. Guenther, S.C. Fry, M.M.F.F. Fuenfgeld, D. Veyel, T. Mettler-Altmann, et al., Synthesis and use of stable-isotope-labeled internal standards for quantification of phosphorylated metabolites by LC-MS/MS, Anal. Chem. 87 (2015) 6896–6904, doi:10.1021/acs.analchem.5b01387.

[88] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, et al., Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation, Anal. Chem. 78 (2006) 567–574, doi:10.1021/ac051495j.

[89] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, M. Oresic, Normalization method for metabolomics data using optimal selection of multiple internal standards, BMC Bioinformatics 8 (2007) 93, doi:10.1186/1471-2105-8-93.

[90] A. Bajoub, T. Pacchiarotta, E. Hurtado-Fernández, L. Olmo-García, R. García-Villalba, A. Fernández-Gutiérrez, et al., Comparing two metabolic profiling approaches (liquid chromatography and gas chromatography coupled to mass spectrometry) for extra-virgin olive oil phenolic compounds analysis: a botanical classification perspective, J. Chromatogr. A 1428 (2016) 267–279, doi:10.1016/j.chroma.2015.10.059.

[91] E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Metabolic profiling of Daphnia magna exposed to environmental stressors by GC–MS and chemometric tools, Metabolomics 12 (2016) 86, doi:10.1007/s11306-016-1021-x.

[92] S. Orchard, L. Montechi-Palazzi, E.W. Deutsch, P.-A. Binz, A.R. Jones, N. Paton, et al., Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23–25, 2007 Ecole Nationale Supérieure (ENS), Lyon, France, Proteomics 7 (2007) 3436–3440, doi:10.1002/pmic.200700658.

[93] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, et al., mzML – a community standard for mass spectrometry data, Mol. Cell. Proteomics 10 (2011) R110.000133, doi:10.1074/mcp.R110.000133.

[94] ASTM E1947–98(2014), Standard specification for analytical data interchange protocol for chromatographic data. <http://www.astm.org/Standards/E1947.htm>, 2016 (accessed 19.01.16) n.d.

[95] R.S. McDonald, P.A. Wilks, JCAMP-DX: a standard form for exchange of infrared spectra in computer readable form, Appl. Spectrosc. 42 (1988) 151–162, doi:10.1366/0003702884428734.

[96] E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using chemometric tools, Protoc. Exch. (2015) doi:10.1038/protex.2015.102.

[97] P.G.A. Pedrioli, Trans-proteomic pipeline: a pipeline for proteomic analysis, Methods Mol. Biol. 604 (2010) 213–238, doi:10.1007/978-1-60761-444-9_15.

[98] D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, ProteoWizard: open source software for rapid proteomics tools development, Bioinformatics 24 (2008) 2534–2536, doi:10.1093/bioinformatics/btn323.

[99] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787, doi:10.1021/ac051437y.
[100] Y. Tikunov, A. Lommen, C.H.R. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, et al., A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles, Plant Physiol. 139 (2005) 1125–1137, doi:10.1104/pp.105.068130.
[101] M. Katajamaa, J. Miettinen, M. Oresic, MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data, Bioinformatics 22 (2006) 634–636, doi:10.1093/bioinformatics/btk039.
[102] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics 11 (2010) 395, doi:10.1186/1471-2105-11-395.
[103] A. Cuadros-Inostroza, C. Caldana, H. Redestig, M. Kusano, J. Lisec, H. Peña-Cortés, et al., TargetSearch – a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data, BMC Bioinformatics 10 (2009) 428, doi:10.1186/1471-2105-10-428.
[104] L. Gatto, K.S. Lilley, MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation, Bioinformatics 28 (2011) 288–289, doi:10.1093/bioinformatics/btr645.
[105] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, et al., Second-order peak detection for multicomponent high-resolution LC/MS data, Anal. Chem. 78 (2006) 975–983, doi:10.1021/ac050980b.
[106] G.G. Siano, I.S. Pérez, M.D.G. García, M.M. Galera, H.C. Goicoechea, Multivariate curve resolution modeling of liquid chromatography-mass spectrometry data in a comparative study of the different endogenous metabolites behavior in two tomato cultivars treated with carbofuran pesticide, Talanta 85 (2011) 264–275, doi:10.1016/j.talanta.2011.03.064.
[107] M. Farrés, B. Piña, R. Tauler, Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS, Metabolomics 11 (2014) 210–224, doi:10.1007/s11306-014-0689-z.
[108] E. Gorrochategui, J. Casas, C. Porte, S. Lacorte, R. Tauler, Chemometric strategy for untargeted lipidomics: biomarker detection and identification in stressed human placental cells, Anal. Chim. Acta 854 (2015) 20–33, doi:10.1016/j.aca.2014.11.010.
[109] W.B. Dunn, I.D. Wilson, A.W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, Bioanalysis 4 (2012) 2249–2264, doi:10.4155/bio.12.204.
[110] F.M. van der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping, J. Proteome Res. 8 (2009) 5132–5141, doi:10.1021/pr900499r.
[111] E.M. Qannari, I. Wakeling, P. Courcoux, H.J. MacFie, Defining the underlying sensory dimensions, Food Qual. Prefer. 11 (2000) 151–154, doi:10.1016/S0950-3293(99)00069-5.
[112] E. Dubin, M. Spiteri, A.-S. Dumas, J. Ginet, M. Lees, D.N. Rutledge, Common components and specific weights analysis: a tool for metabolomics data pre-processing, Chemom. Intell. Lab. 150 (2015) 41–50, doi:10.1016/j.chemolab.2015.11.005.
[113] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Metabolite fingerprinting: detecting biological features by independent component analysis, Bioinformatics 20 (2004) 2447–2454, doi:10.1093/bioinformatics/bth270.
[114] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, et al., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards, Anal. Chem. 75 (2003) 4818–4826, doi:10.1021/ac026468x.
[115] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics, Anal. Chem. 78 (2006) 4281–4290, doi:10.1021/ac051632c.
[116] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 142, doi:10.1186/1471-2164-7-142.
[117] O.M. Khalheim, Scaling of analytical data, Anal. Chim. Acta 177 (1985) 71–79, doi:10.1016/S0003-2670(00)82939-6.
[118] E.M. Kasprzak, K.E. Lewis, Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method, Struct. Multidiscip. Optim. 22 (2014) 208–218, doi:10.1007/s001580100138.
[119] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Anal. Chem. 77 (2005) 6729–6736, doi:10.1021/ac051080y.
[120] H.C. Keun, T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert, E. Holmes, et al., Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, Anal. Chim. Acta 490 (2003) 265–276, doi:10.1016/S0003-2670(03)00094-1.
[121] M.R. Keenan, P.G. Kotula, Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis, Appl. Surf. Sci. 231–232 (2004) 240–244, doi:10.1016/j.apsusc.2004.03.025.
[122] M.R. Keenan, P.G. Kotula, Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images, Surf. Interface Anal. 36 (2004) 203–212, doi:10.1002/sia.1657.
[123] O.M. Kvalheim, F. Brakstad, Y. Liang, Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise, Anal. Chem. 66 (1994) 43–51, doi:10.1021/ac00073a010.
[124] J. Xia, N. Psychogios, N. Young, D.S. Wishart, MetaboAnalyst: a web server for metabolomic data analysis and interpretation, Nucleic Acids Res. 37 (2009) W652–W660, doi:10.1093/nar/gkp356.
[125] J. Xia, R. Mandal, I.V. Sinelnikov, D. Broadhurst, D.S. Wishart, MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis, Nucleic Acids Res. 40 (2012) W127–W133, doi:10.1093/nar/gks374.
[126] Y. Wu, L. Li, Sample normalization methods in quantitative metabolomics, J. Chromatogr. A 1430 (2016) 80–95, doi:10.1016/j.chroma.2015.12.007.
[127] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. 30 (1995) 133–146, doi:10.1016/0169-7439(95)00047-X.
[128] I. Sánchez Pérez, M.J. Culzoni, G.G. Siano, M.D. Gil García, H.C. Goicoechea, M. Martínez Galera, Detection of unintended stress effects based on a metabonomic study in tomato fruits after treatment with carbofuran pesticide. Capabilities of MCR-ALS applied to LC-MS three-way data arrays, Anal. Chem. 81 (2009) 8335–8346, doi:10.1021/ac901119h.
[129] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36, doi:10.1016/j.aca.2012.12.028.
[130] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964, doi:10.1039/c4ay00571f.
[131] E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler, Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling, Electrophoresis 36 (2015) 2324–2335, doi:10.1002/elps.201500027.
[132] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, Anal. Bioanal. Chem. 407 (2015) 8835–8847, doi:10.1007/s00216-015-9042-2.
[133] C. Bedia, N. Dalmau, J. Jaumot, R. Tauler, Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors, Environ. Res. 140 (2015) 18–31, doi:10.1016/j.envres.2015.03.014.
[134] K. Podwojski, A. Fritsch, D.C. Chamrad, W. Paul, B. Sitek, K. Stuhler, et al., Retention time alignment algorithms for LC/MS data must consider non-linear shifts, Bioinformatics 25 (2009) 758–764, doi:10.1093/bioinformatics/btp052.
[135] J.T. Prince, E.M. Marcotte, Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping, Anal. Chem. 78 (2006) 6140–6152, doi:10.1021/ac0605344.
[136] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395, doi:10.1145/358669.358692.
[137] T.G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, L.M.C. Buydens, Warping methods for spectroscopic and chromatographic signal alignment: a tutorial, Anal. Chim. Acta 781 (2013) 14–32, doi:10.1016/j.aca.2013.03.048.
[138] X. Shao, C. Pang, Q. Su, A novel method to calculate the approximate derivative photoacoustic spectrum using continuous wavelet transform, Fresenius. J. Anal. Chem. 367 (2000) 525–529, doi:10.1007/s002160000404.
[139] R. Bro, PARAFAC tutorial and applications, Chemom. Intell. Lab. 38 (1997) 149–171, doi:10.1016/S0169-7439(97)00032-4.
[140] S.A. Bortolato, J.A. Arancibia, G.M. Escandar, A.C. Olivieri, Time-alignment of bidimensional chromatograms in the presence of uncalibrated interferences using parallel factor analysis, Chemom. Intell. Lab. 101 (2010) 30–37, doi:10.1016/j.chemolab.2009.12.001.
[141] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, PARAFAC2 – Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemometrics 13 (1999) 275–294 <http://www.scopus.com/inward/record.url?eid=2-s2.0-0001718376&partnerID=tZOtx3y1>.
[142] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 – Part II. Modeling chromatographic data with retention time shifts, J. Chemometrics 13 (1999) 295–309 <http://www.scopus.com/inward/record.url?eid=2-s2.0-0000095845&partnerID=tZOtx3y1>.
[143] B. Khakimov, J.M. Amigo, S. Bak, S.B. Engelsen, Plant metabolomics: resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods, J. Chromatogr. A 1266 (2012) 84–94, doi:10.1016/j.chroma.2012.10.023.
[144] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV-vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Anal. Chim. Acta 842 (2014) 11–19, doi:10.1016/j.aca.2014.07.007.
[145] P. Comon, Independent component analysis, A new concept?, Signal Process. 36 (1994) 287–314, doi:10.1016/0165-1684(94)90029-9.
[146] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry?, TrAC Trends Anal. Chem. 31 (2012) 134–143, doi:10.1016/j.trac.2011.07.010.
[147] X. Zhang, R. Tauler, Measuring and comparing the resolution performance and the extent of rotation ambiguities of some bilinear modeling methods, Chemom. Intell. Lab. 147 (2015) 47–57, doi:10.1016/j.chemolab.2015.08.005.
[148] P.A. Højen-Sørensen, O. Winther, L.K. Hansen, Mean-field approaches to independent component analysis, Neural Comput. 14 (2002) 889–918, doi:10.1162/089976602317319009.
[149] Y. Liu, K. Smirnov, M. Lucio, R.D. Gougeon, H. Alexandre, P. Schmitt-Kopplin, MetICA: independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics, BMC Bioinformatics 17 (2016) 114, doi:10.1186/s12859-016-0970-4.

[150] D.I. Broadhurst, D.B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics 2 (2006) 171–196, doi:10.1007/s11306-006-0037-z.
[151] R. Rousseau, B. Govaerts, M. Verleysen, B. Boulanger, Comparison of some chemometric tools for metabonomics biomarker identification, Chemom. Intell. Lab. 91 (2008) 54–66, doi:10.1016/j.chemolab.2007.06.008.
[152] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics 9 (2013) 280–299, doi:10.1007/s11306-012-0482-9.
[153] M.L. Head, L. Holman, R. Lanfear, A.T. Kahn, M.D. Jennions, The extent and consequences of P-hacking in science, PLoS Biol. 13 (2015) doi:10.1371/journal.pbio.1002106 e1002106.
[154] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. 2 (1987) 37–52, doi:10.1016/0169-7439(87)80084-9.
[155] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometrics 17 (2003) 166–173, doi:10.1002/cem.785.
[156] S. Le Cessie, J.C. Van Houwelingen, Logistic regression for correlated binary data. <http://www.jstor.org/stable/2986114?seq=1#page_scan_tab_contents>, 2016 (accessed 19.01.16) n.d.
[157] C.S.L.J. Breiman, R. Freidman, R. Olsen, Classification and regression trees, Belmont, CA, 1984.
[158] T. Rajalahti, R. Arneberg, F.S. Berven, K.-M. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, Chemom. Intell. Lab. 95 (2009) 35–48, doi:10.1016/j.chemolab.2008.08.004.
[159] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, Anal. Chem. 81 (2009) 2581–2590, doi:10.1021/ac802514y.
[160] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. 58 (2001) 109–130, doi:10.1016/S0169-7439(01)00155-1.
[161] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048, doi:10.1093/bioinformatics/bti476.
[162] D.J. Vis, J.A. Westerhuis, A.K. Smilde, J. van der Greef, Statistical validation of megavariate effects in ASCA, BMC Bioinformatics 8 (2007) 322, doi:10.1186/1471-2105-8-322.
[163] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. 78 (2005) 103–112, doi:10.1016/j.chemolab.2004.12.011.
[164] R. Gosselin, D. Rodrigue, C. Duchesne, A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemom. Intell. Lab. 100 (2010) 12–21, doi:10.1016/j.chemolab.2009.09.005.
[165] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemometrics 29 (2015) 528–536, doi:10.1002/cem.2736.
[166] A. Checa, C. Bedia, J. Jaumot, Lipidomic data analysis: tutorial, practical guidelines and applications, Anal. Chim. Acta 885 (2015) 1–16, doi:10.1016/j.aca.2015.02.068.
[167] C.M. Andersen, R. Bro, Variable selection in regression-a tutorial, J. Chemometrics 24 (2010) 728–737, doi:10.1002/cem.1360.
[168] K.J. Mielke, P.W. Berry Jr., Permutation Methods: A Distance Function Approach, Springer, New York, 2001.
[169] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Natl. Cancer Inst. 95 (2003) 14–18 <http://www.scopus.com/inward/record.url?eid=2-s2.0-0037245343&partnerID=tZOtx3y1>.
[170] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 6562–6566, doi:10.1073/pnas.102102699.
[171] S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts, C.G. de Koster, Assessing the statistical validity of proteomics based biomarkers, Anal. Chim. Acta 592 (2007) 210–217, doi:10.1016/j.aca.2007.04.043.
[172] E. Bradley, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, 1993.
[173] F. Westad, F. Marini, Validation of chemometric models – a tutorial, Anal. Chim. Acta 893 (2015) 14–24, doi:10.1016/j.aca.2015.06.056.
[174] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 97–106, doi:10.1002/wics.51.
[175] L. Xu, Q.-S. Xu, M. Yang, H.-Z. Zhang, C.-B. Cai, J.-H. Jiang, et al., On estimating model complexity and prediction errors in multivariate calibration: generalized resampling by random sample weighting (RSW), J. Chemometrics 25 (2011) 51–58, doi:10.1002/cem.1323.
[176] N.L. Afanador, T.N. Tran, L.M.C. Buydens, Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression, Anal. Chim. Acta 768 (2013) 49–56, doi:10.1016/j.aca.2013.01.004.
[177] W.B. Dunn, A. Erban, R.J.M. Weber, D.J. Creek, M. Brown, R. Breitling, et al., Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics, Metabolomics 9 (2012) 44–66, doi:10.1007/s11306-012-0434-4.
[178] P. Li, H.A. Senthilkumar, S.-B. Wu, B. Liu, Z. Guo, J.E. Fata, et al., Comparative UPLC-QTOF-MS-based metabolomics and bioactivities analyses of Garcinia oblongifolia, J. Chromatogr. B 1011 (2016) 179–195, doi:10.1016/j.jchromb.2015.12.061.
[179] A. Nordström, G. O'Maille, C. Qin, G. Siuzdak, Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum, Anal. Chem. 78 (2006) 3289–3295, doi:10.1021/ac060245f.
[180] Y. Konishi, T. Kiyota, C. Draghici, J.-M. Gao, F. Yeboah, S. Acoca, et al., Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products, Anal. Chem. 79 (2007) 1187–1197, doi:10.1021/ac061391o.
[181] M. Wrona, T. Mauriala, K.P. Bateman, R.J. Mortishire-Smith, D. O'Connor, "All-in-one" analysis for metabolite identification using liquid chromatography/hybrid quadrupole time-of-flight mass spectrometry with collision energy switching, Rapid Commun. Mass Spectrom. 19 (2005) 2597–2602, doi:10.1002/rcm.2101.
[182] Y.-Y. Zhao, R.-C. Lin, UPLC-MS(E) application in disease biomarker discovery: the discoveries in proteomics to metabolomics, Chem. Biol. Interact. 215 (2014) 7–16, doi:10.1016/j.cbi.2014.02.014.
[183] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res. 40 (2012) D109–D114, doi:10.1093/nar/gkr988.
[184] P.D. Karp, Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, Nucleic Acids Res. 33 (2005) 6083–6089, doi:10.1093/nar/gki892.
[185] R. Caspi, MetaCyc: a multiorganism database of metabolic pathways and enzymes, Nucleic Acids Res. 34 (2006) D511–D516, doi:10.1093/nar/gkj128.
[186] A.R. Pico, T. Kelder, M.P. van Iersel, K. Hanspers, B.R. Conklin, C. Evelo, WikiPathways: pathway editing for the people, PLoS Biol. 6 (2008) e184, doi:10.1371/journal.pbio.0060184.
[187] T. Kelder, M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, et al., WikiPathways: building research communities on biological pathways, Nucleic Acids Res. 40 (2012) D1301–D1307, doi:10.1093/nar/gkr1074.
[188] Y. Chu, H. Jiang, J. Ju, Y. Li, L. Gong, X. Wang, et al., A metabolomic study using HPLC-TOF/MS coupled with ingenuity pathway analysis: intervention effects of Rhizoma Alismatis on spontaneous hypertensive rats, J. Pharm. Biomed. Anal. 117 (2016) 446–452, doi:10.1016/j.jpba.2015.09.026.
[189] A. Perl, R. Hanczko, Z.-W. Lai, Z. Oaks, R. Kelly, R. Borsuk, et al., Comprehensive metabolome analyses reveal N-acetylcysteine-responsive accumulation of kynurenine in systemic lupus erythematosus: implications for activation of the mechanistic target of rapamycin, Metabolomics 11 (2015) 1157–1174, doi:10.1007/s11306-015-0772-0.
[190] J.B. Coble, C.G. Fraga, Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery, J. Chromatogr. A 1358 (2014) 155–164, doi:10.1016/j.chroma.2014.06.100.
[191] O. Alter, P.O. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 3351–3356, doi:10.1073/pnas.0530258100.
[192] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, J. Trygg, Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data, Plant J. 52 (2007) 1181–1191, doi:10.1111/j.1365-313X.2007.03293.x.
[193] T. Löfstedt, J. Trygg, OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemometrics 25 (2011) 441–455, doi:10.1002/cem.1388.
[194] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (2013) 822–833, doi:10.3758/s13428-012-0295-9.
[195] J. Kuligowski, D. Pérez-Guaita, Á. Sánchez-Illana, Z. León-González, M. de la Guardia, M. Vento, et al., Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE), Analyst 140 (2015) 4521–4529, doi:10.1039/c5an00706b.
[196] E. Acar, R. Bro, A.K. Smilde, Data fusion in metabolomics using coupled matrix and tensor factorizations, Proc. IEEE 103 (2015) 1602–1620, doi:10.1109/JPROC.2015.2438719.
[197] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemometrics 16 (2002) 119–128, doi:10.1002/cem.695.
[198] S. El Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, H.-W. Uh, Evaluation of O2PLS in Omics data integration, BMC Bioinformatics 17 (Suppl. 2) (2016) 11, doi:10.1186/s12859-015-0854-z.
[199] T. Löfstedt, M. Hanafi, G. Mazerolles, J. Trygg, OnPLS path modelling, Chemom. Intell. Lab. 118 (2012) 139–149, doi:10.1016/j.chemolab.2012.08.009.
[200] V. Srivastava, O. Obudulu, J. Bygdell, T. Löfstedt, P. Rydén, R. Nilsson, et al., OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI-superoxide dismutase Populus plants, BMC Genomics 14 (2013) 893, doi:10.1186/1471-2164-14-893.
[201] L. Blanchet, A. Smolinska, Data fusion in metabolomics and proteomics for biomarker discovery, Methods Mol. Biol. 1362 (2016) 209–223, doi:10.1007/978-1-4939-3106-4_14.

2.2. SCIENTIFIC RESEARCH

In order to provide a data analysis methodology for metabolomic LC-MS data sets that can be implemented and properly used by researchers, a detailed description of all the steps involved in the process has been supplied in this Thesis. Such description has been included in the scientific article II, elaborated as a protocol, entitled "*A protocol for LC-MS metabolomics data processing using chemometric tools*" (Section 2.2.1). In this protocol, the functions to perform data compression are provided together with an example data set used for the illustration of the developed methodology. In a further step, the basics and fundamentals of the previously developed methodology are for the first time presented and described in the scientific article III entitled "*ROIMCR: a powerful data analysis strategy for LC-MS metabolomic data sets*" (Section 2.2.2).

2.2.1. SCIENTIFIC ARTICLE II

A protocol for LC-MS metabolomic data processin using chemometric tools

E. Gorrochategui, J. Jaumot, R. Tauler

Nature Protocol Exchange (2015) doi:10.1038/protex.2015.102

PROTOCOL EXCHANGE | *COMMUNITY CONTRIBUTED*

A protocol for LC-MS metabolomic data processing using chemometric tools

Eva Gorrochategui, Joaquim Jaumot & Romà Tauler

CHEMAGEB -Chemometrics and Omics Group-, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Spain

Abstract

Liquid chromatography- mass spectrometry (LC-MS) is a powerful methodology for metabolomics. However, **LC-MS data processing** comes out as the "bottleneck" of omic sciences due to its complexity. The present protocol, easy to execute in MATLAB environment, covers all data analysis steps (conversion and import, compression and processing) of LC-MS data sets and it is specifically designed for users with limited background in **chemometric and data analysis tools**. Data conversion and import are described for most important LC-MS manufacturers (*i.e.*, *Waters*, *Thermo Fischer*, *Agilent*, *AB Sciex* and *Bruker*), data compression consists on the search of "regions of interest" (**ROI**) and data processing is based on the use of Multivariate Curve Resolution-Alternating Least Squares (**MCR-ALS**), a powerful chemometric tool that allows chromatographic resolution. Results are rapidly achieved (usually < 15 min per sample), and they are easy to interpret and evaluate both in terms of chemistry and biology.

Subject terms: Computational biology Lipidomics Metabolomics

Keywords: metabolomics LC-MS data processing ROI chemometrics MCR-ALS

Introduction

Metabolomics is a field that aims at the study of the abundance and/or structural characterization of a large range of metabolites of organisms that have suffered unknown alterations due to exposure to environmental stressors.

Several analytical methods have been developed to perform metabolomic studies. Among them, mass spectrometry methods, coupled to chromatographic techniques have evolved into a novel and powerful technology due to their ability for multiparallel analysis of low molecular weight compounds in biological systems. Regarding chromatographic techniques, liquid chromatography (LC) is nowadays preferred to gas chromatography (GC), since the latter is restricted to volatile compounds, often requiring chemical derivatisation.

Data sets obtained with LC-MS technology contain large amount of information. Therefore, data processing is necessary to detect variations among omic profiles. However, one of the first steps required previous to data processing is the reduction of the dimensions of the original data sets, *i.e.*, data compression. Recently, a novel data compression method has been introduced in the *centWave*

algorithm of *XCMS* software[1]. This method is based on the concept of considering analytes as a region of data points with a high density ranked by a specific "data void", first presented by Stolt et al.[2]. These regions where analytes are found are called *regions of interest* (*ROI*). This data compression method appears as a better alternative to the classical binning procedure[1] since no loss of spectral accuracy is derived from a ROI search whereas a loss of resolution occurs after binning, which performs a compression in the *m/z*-mode dimension.

Several feature detection packages for omic LC-MS data have been developed in the last years (*e.g.*, *MarkerLynx* (*Waters*), *MetAlign*[3], *XCMS*[4] and *Mzmine*[5]). However, a powerful alternative to these packages is the use of chemometric tools[6],[7]. In fact, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[8] methods can properly resolve the profiling problem in omic data sets without the necessity of previous chromatographic alignment or shaping, which are required in most of the existent feature detection packages and represent the highest source of error.

In the present study we provide a detailed protocol and MATLAB functions (see Supplementary MATLAB functions) for LC-MS omic data analysis (including data conversion, import, pre-processing (*i.e.*, data compression) and processing steps). The distinct data analysis steps together with a brief description of the functions hereby used are provided in the **PROCEDURE** section.

Target audience and level of expertise needed to implement the protocol

The present protocol targets scientists who are using LC-MS techniques in metabolomic studies and want to analyze their own data but are not specialized in data analysis tools (including chemometric tools). In addition, this protocol is also valid for scientists using other mass spectrometry techniques such as CE-MS or mass spectrometry imaging. The required minimum skill level of users is low: only a basic understanding of what kind of information an LC-MS chromatogram provides is necessary. However, skilled chemometricians will also take advantage of the streamlined workflow.

Experimental Design

Biological samples

The selected samples for the illustration of the different steps of the protocol were the extracted lipids of a human placental choriocarcinoma (JEG-3) cell line, obtained from American Type Culture Collection (ATCC HTB-36), after exposure to tributyltin (TBT) or to the carrier solvent (DMSO) in the case of vehicle controls. Data from lipids of exposed and non-exposed culture cells were acquired using an Acquity UHPLC system (Waters, USA) connected to a Time of Flight (LCT Premier XE) detector under positive electrospray ionization (ESI (+)).

However, the protocol described in the present study is versatile and well suited to different kinds of chromatographic (*e.g.*, UHPLC, HPLC) coupled to mass spectrometric (*e.g.*, TOF, Orbitrap) data of diverse target molecules (*e.g.*, metabolites (including lipids)) coming from a big range of sample types.

Equipment

Hardware

• Standard-equipped PC or Mac with minimum system requirements to run the software (see below) and enough free disk space for saving the results. For optimal viewing of MCR-ALS interface, a screen resolution of 1,920×1,200 is recommended. Low-resolution screens can result in cropping of

text. MCR-ALS results as text are displayed in the MATLAB Command window, in case they are cropped in low-resolution screens (**Table 1**).

Software

• For data analysis and visualization: MATLAB R2013b (The MathWorks Inc., Natick, MA, USA) or newer versions are recommended. However, older versions are also valid.

• Statistics Toolbox™ for MATLAB and Bioinformatics Toolbox™ are required.

• Vendor software: Specific vendor software (*e.g.*, *Waters/ Micromass MassLynx*™, *Thermo Fischer Scientific Xcalibur*) are required for initial conversion of raw data formats into open data formats. Otherwise, the external program *ProteoWizard* can also be employed. However, the data provided by the authors for testing the protocol (see Supplementary Data) have already been converted and do not require the use of any of these software.

• For the TIMING section in the protocol, the following MATLAB version has been used: version 8.2.0 (R2013b), Win (64 bit).

Data files and MATLAB functions

• **Input data**. The following input data files were used in the **ANTICIPATED RESULTS** section of the present protocol (and are available as Supplementary Data): *Control1.mat*, *Control2.mat*, *Control3.mat*, *TBT1.mat*, *TBT2.mat* and *TBT3.mat*. All of them are MATLAB files obtained after conversion of their initial vendor formats (*Waters*) into open data formats (*i.e.*, *netCDF formats*) using *Databridge* interface and further import using Bioinformatics Toolbox™. Each file contains three variables in MATLAB workspace: vector *time*, containing information of all retention times, variable *mzCDFStruct*, containing information of the sample and the cell structure *peaks*, cell array providing information of m/z values and corresponding MS intensities measured by the mass spectrometer at each of the scans. Control samples including *Control1.mat*, *Control2.mat* and *Control3.mat* (1, 2 and 3 indicate replicate number) correspond to LC-MS data of extracted lipids of human placental choriocarcinoma cells (JEG-3) 24-h exposed to DMSO. TBT samples including *TBT1.mat*, *TBT2.mat* and *TBT3.mat* (again 1, 2 and 3 are indicators of the number of replicates), correspond to LC-MS data of extracted lipids of JEG-3 cells 24-h exposed to TBT. **All these data can be used for testing the protocol and as formatting guides for own data**.

• **MATLAB functions**. The following MATLAB functions are provided as Supplementary MATLAB functions to test the protocol whether with the provided data or with user's own data: *ROIpeaks.mat*, *ROIplot.mat*, *MSroiaug.mat* and *plotprofilestable.mat*.

! **CAUTION** Folder and file names must contain standard alphanumeric characters only (*e.g.*, unaccented Latin letters, numbers and underscore). No special characters are allowed. Do not use capitals; only use small letters for file extensions (*e.g.*, jpg and not .JPG, .txt and not .TXT, .mat and not .MAT and so on), as MATLAB is case-sensitive.

Procedure

PROCEDURE

Δ **CRITICAL** Many of the steps are prefaced by the term "optional". This means that technically future steps are not dependent on these having been performed. However, they can provide additional information, or they can alter the outcome of the analysis.

[Steps from 1 to 13 are necessary to import data into MATLAB environment. For data already in MATLAB format including Supplementary Data ***move directly to Step 14]***

*** Data conversion steps**

The following data conversion procedures are described for the distinct LC-MS vendors. In all cases, an external software called *ProteoWizard* can be used for data conversion (**option A**). On the other hand, specific vendor softwares can also be used with the same purpose. In this protocol we show two examples of data conversion using the specific vendor softwares of *Waters* and *Thermo Fisher* Corporations (**options B** and **C**).

(**A**) **Waters / Thermo Fisher / Agilent / AB Sciex / Bruker vendors (using ProteoWizard software)**

1| Install *Proteowizard* software as described in the web (**proteowizard.sourceforge.net**).
2| Go to MSConvert options, as shown in **Figure 1**.
3| Click ´Browse´ and select the source folder of the raw data files (.d) to convert. Multiple files can be selected at once, to be converted in batch mode.
4| Click the button ´Add´.
5| Select the output directory.
6| Select the output format (*mzXML* or *txt*).
7| Click ´Start´ to begin file conversion.

(**B**) **Waters Corporation (using MassLynx software)**

1| Open the *Databridge* interface of the *MassLynx* file converter as shown in **Figure 2**.
2| Click ´Select´ and browse the raw data files (.raw) to convert by searching on the directory where i is stored.
3| Click ´Options´ and specify the source of the raw files (*MassLynx*) and the target output format which must be *netCDF* for *cdf* files or *ASCII* for *txt* files.
4| Indicate the output directory where the new file will be stored and indicate the filename. Although the filename is already prefilled with the same name of the raw data file, it can be changed.
5| Click ´Convert´ to begin file conversion. A new box will appear indicating the % of completeness of the data conversion process.
? TROUBLESHOOTING

(**C**) **Thermo Fisher vendor (using Xcalibur software)**

1| Go to ´Tools > File Converter´as shown in **Figure 3**.
2| Specify the source data type.
3| Click ´Browse´ and select the source folder of the raw data files (.raw) to convert.
4| Select the desired files to convert. Multiple files can be selected at once, and all files are selected automatically by clicking on the button 'Select All'.
5| Click the button ´Add Job(s)´.
6| Select the destination path and data type, *ANDI Files* for *cdf* format or *Text Files* for the *txt* format.
7| Click ´Convert´ to begin file conversion.

protocol exchange
Login
Home | Browse | Share protocol | Lab groups | About | Contact

2/2/2018 A protocol for LC-MS metabolomic data processing using chemometric tools : Protocol Exchange

*** Starting and Data import steps using Bioinformatics Toolbox^{TM}**

8| Start MATLAB.

9| Navigate to the folder containing converted data files in *cdf* or *mzXML* formats, using the ´Current Folder´ panel in MATLAB.

10| Run the function *InfoStruct= mzcdfinfo* (*File*) or *InfoStruct= mzxmlinfo* (*File*) in the 'Command Window' panel in MATLAB. *InfoStruct* variable will appear in the workspace.
mzcdfinfo and *mzxmlinfo* functions extract the information of the *netCDF* or *mzXML* files, respectively, returning a MATLAB structure, named *InfoStruct*.

11| (Optional) Before going forward with the remaining procedure, have a look at the *InfoStruct* variable generated.
InfoStruct variable contains the following fields: Filename (name of the file), FileTimeStamp (date time stamp of the file), FileSize (size of the file in bytes), NumberOfScans (number of scans in the file), StartTime (run start time), EndTime (run end time), TimeUnits (units for time), GlobalMassMin (minimum m/z value in all scans), GlobalMassMax (maximum m/z value in all scans), GlobalIntensityMin (minimum intensity value in all scans), GlobalIntensityMax (maximum intensity value in all scans) and ExperimentType (indicates if data is raw or centroided).

12| Run the function *mzCDFStruct= mzcdfread* (*File*) or *mzXMLStruct= mzxmlread* (*File*) in the ´Command Window´ panel in MATLAB. *mzCDFStruct* or *mzXMLStruct* variables will appear in the workspace.
mzcdfread and *mzxmlread* functions read MS data from the *netCDF* or *mzXML* files and give as an output argument a MATLAB structure (*i.e.*, *mzCDFStruct* or *mzXMLStruct*) containing information of the LC-MS data.

13| Run the function [*Peaks*, *Time*]= mzcdf2peaks (*mzCDFStruct*) or [*Peaks*, *Time*]= mzxml2peaks (*mzXMLStruct*) in the 'Command Window' panel in MATLAB. A cell array named *peaks* and a vector named *time* will appear in MATLAB workspace.
These functions extract peak information from the MATLAB structures *mzCDFStruct* or *mzXMLStruct* created by *mzcdfread* or *mzxmlread* functions, respectively. The cell array named *peaks* contains mass/charge (m/z) and ion intensity values at each of the scans and the vector *time* gives information of the retention times associated with the LC-MS data set.
Δ **CRITICAL STEP** LC-MS data must be in *netCDF* or *mzXML* formats for their import. No other formats are accepted. Other file types need to be re-formatted (go back to **Step 1**) to match the input file requirements of Bioinformatics Toolbox^{TM} data import tools.

- -

*[The **following steps can be directly applied to the*** Supplementary Data ***provided with the present study**. **If used with new data**, **it must be imported to MATLAB obeying the steps previously explained** (**Steps 1 to 13**). **However**, **for a better understanding of the protocol the authors recommend to perform a first trial of ROI functions on the prepared*** Supplementary Data *]*

*** Data compression steps using ROI search**

14| Download the ROI package provided as Supplementary MATLAB functions (containing

https://www.nature.com/protocolexchange/protocols/4347

ROIpeaks, *ROIplot*, *MSroiaug* and *plotprofilestable* functions) and save it in a folder.

15| Go to ´Set path´ panel and add this folder to MATLAB search path.

16| Run the *ROIpeaks* function [*mzroi*, *MSroi*, *roicell*]= *ROIpeaks* (*peaks*, *snthresh*, *mzerror*, *minroi*, *nrows*, *time*) in the ´Command Window´ panel in MATLAB to search ROIs in the first sample. *ROIpeaks* function allows building an MS data matrix from variable *peaks* by selecting only the regions of interest. The implementation of this function requires the input of two variables containing information of the sample, *peaks* and *time*, together with the following parameters: *snthresh* (chromatographic signal-to-noise threshold, commonly between 0.1-1% maximum MS intensity, used to filter significative MS intensities), *mzerror* (admissible mass deviation, typically set to a generous multiple of the mass accuracy of the mass spectrometer, *e.g.*, 0.05 Da/e), *minroi* (minimum number of retention times to be considered in a ROI, normally between 5 and 12 seconds in UHPLC systems and between 20 and 50 seconds in HPLC systems) and *nrows* (number of cells/rows/spectra of the variable *peaks* desired to be processed). The output parameters of *ROIpeaks* function are *MSroi* (newly arranged matrix of dimensions (*num.of.scans* (m) x *nROI*), containing the MS spectra of every scan in its rows, and the chromatograms of every ROI in its columns), *mzroi* (vector containing final m/z values of all ROIs, calculated as the mean of all m/z classified within the same ROI), and *roicell* (cell array {nROI x 5}, containing nROI x 5 cells, providing information of m/z values (1), retention times (2), intensities (3), scan numbers (4) and mean m/z values of ROIs (5)).

When the process is finished a message indicating final number of ROIs and elapsed time will be displayed on the 'Command Window' screen. Variables *MSroi*, *mzroi* and *roicell* will appear in the workspace and two plots, one displaying *MSroi* respect to time and the other displaying the sum of *MSroi* respect to *mzroi* values will be automatically generated (see **Figure 4** and Supplementary Results 1).

? TROUBLESHOOTING

17| Run the function *ROIplot* (*roicell*(*n*)) in the 'Command Window'.

ROIplot function allows the evaluation of the ROI previously obtained, to avoid having multiple or halving peaks. The input variable of this function is the previously obtained *roicell* and the graphical output representations correspond to the chromatographic shapes of the obtained ROIs as well as the distribution of the distinct m/z values defining the same ROI (see **Figure 5**, where these plots are shown for a particular ROI). In this function, *n* specifies the particular ROI for which the results are shown. If no *n* is specified, results will be shown for all ROI values.

? TROUBLESHOOTING

Δ **CRITICAL STEP** Selecting the correct values for parameters *snthresh*, *mzerror* and *minroi* determines the outcome of the ROI search. Different values of parameters should be tested to see whether the elution profiles of the obtained ROI are meaningful or not. In the case of uncertainties, consult a mass spectrometry expert to avoid misinterpretation of the results.

18| Modify the values of input parameters, if necessary, and repeat the ROI search described in **Step 16** using the new values. Repeat this step the number of times required to obtain ROIs that fit original MS data.

19| (Optional) Change the name of the output variables in the workspace by right-clicking on them and selecting ´rename´ in the opening context-sensitive menu, to a name which makes reference to the sample group and number of replicate (*i.e.*, *MSroiC1* indicating that this variable corresponds to compressed data of the first replicate of a control sample).

20| (Optional) Save all the variables generated in the workspace, using 'Save workspace' button.

Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed. It is highly recommendable to select a name indicating the sample and the type of ROI information that it contains (*e.g.*, *ROIC1*, indicating that the .mat file provides information of individual ROI search of Control 1 sample).

∆ **CRITICAL STEP** Only alphanumeric filenames (*i.e.*, only unaccented Latin letters and numbers and underscore are allowed; special characters are not accepted). It is also important to have filenames that are representative for the sample.

21| (Optional) Save also the generated plots using their respective ´File/Save as…´ buttons, located above of each plot. A save dialog opens, with prefilled values for filename, format and location, which can be changed.

? TROUBLESHOOTING

22| Close figure windows individually.

! **CAUTION** Unsaved plots cannot be recovered after closing their respective windows.

23| Type ´clear all´at the MATLAB Command Window prompt to clear the MATLAB workspace and memory from all variables.

! **CAUTION** Unsaved data cannot be recovered after this step.

24| Type ´clc´ at the MATLAB Command Window prompt to clear the Command Window.

25| For data conversion and import of a new sample return to **Step 1**. For data compression of a new sample return to **Step 16**.

*** Steps to generate augmented data matrices**

26| Navigate to the folder containing the .mat files generated in the ROI search of individual data matrices (**Step 16**), using the ´Current Folder´ panel in MATLAB.

∆ **CRITICAL STEP** Search of ROI among samples and generation of augmented data matrices is only possible when previous ROI search of individual data matrices has been performed.

27| Double-click on two .mat files of two distinct samples (*e.g.*, *ROIC1.mat* and *ROIC2.mat*) to load them into MATLAB workspace. The loaded .mat files contain the variables *MSroiC1*, *MSroiC2*, *mzroiC1*, *mzroiC2*, *Time1* and *Time2* (see Supplementary Results 1) necessary for the ROI search among the two samples.

? TROUBLESHOOTING

28| Run the command [*Msroiaug*, *mzroiaug*, *timeaug*]= *Msroiaug* (*MsroiC1*, *MsroiC2*, *mzroiC1*, *mzroiC2*, *mzerror*, *thresh*, *TimeC1*, *TimeC2*) in the ´Command Window´ panel in MATLAB. A pairwise search of ROI among samples Control 1 and Control 2 is being performed, evaluating common and uncommon ROI values and finally considering both of them. At the end of this search, three new variables are generated in the workspace: *MSroiaug*, *mzroiaug* and *timeaug* and the same plots described in **Step 16** are again generated (see Supplementary Results 2). In this search, the parameter *mzerror* is used to define the admissible mass difference between two *mzroi* values to be considered the same (± *mzerror*/2).

? TROUBLESHOOTING

! CAUTION The two input *MSroi* matrices must have the same rt-mode dimensions to enable the search.

29| (Optional) Change the name of the output variables in the workspace by right-clicking on them and selecting ´rename´ in the opening context-sensitive menu, to a name which makes reference to the sample group and number of replicate (*e.g.*, *MSroiaugC1C2* indicating that the ROI search has

been conducted among Control 1 and Control 2 samples).

30| (Optional) Save all the variables generated in the workspace, using 'Save workspace' button. Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed. It is highly recommendable to select a name indicating the sample and the type of ROI information that it contains (*e.g.*, *ROIC1C2.mat* , indicating that the .mat file contains information of ROI search among Control 1 and Control 2 samples).

31| (Optional) Save also the generated plots using their respective ´File/Save as...´ buttons, located above of each plot. A save dialog opens, with prefilled values for filename, format and location, which can be changed.

? TROUBLESHOOTING

32| Close figure windows individually.

! CAUTION Unsaved plots cannot be recovered after closing their respective windows.

33| (Optional) Remove all variables containing information of ROI search of individual data matrices in the workspace (*e.g.*, *MsroiC1* , *MsroiC2* and so on) by right-clicking on the file and selecting ´delete´ in the open context-sensitive menu.

34| Load the .mat file containing information of the individual ROI search of a third sample classified as a control (*e.g.*, *ROIC3.mat*) by double-clicking on it.

35| Run the command [*Msroiaug*, *mzroiaug*, *timeaug*]= *MSroiaug* (*MSroiaugC1C2*, *MSroiC3*, *mzroiaugC1C2*, *mzroiC3*, *mzerror*, *thresh*, *timeaugC1C2*, *timeC3*)´ in the ´Command Window´ panel in MATLAB. A pairwise search of ROI among the previous generated *MSroiaugC1C2* matrix and the new *MSroiC3* matrix is being performed. At the end of this search, three new variables are generatec in the workspace: *MSroiaug* , *mzroiaug* and *timeaug* and the same plots described in **Step 16** are again generated but for the augmented case (see **Figure 6a**).

? TROUBLESHOOTING

36| (Optional) Change the name of the output variables in the workspace by right-clicking on them and selecting ´rename´ in the opening context-sensitive menu, to a name which makes reference to the sample group and number of replicate (*e.g.*, *MSroiaugC1C2C3* indicating ROI search has been conducted among Control 1, Control 2 and Control 3 samples).

37| (Optional) Save all the variables generated in the workspace, using 'Save workspace' button. Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed. It is highly recommendable to select a name indicating the sample and the type of ROI information that it contains (*e.g.*, *ROIC1C2C3.mat*, indicating that the .mat file contains information of ROI search among Control 1, Control 2 and Control 3 samples).

38| (Optional) Save also the generated plots using their respective ´File/Save as...´ buttons, located above of each plot. A save dialog opens, with prefilled values for filename, format and location, which can be changed.

? TROUBLESHOOTING

39| Close figure windows individually.

! CAUTION Unsaved plots cannot be recovered after closing their respective windows.

40| Type ´clear all´at the MATLAB Command Window prompt to clear the MATLAB workspace and memory from all variables.

! CAUTION Unsaved data cannot be recovered after this step.

41| Type ´clc´ at the MATLAB Command Window prompt to clear the Command Window.

42| Repeat **Steps 26** to **41** to find ROI values among other samples classified as another group

(*e.g.*, samples *TBT1* , *TBT2* and *TBT3*) to obtain information of common and uncommon ROI of the three stressed samples (*e.g.*, *ROIT1T2T3*). The plots obtained will be analogue to the ones represented in **Figure 6** but for the TBT-exposed samples in this case.

43| Repeat **Steps 33** to **39** with ROI values obtained independently for controls and stressed samples (*e.g.*, *ROIC1C2C3* and *ROIT1T2T3*) to find ROI values among the two groups of samples. Final obtained *MSroiaug* matrix (*MSroiaugC1C2C3T1T2T3*) is the column-wise compressed data matrix ready for the MCR-ALS analysis. The graphical outputs of the ROI search among the six samples are represented in **Figure 6b**.

44| Clear all variables individually in the workspace, by right-clicking on the file and selecting ´delete´ in the open context-sensitive menu, except final *MSroiaugC1C2C3T1T2T3* matrix and vectors *timeaugC1C2C3T1T2T3* and *mzroiaugC1C2C3T1T2T3*, which should be saved.

*** Data analysis steps for MCR-ALS method**

45| Download the freely available MCR-ALS GUI 2.0 and save it in a folder. In this web page information of MCR-ALS code, related tutorials and data sets for practicing can be found.

! CAUTION Although distinct programs can be downloaded from this webpage (MCR-ALS GUI 2.0, MCR-ALS Toolbox 1.0, MCR-ALS command line, MCR-ALS GUI and MCR-Bands), the newest version (MCR-ALS GUI 2.0) is the one used in this protocol.

! CAUTION For requirements regarding software description together with information of new features and applications of the latest version, please refer to another study[9].

46| Go to ´Set path´ panel and add this folder to MATLAB search path.

47| Type ´mcr_main´ at the MATLAB Command Window prompt to call the necessary auxiliary routines for the MCR-ALS analysis. The main window of MCR-ALS Toolbox is launched immediately (see **Figure 7**).

48| Select the data for MCR-ALS analysis by clicking on the ´Select a data matrix´ drop-down button (*e.g.*, *MSroiaugC1C2C3T1T2T3* , provided in the Supplementary Results 2). A new variable named ´mcr_str´ is generated in the workspace.

? TROUBLESHOOTING

49| Select the number of components of the initial estimation by clicking one of these buttons: ´Manual´ or ´SVD´. ´Manual´ button is used when prior knowledge about the correct number of components is available and ´SVD´ button is used when this estimate is performed considering the number of largest singular values obtained by the Singular Value Decomposition algorithm. In this case, the more appropriate option is the initial estimation through ´SVD´.

50| Click in ´SVD´ button and select the number of components following one of these two options. The first is to use the drop-down menu of Eigenvalues, listed in ascending order of component number (lower Eigenvalue for higher component number). The text box below automatically updates to show the selected number of components in red. The second option is to type the number of components into the text box with the same name and hit enter. The drop-down menu above automatically updates to show the corresponding singular value.

! CAUTION When having data matrices with one of the dimensions large (> 10000 elements), calculation of only few singular values is recommended to avoid computer memory problems. (Optional) For a proper choice of the number of components you can zoom in the ´EigenValues´ Representation and inspect when the rate of the decline between two consecutive values is much lower than for the previous pair of smaller eigenvalues.

After selecting the number of components, the ´EigenValues´ representation remains unchanged but the ´EigenVectors´ representation automatically shows the selected number of components for each of the individual matrices conforming the column-wise augmented data matrix (in this example case, the number of matrices is 6).

Δ **CRITICAL STEP** Selecting the correct number of components finally determines the outcome of the analysis. Distinct numbers of components should be tested and the results should be evaluated to see which gives the best result in terms of data fitting and chemistry and biology.

! CAUTION In an MCR-ALS analysis, adding a new component is not an additive process. In other words, it does not leave the original components intact, but it recalculates all components (see INTRODUCTION).

51| (Optional) Copy the box showing the number of components selected in the initial estimation by clicking the button ´Copy´.

52| Click ´OK´ button to return to the main screen of MCR-ALS program.

53| Start the initial estimation of one of the two factor matrices (**C** for concentrations or $\mathbf{S}^T$ for spectra) by selecting one of these three options: ´Manual´, if they are already available, ´Pure´ for determining initial estimates either of **C** or $\mathbf{S}^T$ by means of a purest variable detection method, or "EFA" by means of Evolving Factor Analysis[10], only suitable for the case of analyzing evolving processes. In this example, pure estimates will be used, which is calculated using a purest variable selection method (like in the SIMPLISMA method[11]).

54| Click in ´Pure´ button and select the direction of the variable selection (either concentrations or spectra) by using the drop-down menu of the ´Pure variable detection method´ box.

55| (Optional) Change the noise allowed (in percentage) for the calculation of initial estimates in the text box labeled ´Noise allowed (%)´ and hit enter. Although the default value of 1% is generally safe, different values can be tested and their effect evaluated in the ´Pure Spectra Estimation (Initial Values)´ plot. In this example, 10% of noise will be used to avoid selection of noisy variables.

56| Click ´Do´ button and examine the obtained ´Purest variables´ representation to see whether chromatographic/spectra profiles are reasonable or not (*e.g.*, whether they contain only noise or they are very similar to each other (can indicate that too many components were selected), whether they show every band in the spectra with equal weight (can indicate too few components selected), or whether they contain artifacts (can indicate improper pre-processing)).

The list of purest variables is immediately shown in a box emplaced in the left.

Δ **CRITICAL STEP** It is important to see whether the pure spectral estimates are meaningful or not, as this can help in selecting the correct number of components. If the addition of a new component (**Step 50**) does not result in a significantly different new spectral estimate, it is likely that the new component is not required and will not be well resolved.

57| Click ´OK´ button to return to the main screen of MCR-ALS program.

58| Initiate the optimization process by clicking ´Continue´ button, at the bottom of the main interface box. A summary screen in which the top plots represent the row and column profiles of the experimental data, the middle plots show the initial estimate and the **C** and $\mathbf{S}^T$ profiles obtained by a least-squares step and the bottom plots represent the score and loading plots obtained by PCA of the analyzed data matrix **D** with the previously selected number of components will appear (see **Figure 8**).

59| Modify the number of matrices simultaneously analyzed by writing the number in a text box above the plots. In this example, this number is "6".

! CAUTION The default value for the number of matrices is "1" since default conditions are established for a single data matrix analysis.

60| Click ´Continue´ button to proceed with the definition of the data set.

61| In the ´Definition of the data set´ window, define the type of multiset data structure by selecting the correct option in the drop-down menu: column-wise augmented data matrix (**C** direction), row-wise augmented data matrix (**S** direction) or column- and row-wise augmented data matrix (**C & S** directions). In this example, the ´column-wise augmented data matrix´ is selected, with 6 submatrices all having the same number of rows (see **Figure 9**).

62| Click ´OK´ button to proceed with the selection of constraints for ALS optimization.

In this new version of the interface, there are two differentiated screens for the choice of constraints, one for the profiles linked to the row mode (*i.e.* concentration profiles, **C** matrix) and another for the profiles related to the column mode (*e.g.*., spectral profiles, $\mathbf{S}^T$ matrix).

63| In the ´Constraints: row mode (concentrations and multiple experiments)´ window indicate whether the same constraints will be applied to all **C** submatrices or not (see **Figure 10a**).

At the top of the screen, a panel regarding the multiset data structure is presented. It contains the total number of **C** submatrices included in the augmented data set, an option to apply the same constraints to all **C** submatrices, or the possibility to change the constraints according to the different **C** submatrices. Finally, at the right corner, the possibility to apply the constraint of correspondence among species by selecting which components are present in every considered **C** submatrix is offered. In this case, the same constraints will be applied to all **C** submatrices.

! CAUTION In the selection of row constraints for augmented data matrices is not allowed the possibility to deal with multiple and different constraints for every analyzed **C** submatrix.

64| Select the constraints among the four common options (non-negativity, unimodality, closure and equality constraints) and more advanced constraints (such as correlation or kinetic hard-modeling). In the present example, only non-negativity constraints are applied when selecting the option ´forced to zero´ in the drop-down menu.

Δ CRITICAL STEP The implementation of non-negativity constraints through the ´forced to zero´ option is recommended to speed up the calculation.

65| Click ´Continue´ button to proceed with the selection of constraints of $\mathbf{S}^T$ matrix.

66| In the ´Constraints: column mode (spectra and single technique)´ window select the constraints for $\mathbf{S}^T$ matrix among the four common options: non-negativity, unimodality, closure and equality constraints (see **Figure 10b**). In this example, non-negativity constraints through ´forced to zero´ option are implemented.

67| Click ´Continue´ button.

! CAUTION When no closure is selected (*e.g.*, no mass balance in concentrations) constraints, a new window appears to offer the possibility of normalizing the resolved spectra profiles (*e.g.*, normalizing them to have equal height, total sum norm or Euclidean norm) prior to starting ALS optimization. This is recommended to avoid scale instabilities during the evolution of the ALS optimization and it fixes the possible intensity ambiguities. In this example, ´spectra equal height´ was selected.

68| Select general optimization parameters (*e.g.*, the number of iterations or convergence criterion) and the name of output variables in the ´Parameters/Output of ALS optimization´ screen (see **Figure 11**). In this case, a total of 50 iterations are selected (default value) and the convergence criterion is set to 1%.

69| Select the box to enable the graphical output of the results and click ´Continue´ button. Suggestions for a variable name for concentrations and spectra matrices resulting from MCR-ALS analysis are *copt* and *sopt*, respectively.

70| Evaluate the results shown in ´ALS optimization´ screen including information about the convergence, lack of fit and explained variance (see **Figure 12**).

71| (Optional) Click ´Information´ button to obtain more detailed information about the evolution of the ALS optimization (*e.g.*, plots of explained variance, lack of fit, logarithm of the sum of squares residuals and evolution of concentration/spectra profiles).

72| (Optional) Save all the variables generated in the workspace (see Supplementary Results 3), using 'Save Workspace' button. Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed.

73| (Optional) Save also the generated plots using their respective ´File/Save as…´ buttons, located above of each plot. A save dialog opens, with prefilled values for filename, format and location, which can be changed.

? TROUBLESHOOTING

74| Close figure windows individually.

! CAUTION Unsaved plots cannot be recovered after closing their respective windows.

75| Clear all MATLAB variables in the workspace, by right-clicking on the file and selecting ´delete´ in the open context-sensitive menu, except from variables *MSroiaugC1C2C3T1T2T3* , *timeaugC1C2C3T1T2T3* , *copt* , *sopt* and vector *mzroiaugC1C2C3T1T2T3* .

! CAUTION Unsaved data cannot be recovered after this step.

76| (Optional) Rename those variables using shorter names (*e.g.*, *x* , *time* , *c* , *s* and *mz* instead of *MSroiaugC1C2C3T1T2T3* , *timeaugC1C2C3T1T2T3* , *copt* , *sopt* and vector *mzroiaugC1C2C3T1T2T3*, respectively).

77| Type ´clc´ at the MATLAB Command Window prompt to clear the Command Window.

*** Steps to evaluate concentration and spectral profiles of MCR-ALS components**

78| Create two new variables in the workspace named as *nexp* and *ncontrol* containing information about the number of experiments and the number of controls of the data sets (in this case, *nexp*=6 and *ncontrol*=3) by typing *nexp*=6 and *ncontrol*=3 in the ´Command Window´ prompt.

79| Run the command [*area*, *height*, *table*, *table2*]= *plotprofilestable* (*x*, *c*, *s*, *time*, *mz*, *nexp*, *ncontrol*) in the ´Command Window´ panel in MATLAB. Two graphical outputs will be obtained for the first component. In addition, the results of the corresponding statistical evaluation will appear in the MATLAB Command Window followed by the message: "select forward backward plot 1/0". In order to proceed with the evaluation of next MCR-ALS component introduce "1" in the Command Window, otherwise write "0".

! CAUTION If Step 76 was skipped, the input variable names of this function must be changed to those used to define the corresponding MATLAB variables.

80| (Optional) Save the two generated figures, resulting from the analysis of the first component (see **Figures 13a** and **13b**), using their respective ´File/Save as…´ buttons, located above of each plot. A save dialog opens, with prefilled values for filename, format and location, which can be changed.

? TROUBLESHOOTING

81| Close figure windows individually.

! CAUTION Unsaved plots cannot be recovered after closing their respective windows.

82| (Optional) Copy the information provided in the ´Command Window´ panel (*lof*, *fit*, *Rsquare*, etc) and paste it in another document to further save it (see **Figure 13c**).
83| Click to any button to obtain the same plots and results for the next component.
84| Repeat **Steps 80** to **83** until the last component.
85| Once obtained the results for all components click once more any key of the computer keyboard. Statistic results when considering all components simultaneously will be presented in the ´Command Window´ panel (see **Figure 13d**).
86| (Optional) Copy the information provided in the ´Command Window´ panel (*lof*, *fit*, *Rsquare*) and paste it in another document to further save it.
87| Click again any key to obtain in the workspace the two tables containing statistical information (Table and Table2).
88| Click ´Save workspace´ button of the upper panel of MATLAB to save all variables contained in the Workspace. Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed. It is highly recommendable to select a name indicator of samples and type analysis (*e.g.*, *MCR-ALS.mat*, indicating that the .mat file contains information of the MCR-ALS analysis.
89| Type ´clear all´at the MATLAB Command Window prompt to clear the MATLAB workspace and memory from all variables.
! CAUTION Unsaved data cannot be recovered after this step.
90| Type ´clc´ at the MATLAB Command Window prompt to clear the Command Window.

Timing

The timing required for the distinct steps of LC-MS data analysis described in the **PROCEDURE** section is variable but is usually between 2 and 4 min per sample for data compression and import, about 2 min for data compression following a ROI search and between 5 and 10 min per sample for the MCR-ALS analysis.

Troubleshooting

Troubleshooting advice can be found in **Table 1**.

Anticipated Results

Although the data used as example in the present protocol was also used in a previous study by the authors[6], the results hereby presented were not included in the original publication and are specifically selected now in order to demonstrate the key features of the present protocol. These results include data compression through ROI search of individual data matrices (see Supplementary Results 1), data matrix augmentation through ROI search among data matrices (see Supplementary Results 2) and MCR-ALS analysis of the obtained MSroi augmented matrix (see Supplementary Results 3).
The utilized data to obtain all these results are provided as Supplementary Data and the functions used are supplied as Supplementary MATLAB functions.

References

1. Tautenhahn, R., Böttcher, C., Neumann, S. Highly sensitive feature detection for high resolution LC-MS. *BMC Bioinf.* 9, 504 (2008).

2. Stolt, R., Torgrip, R., Lindberg, J, Csenki, L., Kolmert, J., Schuppe-Koistinen I., Jacobsson, S. Second-Order Peak Detection for Multicomponent High-Resolution LC-MS Data. *Anal. Chem.* 78: 975-983 (2006).

3. Tikunov, Y., Lommen, A., Vos, Cd., Verhoeven, H., Bino, R., Hall, R., Bovy, A. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant. Physiol.* 139(3): 1125-37 (2005).

4. Smith, C., Want, E., O'Maille, G., Abagyan, R., Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal. Chem.* 78: 779-787 (2006).

5. Katajamaa, M., Oresic, M. Processing methods for differential analysis of LC-MS profile data. *BMC Bioinf.* 6: 179 (2005).

6. Gorrochategui, E., Casas, J., Porte, C., Lacorte, S., Tauler, R. Chemometric strategy for untarget lipidomics: biomarker detection and identification in stressed human placental cells. *Anal. Chim. Acta*. 854: 20-33 (2015).

7. Farrés, M., Piña, B., Tauler, R. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS. *Metabolomics* 11: 210-224 (2015).

8. Tauler, R. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.* 30: 133-146 (1995).

9. Jaumot, J., De Juan, A., Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemometr._ Intell_._ Lab_. Syst.* 140: 1-12 (2015).

10. Maeder, M. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.* 59: 527-530 (1987).

11. Bu, D., Brown, C.W. Self-modeling mixture analysis by interactive principal component analysis. *Appl. Spectrosc.* 54:1214–1221 (2000).

Acknowledgements

The research leading to these results has received funding from the **European Research Council** under the European Union's Seventh Framework Programme (FP/2007-2013) / **ERC Grant Agreement n. 320737**. First author acknowledges the Spanish Government (Ministerio de Educación, Cultura y Deporte) for a predoctoral FPU scholarship.

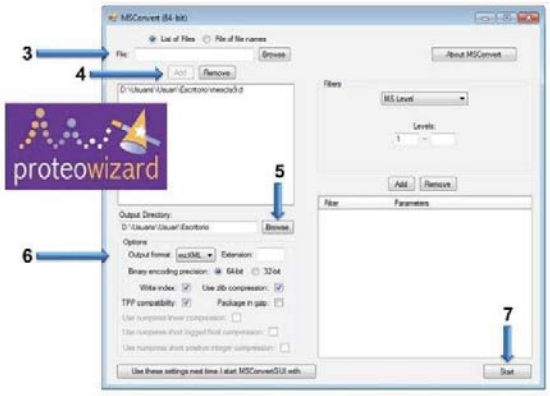
protocol exchange

Login

Home | Browse | Share protocol | Lab groups | About | Contact

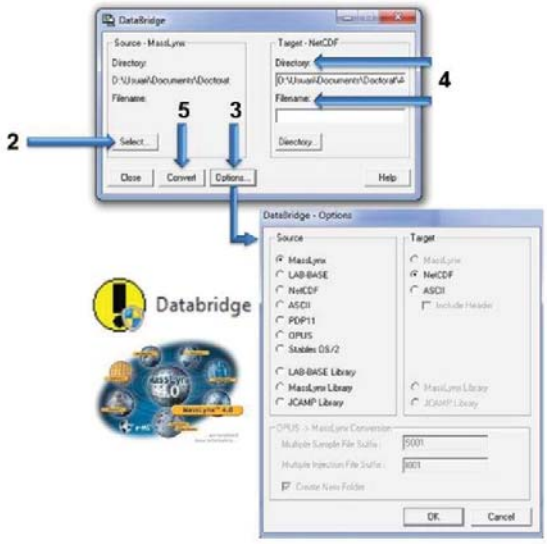
Figures

Figure 1: Data conversion interface of *ProteoWizard* software: MSConvert.



The input fields or icons of the software are numbered according to their corresponding **PROCEDURE** steps. First, files to be converted should be selected (**3** and **4**). Then, output folder (**5**) and options related to the conversion process such as the output format or the binary encoding should be selected (**6**) previous to the beginning of the conversion (**7**).

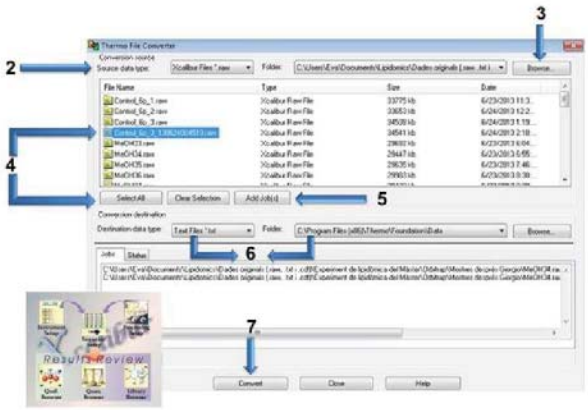
Figure 2: Data conversion interface of Waters vendor: Databridge tool from MassLynx softwar



The input fields or icons of the software are numbered according to their corresponding **PROCEDURE** steps. First, source files should be selected (**2**) and options related to the conversion process predefined (**3**): source data type (usually *MassLynx*.raw) and target output type (*netCDF* is recommended for further work). Target folder and filenames (**4**) should be indicated prior to conversion.

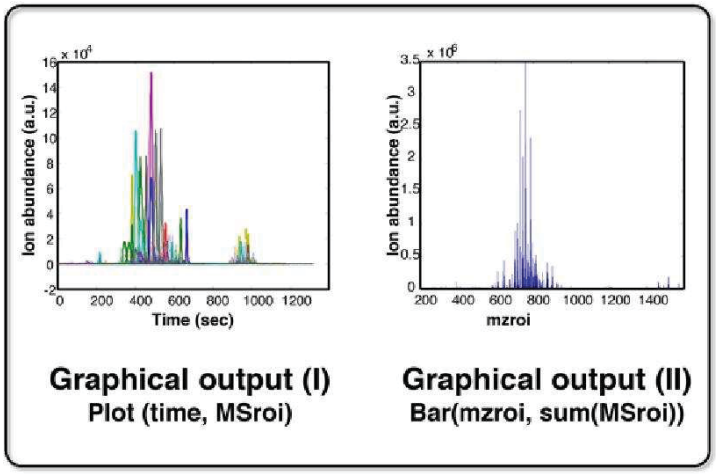
https://www.nature.com/protocolexchange/protocols/4347

Figure 3: Data conversion interface of Thermo Fischer vendor: File converter tool from Xcalibur software.



The input fields or icons of the software are numbered according to their corresponding **PROCEDURE** steps. First, source files should be selected from the available formats (usually, .raw) (**2**). Files selected are added to the job queue by clicking the "Add Job(s)" button (**5**). Format and folder of output converted files should be also indicated (**6**), and it is recommended to select the ANDI (.cdf) data type.

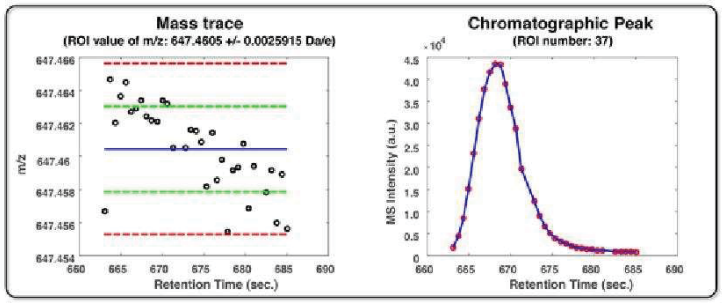
Figure 4: Graphical outputs obtained after a ROI search using ROIpeaks function in sample Control 1



(**I**) plot of *MSroi* respect to time (new compressed chromatogram) and (**II**) bar plot of the sum of *MSroi* intensities respect to *mzroi* values (new MS spectra).

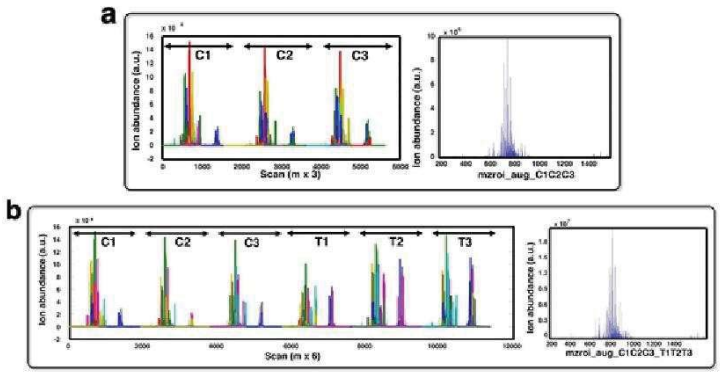
https://www.nature.com/protocolexchange/protocols/4347

Figure 5: Representation of a chromatographic elution profile and the corresponding mass trace of a particular ROI



In the mass trace representation, the continuous blue line represents the mean value of *mzroi*, green dotted-lines represent the mass deviation intervals and red dotted-lines two times the mass deviation interval.

Figure 6: Results of a ROI search among matrices when using MSroiaug function



(**a**) ROI search among *Control1* (*C1*), *Control2* (*C2*) and *Control3* (*C3*) and (**b**) ROI search among the six samples. **Note**: The results presented in this figure were obtained for a ROI search among samples fixing a threshold value of 750 a.u. and an mzerror of 0.05 Da/e. In addition the index *m* is used to represent the number of scans of one sample.

https://www.nature.com/protocolexchange/protocols/4347

Figure 7: MCR-ALS GUI 2.0 main window (mcr_main) and other subwindows corresponding tc SVD calculations and pure variable Initial ALS estimations

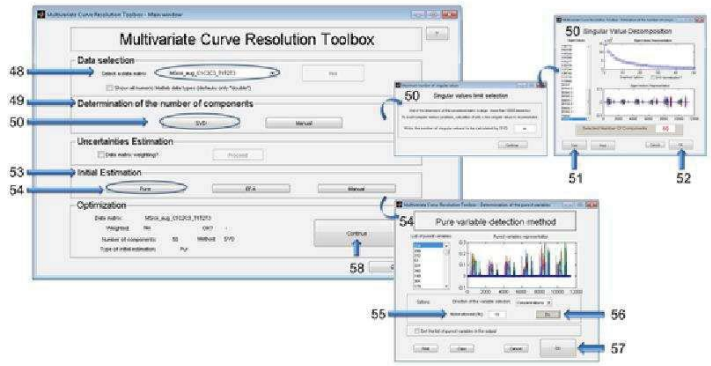
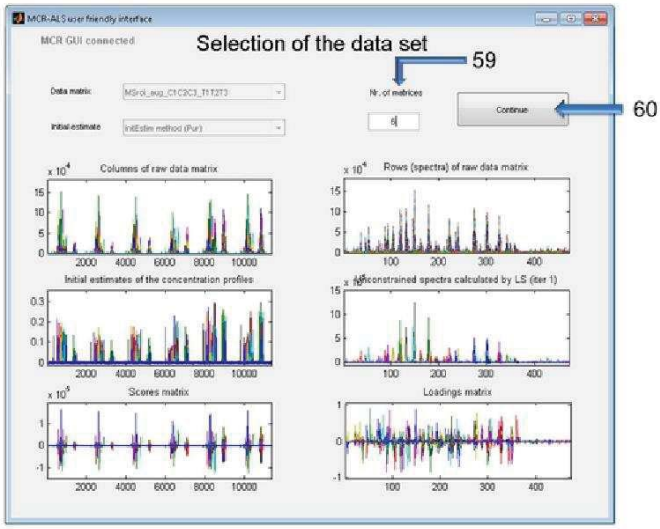


Figure 8: Data set selection. MCR-ALS GUI window.

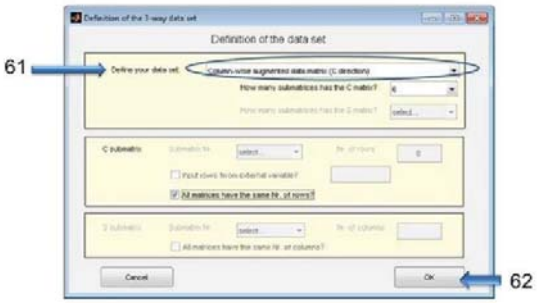


MCR-ALS window plots representing row (top right) and column (top left) profiles of the experimental data matrix **D**; plots showing $\mathbf{S}^T$ spectra initial estimates (middle left), and **C** concentration profiles (middle right) estimated by least-squares; and PCA scores (bottom left) and loadings (bottom right) plots of the experimental data matrix **D** using the selected number of components. A critical issue in this screen is the selection of the correct number of experiments analyzed simultaneously in the edit box (**59**). In this example, six matrices are analyzed at the same time.

https://www.nature.com/protocolexchange/protocols/4347

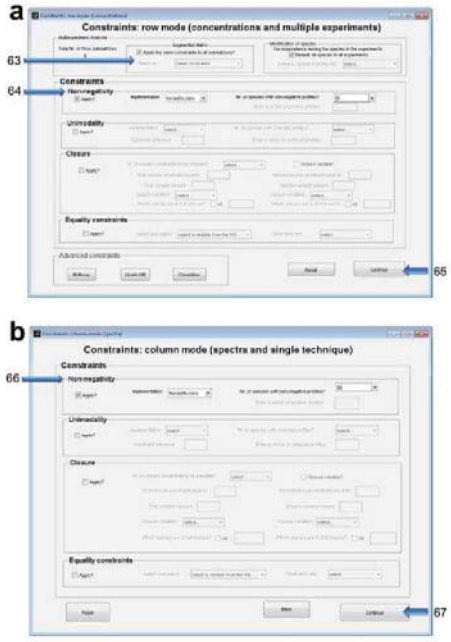
2/2/2018 A protocol for LC-MS metabolomic data processing using chemometric tools : Protocol Exchange

Figure 9: Selection of the type (column- or row-wise) of augmented data matrix MCR-ALS window.



MCR-ALS window of the definition of the data set in the case of considering simultaneously more than one matrix. In the omic case, column-wise augmented data set should be selected (**61**) as several experiments monitored with the same technique (usually, MS) are considered. If all the considered experiments have the same number of rows, it is recommended to click the checkbox "All matrices have the same Nr. of rows?" that facilitate the input of the information related to the number of rows of each considered matrix.

Figure 10: MCR-ALS constraints selection windows.



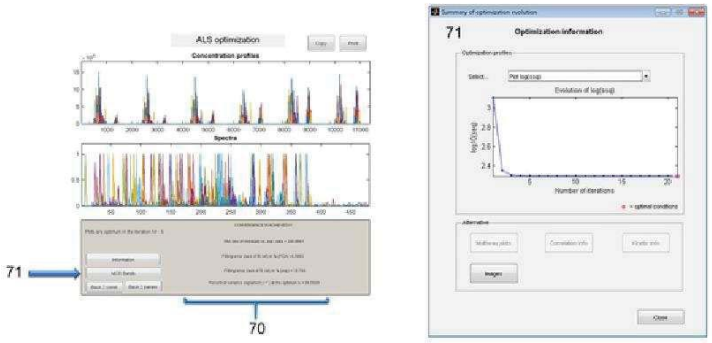
Selected constraints (**a**) for **C** matrix (concentration/rows profiles), and (**b**) for $\mathbf{S}^T$ matrix (spectra/columns profiles). In the case of the omic studies, only non-negativity constraints should be applied to both **C** and $\mathbf{S}^T$ profiles enforcing that the resolved chromatographic elution and MS spectra are positive (**64** and **66**). If multiple matrices are analyzed simultaneously, it is possible to apply the same constraints to all the experiment by clicking the appropriate checkbox (**63**). If it is not checked, then each matrix could have different constraints.

https://www.nature.com/protocolexchange/protocols/4347

Figure 11: MCR-ALS optimization parameters and outputs window.

MCR-ALS window showing general parameters for ALS optimization (default parameters are 50 iterations and a convergence criterion of 0.1% of percentage of change of the standard deviation of residuals between consecutive iterations, however in the presented example the latter value has been changed to 1% to allow a faster iteration) and output variable names for MCR-ALS results (**68**).

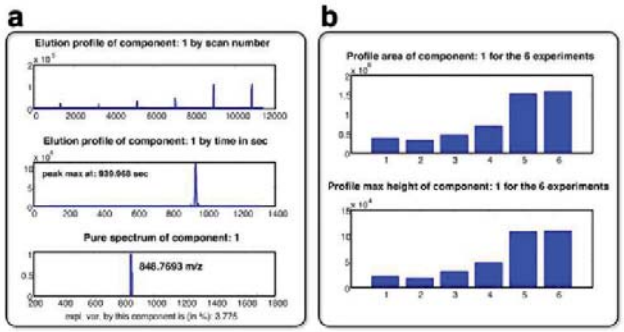
Figure 12: MCR-ALS optimization results and other related information window.



MCR-ALS window showing the results of ALS optimization (**70**). Final screen with information about the optimation process: number of iterations, convergence/or divergence, standard deviation of residuals respect experimental data, fitting error of the model considering both experimental and PCA reproduced data, and the percent of variance explained (R^2) Additional information related to the optimization process can also be obtained such as the evolution of the logarithm of the sum of squares or the lack of fit of the model (**71**).

protocol exchange
Login
Home | Browse | Share protocol | Lab groups | About | Contact

Figure 13: Results of the statistical evaluation performed on MCR-ALS components to determine whether they present significant differences among controls and stressed samples



c

Analysis of profiles for component: 1
lof (%) = 98.0943
fit (%) = 1.9057
Rsquare is 3.774987771764
time, mz: 939.968 846.7693
Area, height: 5245597.4287 109879.40198
Areas: 415667.7487 358443.31543 530667.71828 743128.56909 1573791.7461 1623898.3311
Heights 21635.1843 18002.91194 30828.51867 47848.65725 106979.6545 109879.402
Mean areas in control and treated samples are: 434926.26081 1313606.2154
Areas fold change is: 1 3.0203
Mean heights in control and treated samples are: 23488.8716 86235.9046
Height fold change is: 1 3.7565
Paired t-test results
H = 0 diff in peak areas control/test samples (y/n 1/0) with p significance: 0.087125
H = 0 diff in peak heights control/test samples(y/n 1/0) with p significance: 0.07962
Two-samples t-test results
H = 1 diff in peak areas control/test samples (y/n 1/0) with p significance: 0.038807
H = 1 diff in peak heights control/test samples(y/n 1/0) with p significance: 0.034583
Rank-sum test results
H = 0 diff in peak areas control/test samples (y/n 1/0) with p significance: 0.1
H = 0 diff in peak heights control/test samples(y/n 1/0) with p significance: 0.1

d

Total explained variance / % lack of fit fit
lof (%) = 18.706
fit (%) = 81.294
Rsquare is 96.500854331747
Total sum of individual explained variances is (%): 99.8236
Total overlapped variance is (%): 3.3228

(**a**) elution and pure spectra profiles of component 1, (**b**) profile areas and max height of component 1 among the six samples, (**c**) results derived from the statistical evaluation of component 1 and (**d**) results derived from the statistical evaluation of all the components.

https://www.nature.com/protocolexchange/protocols/4347

protocol exchange
Login
Home | Browse | Share protocol | Lab groups | About | Contact

Table 1: Troubleshooting table.

Step	Problem	Possible reason	Solution
(B)5	A box appears with this error message: "*The source file contains 2 functions. netCDF only supports one function per file. The generated NetCDF files will be named according to the function number they are derived from. Continue?*"	The data conversion tool detects that two functions are saved in the same folder. First one contains acquired data of the sample and second one of the reference compound (leucine).	Click on the button 'accept' at the bottom of the out-coming box and only consider one of the two netCDF files generated (that one named as *file1.CDF*). The other one can be suppressed.
16	Index exceeds matrix dimensions. *This error message appears in the 'Command Window' of MATLAB.*	The dimensions of vector time and the data matrix differ.	Make sure that the number of retention times selected to be processed coincide with the dimensions of Peaks variable.
16, 17, 28, 35	Undefined function or variable in a MATLAB script. *This error message appears in the 'Command Window' of MATLAB.*	Names of variable inputs of a script written in the 'Command Window' do not match variable names in workspace.	Ensure the consistency among the variable names in the workspace and the ones written in the command line.
21, 31, 38, 73, 80	Saved plots are of suboptimal quality	File format is suboptimal	Change the file type in the save dialog box
	Text is cropped in the GUI	Low-resolution computer screen	If possible, increase the screen resolution. Maximize the GUI window.
27	File does not load	Improper file formatting or name	Make sure that file is in the correct format (see the Data import section of this PROCEDURE), and that the folder names do not contain special characters.
48	Script is extremely slow to start	Insufficient computing power	Shut down unnecessary processes

Associated Publications

This protocol is related to the following articles:

- Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells

https://www.nature.com/protocolexchange/protocols/4347

2.2.2. SCIENTIFIC ARTICLE III

ROIMCR: a powerful data analysis strategy for LC-MS metabolomic data sets

E. Gorrochategui, J. Jaumot, R. Tauler

Submitted for publication

METHODOLOGY ARTICLE

ROIMCR: a powerful data analysis strategy for LC-MS metabolomic data sets

Eva Gorrochategui[a], Joaquim Jaumot[a], Romà Tauler[a,*]

egmqam@cid.csic.es
joaquim.jaumot@idaea.csic.es
roma.tauler@idaea.csic.es

[a] *Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, 08034, Catalonia, Spain.*
* Corresponding author: Romà Tauler, Tel.: +34-934006140

ABSTRACT

Background: The analysis of LC-MS metabolomic data sets appears as a challenging task in a wide range of disciplines since it demands for highly-extensive processing of a vast amount of data. Different LC-MS data analysis packages (*e.g.*, *XCMS*, *MZmine* and *MetAlign*) have been developed in the last years in an attempt to facilitate this analysis. However, most of these strategies involve chromatographic alignment and peak shaping and often associate each "feature" (*i.e.*, chromatographic peak) to a unique *m/z* measurement. Thus, the development of an alternative data analysis strategy applicable to most types of MS data sets, which properly addresses these issues, is still a challenge in the metabolomics field.

Results: Here we present an alternative approach called ROIMCR to: i) compress massive LC-MS data while transforming their original structure into a data matrix of reduced dimensions without missing relevant information through the search of regions of interest (ROI) in the *m/z* domain and ii) resolve compressed data to find their contributing pure components without previous alignment nor peak shaping by applying Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) analysis. For the first time, the basics of the ROIMCR method are presented in detail. The functions for ROI compression have been already provided in a protocol written by the authors available at https://www.nature.com/protocolexchange/protocols/4347 and the already existing MCR-ALS interface is accessible at www.mcrals.info. Data analysis is performed under MATLAB (The MathWorks, Inc., www.mathworks.com) programming and computing environment. An example of the use of ROIMCR methodology is provided with LC-MS data generated in a lipidomic study.

Conclusion: The methodology hereby presented combines the benefits of data compression based on the search of ROIs (*i.e.*, no loss of spectral accuracy) with the benefits of MCR-ALS analysis (*i.e.*, powerful data resolution without the necessity of performing neither peak alignment nor peak shaping). The presented method is a powerful alternative to other existing data analysis approaches that do not use MCR-ALS analysis to resolve LC-MS data. Moreover, it is also an improved version of other MCR-ALS based approaches that use less-powerful data compression strategies such as *binning* and *windowing*. Overall, the presented strategy demonstrates the usefulness of chemometrics in data analysis and it is a valuable addition to the untargeted metabolomic research.

Keywords: LC-MS, Data analysis, Data compression, Data resolution, Regions of interest (ROI), MCR-ALS, Metabolomics, Lipidomics, Chemometrics, Untarget.

1. Background

The challenge of data analysis is one of the main concerns of metabolomic liquid chromatography coupled to mass spectrometry (LC-MS) studies[1]. Lots of software packages exist for MS-based metabolomic data analysis, including propriety commercial, open-source, and online workflows[2]. Some commercial tools provided by major vendors of MS and omics high throughput analytical instruments and equipment include MassHunter (Agilent technologies), SIEVE (Thermo Scientific) and Progenesis QI (Waters). Among open-source software some of the most used include XCMS[3] (and XCMS-based Metabox[4], metaX[5]), CAMERA[6], MAIT[7], MetaboAnalyst[8], Workflow4Metabolomics[9], MZmine[10] and MetAlign[11]. However, none of these approaches can be singled out as the best strategy and the methodological discrepancies existing among them make LC-MS data analysis an unresolved problem in the bioinformatics field.

Data analysis of high resolution LC-MS based metabolomic data sets usually begins with their compression, required to reduce them into formats that are manageable with computers (without compromising the original information comprised within) and prevent errors linked to the restricted memory capacity of the computers. The high-dimensional nature of LC-MS based metabolomic data sets is attributed to the superior number of measurements (*m/z* values) related to the number of observations (samples). Apart from compressing data, in this first step, the conversion of raw data into a matrix representation is also required to obtain a well-structured variable to work with. The generated data matrices (x,y) are arranged with retention times in the rows (x-direction) and *m/z* values in the columns (y-direction). A classical procedure used for data compression and matrix transformation is the one referred to as *binning*. With the *binning* procedure, high-resolution raw mass spectra are converted into a matrix representation by dividing the *m/z* axis into parts with a specific *bin* size, generally set to a multiple of the mass accuracy of the mass spectrometer. However, a significant disadvantage of *binning* is the complication related to the right choice of the bin size for a specific data set, being the selection of the *m/z* bin size intensely associated with the recovery of the proper elution profile shape. If the selected bin size is excessively small, chromatographic peaks can fluctuate between bins and therefore not be determined because of the absence of the chromatographic shape of the peak. If the bin size is excessively big, various peaks may occur in the same bin, and tiny peaks might disappear by the elevated noise level[12].

One more major drawback of *binning* is the reduction of spectral accuracy originated from the compression of data made in the *m/z*-mode dimension, which goes in detriment of final identification of metabolites. Moreover, in most cases the compression performed with *binning* is not sufficient and further *windowing* (*i.e.*, selecting continuous regions in the rows (time) direction or the column (*m/z*) direction to be analyzed independently) is necessary. Nevertheless, when performing *windowing*, the whole process is more tedious and prolonged in time, since one sample has to be analyzed by parts.

A better alternative compression strategy to *binning* and *windowing* is based on the idea of assuming analyte signals as a domain of data points with a high density arranged by a particular "data void", first presented by Stolt et al.[13] These regions where analytes are found are called *regions of interest* (*ROI*) and are searched according to specific criteria (*i.e.*, particular threshold intensity, admissible mass error and minimum number of occurrences). Overall, ROI compression strategy consists on considering data included in these regions while rejecting the other. This strategy has already been implemented in the *centWave* algorithm of *XCMS* software[12]. The result of the search of ROIs in a sample is a set of mass traces of distinct dimensions that have to be finally reorganized into a data matrix. Differing from the *binning* procedure, no reduction of spectral resolution occurs as a result of an ROI compression since no bin size has to be fixed. Thus, ROI compression strategy allows taking full advantage of all the benefits offered by high resolution MS techniques. At present, most of the current metabolomic data analysis software tools use ROI compression, as a previous step to peak detection and/or integration.

Following data compression, next crucial step in LC-MS based metabolomic data analysis is data resolution. Most of the existing LC-MS data analysis approaches require two steps (*i.e.*, chromatographic alignment and peak shaping) before peak resolution. Alignment methods look for matching peaks over various chromatographic runs and peak shaping methods model peaks to have a delimited and more regular shape, habitually through the application of continuous wavelet transformations (CWT) and optionally Gauss-fitting[14]. Therefore, preliminary peak correction appears as an indispensable step in most of the present data analysis packages and often is linked to a high source of error. In contrast, neither of the two corrections (*i.e.*, peak alignment and shaping) are required when using Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[15] methods,

since the alignment of distinct chromatographic runs is produced in the spectral direction or mode. It is precisely in the case of multirun chromatographic simultaneous analysis where MCR methods are exceptional powerful tools for mixture analysis and resolution. The main goal of MCR-ALS methods is to resolve spectra arising from mixtures of the chemical constituents present in a sample into contributions from the individual components making up the mixtures. That is, MCR-ALS seeks to model the underlying physical processes that generate the data in terms of a sample's composition. MCR-ALS resolved MS spectra profiles can immediately be used to identify the chemical identity of metabolites by comparison with standards or by library searching. In the last years, MCR-ALS methods have emerged as highly effective tools to resolve the elution problem in different application areas, and in particular in LC-MS based metabolomic data sets.In this work, we provide a new data analysis strategy, known as ROIMCR, to compress and resolve LC-MS metabolomic data sets. Data compression is performed without losing spectral accuracy by the search of ROI, and chromatographic peaks are resolved through the application of MCR-ALS analysis.

The main steps involved in data compression and data resolution are represented in **Fig. 1**. As it can be observed in the figure, after a first data compression through the search of ROIs, the obtained ROI profiles are evaluated to see whether they properly agree with original data features or not. Such compression can be performed in a single file or in multiple files, generating in the latter case column-wise augmented ROI data matrices (*i.e.*, matrices containing distinct submatrices related to distinct samples attached one on top of each other). The generated augmented ROI matrices are further analyzed by MCR-ALS. Finally, the ultimate step would be the statistical evaluation of the resolved MCR-ALS components for the discovery of potential biomarkers. A distinct feature of the proposed ROIMCR strategy is its current implementation under the powerful MATLAB computer and visualization environment, which is very much used worldwide in the Chemometrics field and in scientific and technological software development with all its advantages and toolboxes already incorporated.

Moreover, in this study we provide an example of the performance of ROIMCR strategy with a lipidomic LC-MS data set. The illustrating lipidomic data set was generated in an experiment performed in a previous study of the authors[16] in which a human placental chroriocarcinoma cell line (JEG-3) was exposed to the endocrine disruptor chemical tributyltin (TBT). Researchers interested on ROIMCR procedure, can test this strategy using the example data and the MATLAB functions for ROI compression both provided in a protocol written by the authors[17]. That protocol, available at https://www.nature.com/protocolexchange/protocols/4347, provides step-by-step information of the implementation of ROIMCR procedure. In the present work, the description of the basics and fundamentals of the methodology are presented in detail.

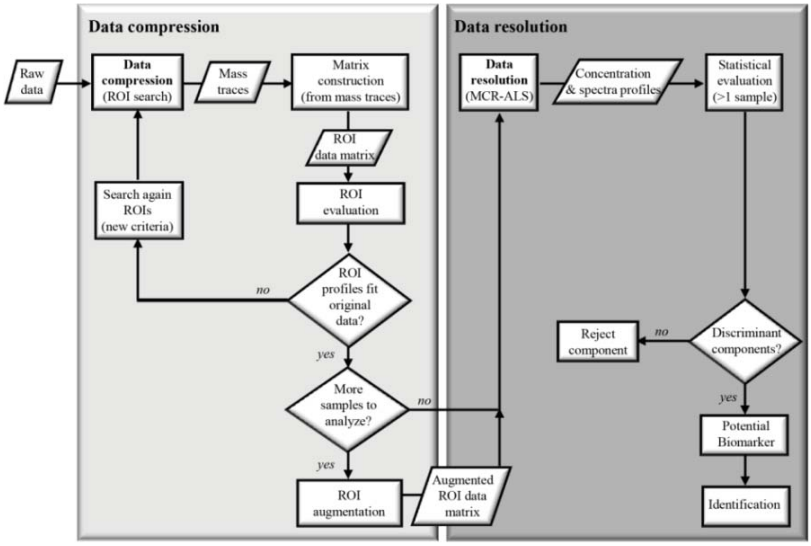


Fig. 1. Schematic representation of data compression and resolution stages of ROIMCR approach. Initially, raw data are compressed through the search of regions of interest (ROI) and the obtained mass traces are reorganized into a matrix representation. Then, ROI profiles are evaluated: if they do not fit original data, the ROI search is repeated but changing initial criteria; on the contrary, if they properly fit original data the obtained ROI matrix is resolved by MCR-ALS. When having more than one sample, following individual ROI searches, column-wise augmented ROI data matrices can be generated and finally analyzed by MCR-ALS. Results of MCR-ALS analysis can be subsequently evaluated by statistical tests to find more significant components in the differentiation among sample groups (*i.e.*, stressed groups *vs.* control groups).

2. Method

A description of the ROI methodology is provided here. In this manuscript, only a brief description corresponding to the MCR-ALS algorithm is presented, since it is a well-stablished chemometric method and its principles and basis have been already described in previous studies of the authors[18,19,20] and they can be found on its official webpage www.mcrals.info.

2.1. ROI search in one sample

The aim of the ROI compression is to scan for regions containing interesting mass traces, *i.e.*, regions that include data of a relevant MS intensity (bigger than a threshold value, **Fig. 2a**), enclosed in a particular mass error range (**Fig. 2b**) and constituted by a minimum number of occurrences (**Fig. 2c**).

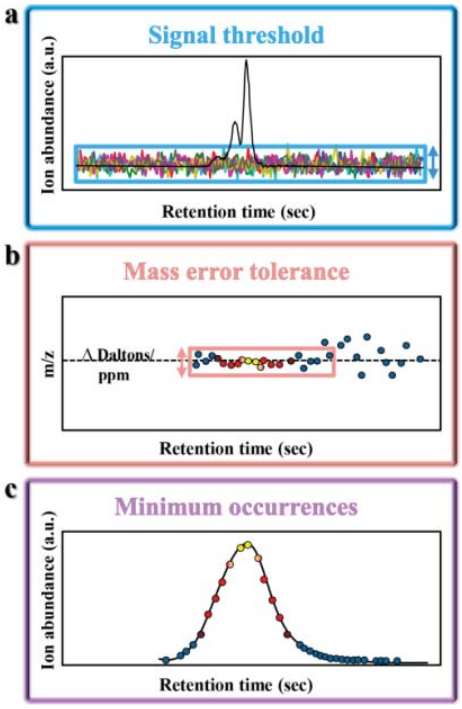


Fig. 2. Parameters necessary to define an ROI. a) Signal threshold, b) Mass error tolerance and c) Minimum occurrences.

These three parameters are the input variables necessary in one ROI search, together with a vector listing the retention times at which the instrument makes the measurements (variable "**time**" in **Fig. 3a**) and a cell array (*i.e.*, array containing data of varying types and sizes in the MATLAB environment) containing the *m/z* values and MS intensities at each retention time (variable "**peaks**" in **Fig. 3a**).

Interestingly, the *m/z* values (and their corresponding MS intensities) measured by the mass spectrometer at each retention time do not follow a regular pattern (*i.e.*, the *m/z* measurements are not equidistant and may differ among samples) and, therefore, the generated vectors enclosed in the cell array containing such information have distinct lengths.

In **Fig. 3a** a representation of the pairs of vectors (*i.e.*, one vector of the pair containing *m/z* values and the other containing MS intensities) including information of one LC-MS sample is shown. As it can be noticed, the length of these vectors varies at the distinct retention times, indicating that the mass spectrometer acquires distinct *m/z* values at each scan.

Once introduced the input parameters, the ROI algorithm performs the ROI search according to the following steps:

1. Search in the first scan for *m/z* values associated with MS intensities higher than a signal threshold value.
2. Search in the same scan clusters of *m/z* values enclosed within a particular mass admissible error.
3. Calculate the mean mass of all the *m/z* values classified inside the same cluster (mzroi).
4. Arrange mean mass values from the lowest to the highest value.
5. Repeat steps 1-4 for the rest of scans, merge and update the calculated mean mass values.
6. Select clusters having a minimum number of occurrences of *m/z* values.
7. Eliminate empty spaces in the final MSROI matrix, substituting them by random values with a threshold mean value, like for instance 1% of the threshold intensity value used in step 1.

The ROI search gives three outputs. A vector containing final mean *m/z* values of ROIs ("**mzroi**" in **Fig. 3b**), a newly arranged data matrix containing the MS spectra of every scan in its rows and the chromatograms of every ROI in its columns ("**MSROI**" in **Fig. 3b**) and a cell array ("**roicell**" in **Fig. 3b**) containing for each ROI, information of their constituting *m/z* values, retention times, MS intensities, scan numbers and the calculated mean *m/z* value.

2.2. ROI search in more than one sample

Since the main purpose of metabolomics is the study of the differences between metabolic profiles across multiple sample groups (*e.g.*, controls *vs.* exposed), final data analysis must consider all samples simultaneously. In fact, MCR-ALS analysis of multiple samples requires the construction of column-wise augmented data matrices, by organizing all samples one above each other. Building such matrices is only possible when dimensions in the *m/z*-mode of all individual data matrices are equal. However, data compression using

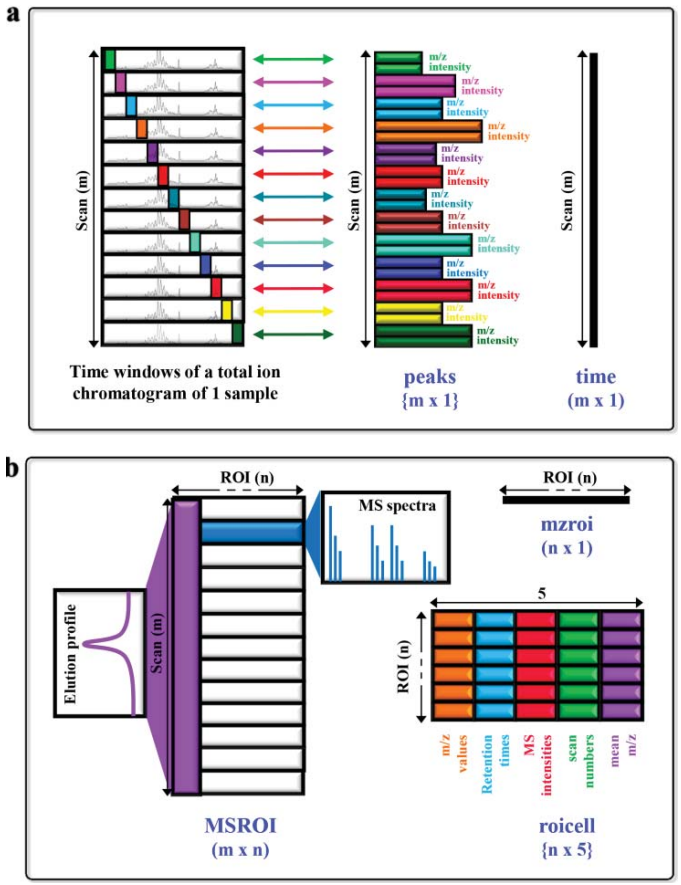


Fig. 3. Schematic illustration of input (a) and output variables (b) of an ROI search when using ROI compression algorithm. Data of the LC-MS chromatogram is described as a {m x 1} cell array (named as peaks), with m cells (equal to the number of retention times), each of them containing two vectors (of variable length among cells), corresponding to the *m/z* and intensity values acquired by the instrument at each of the retention times. Peaks and vector time (m x 1) are the input variables of ROI function together with the parameters required to define an ROI (thresh=750, mzerror=0.05 and minroi=10 are used in this example), resulting in a matrix, a vector and a cell array (MSROI, mzroi and roicell, respectively) after ROI search. nROI is the total number of ROIs obtained (in the example of the figure, nROI= 297). MSROI is a (m x nROI) matrix, containing the MS spectra of every retention time in its rows, and the chromatograms of every ROI in its columns, mzroi is a vector containing mean *m/z* values of ROIs and roicell is a {nROI x 5} cell array, containing nROI x 5 cells (in the example of the figure it would be 297 x 5=1485). Cells comprised in roicell variable from column 1 to column 4 contain single vectors in their structures (containing information of *m/z*, retention times, intensities and scan number of the data enclosed in the same ROI, respectively) whereas cells comprised in the fifth column (roicell {nROI,5}) contain single values (corresponding to mean *m/z* values of ROI).

ROI strategy produces data matrices of *m/z*-mode dimensions equal to the number of ROIs, which can vary between samples. Thus, a final unification of ROIs among samples, considering both common and uncommon mzroi values must be performed.

The following description of the ROI search among samples allows the construction of column-wise augmented data matrices ready for further MCR-ALS analysis. The search of ROI between several files is based on the determination of common and uncommon ROI values of the analyzed data matrices. The steps of the algorithm for ROI search and augmentation are presented here:

1. Check of the matching mzroi values among the several data matrices within +/- mzerror. Consider the new mzroi to be the average of them.
2. Build the corresponding column of the new augmented data matrix with MS intensity values of the coincident mzroi values (if more than one mzroi value is coincident, then consider the average of the MS intensity values).
3. Check for non-matching mzroi values; they are accepted if mzroi≥ tresh. For the non-coincident part, fill up empty values with random values 1% of the threshold intensity value.
4. Eliminate mzroi values that are not coincident and that neither have their MS intensity value higher than the threshold.
5. Reorganize the columns of the new augmented data matrix according to their new mzroi values, from lower to higher mzroi value.

Thus, to perform ROI augmentation, the required input information consists of the arrays of samples to be augmented containing *m/z* values (mzroi matrices) and MS intensities (MSROI matrices), the admissible mass deviation, the threshold intensity value and the vector containing the retention times that must be the same for all samples. The output variables consist of a vector containing final mean *m/z* values of common and uncommon ROIs, the final augmented ROI matrix containing compressed data of all the input files and a vector containing the total number of scans (*i.e.,* sum of the number of retention times of individual samples).

2.3. Multivariate curve resolution-alternating least squares (MCR-ALS)

MCR-ALS method performs a bilinear decomposition of raw data sets, under specific constraints, according to Eq. (1):

$$\mathbf{D}=\mathbf{C}\mathbf{S}^{T}+\mathbf{E} \quad (1)$$

In the equation, matrix **D** (*I* x *J*) exemplifies for instance the spectral data set coming from the output of a second-order instrument. Concerning LC-MS data, **D** matrix includes the MS spectra measured at all chromatographic retention times (*i*=1,… *I*) in its rows, and the elution profiles at the complete range of spectra *m/z* channels (*j*=1,… *J*) in its columns. This matrix is decomposed in the product of two small factor matrices, **C** and $\mathbf{S}^{T}$. The **C** (*I* x *N*) matrix encloses column vectors that agree with the concentration elution profiles of the *N* (*n*=1,… *N*) pure chemical constituents or components of matrix **D**. In $\mathbf{S}^{T}$ (*N* x *J*) matrix, row vectors correspond to the MS spectra of these N pure components. The fraction of **D** that is not described by the bilinear model constitutes the residual matrix, **E** (*I* x *J*). MCR-ALS methods suppose that the measured variance in all samples in the raw data set can be explained using a combination of a small number of chemically significant profiles. Regarding LC-MS data sets, the variance observed in the investigated data matrices is explained by the combination of a number of pure mass spectra (row profiles in the $\mathbf{S}^{T}$ matrix) weighted by the amount of each of them along the elution direction (the associated chromatographic elution peaks, column profiles in **C**).

2.4. MCR-ALS in parallel analysis of multiple samples

MCR-ALS can be implemented across distinct data sets or matrices at a time. For instance, in the case of the analysis of the simultaneous analysis of multiple samples by LC-MS, this is accomplished by generating column-wise data matrices ($\mathbf{D}_{aug}$) including different matrices related to distinct chromatographic runs appended one above each other. Therefore, the MS spectral (column) direction is the same for all them and the data matrix extent is augmented column-wisely in the chromatographic (rows) direction. Resolved pure mass spectra are comparable to all simultaneously analyzed chromatographic runs or experiments ($\mathbf{S}^{T}$) while elution profiles can vary from run to run (experiment to experiment), conforming $\mathbf{C}_{aug}$, as represented in Eq. (2):

$$\mathbf{D}_{aug}=\mathbf{C}_{aug}\mathbf{S}^{T}+\mathbf{E}_{aug} \quad (2)$$

In the MCR-ALS method, bilinear models described in Eq. (1) (single data matrix illustration) or Eq. (2) (augmented data matrix illustration) are resolved by means of an alternating least squares optimization under constraints[3]. When considering metabolomic LC-MS data, the minimum constrains to apply consist on non-negativity for concentration (elution), **C**, and spectra, $\mathbf{S}^{T}$, profiles, and normalization for the second. In this work, due to the sparcity nature of the MCR resolved elution profiles, and especially of MS spectra profiles, no additional constraints were required to achieve reliable results.In the proposed ROIMCR procedure, individual or augmented MSROI data matrices (**D** or $\mathbf{D}_{aug}$) are submitted to MCR-ALS analysis. The application of this method will give the concentration/elution, **C** (or $\mathbf{C}_{aug}$), and MS spectra, $\mathbf{S}^{T}$, profiles of the resolved components. Note that in the MCR-ALS procedure, elution profiles in $\mathbf{C}_{aug}$ are not required to be aligned nor shape modelled and that spectra profiles are the MSROI compressed spectra and have their full instrument mass accuracy.

More details about MCR-ALS method and implementation of different constraints can be found in previous publications[21].

3. Data set

The data set used for the illustration of the current methodology correspond to LC-MS data of extracted lipids of human placental choriocarcinoma cells (JEG-3) 24-h exposed to DMSO (vehicle controls) and to the chemical endocrine disruptor TBT (exposed samples) at a non-lethal dose. Both groups (*i.e.*, controls and exposed) contain three replicates. These data are available at http://cidtransfer.cid.csic.es/descarga.php?enlace1=d5e1de55b1d9b83b1668fc81e151e2ea so that the reader can use them to test the ROIMCR procedure presented here. For details regarding the characteristics of the data, readers are advised to consult https://www.nature.com/protocolexchange/protocols/4347.

4. Implementation

The ROI compression presented in this study has been implemented in the MATLAB command line using the functions available at http://cidtransfer.cid.csic.es/descarga.php?enlace1=3adcd456c13cef5c017b66269651912c. A user friendly graphical interface for ROI compression is under development. The provided MATLAB functions for ROI compression are related to: a) ROI search in one sample; b) evaluation of ROI profiles and c) generation of augmented ROI data matrices. In addition, the statistical evaluation of the concentration profiles obtained after MCR-ALS analysis may be performed. Regarding the implementation of MCR-ALS, its interface available at www.mcrals.info is utilized.

5. Results and discussion

Although the data used as example in the present protocol was also used in previous studies by the authors[16,17], the results hereby presented were not presented in the previous publications and are specifically selected now in order to demonstrate the key features of ROIMCR methodology. These results include data compression through ROI search in individual data matrices, data matrix augmentation through ROI search among data matrices and MCR-ALS analysis of the obtained augmented ROI matrix. The readers interested in the LC-MS data conversion and MATLAB import procedure, are advised to consult https://www.nature.com/protocolexchange/protocols/4347.

5.1. ROI search

5.1.1. Optimization of ROI parameters

As previously stated in the Methods section, some parameters need to be optimized previously to the search of ROIs. In the example shown in **Table 1**, the results of the ROI search are shown after fixing distinct values of one of the three input parameters, while maintaining the other two unchanged. In all cases, three distinct values are tested for parameter: 10 times higher the recommended value, the recommended value, and 10 times lower the suggested value. In the first case, where the influence of the threshold on ROI search is evaluated, the three options tested corresponded to threshold values of 7500, 750 and 75 a.u. The recommended threshold value should be adjusted between 0.1- 1% maximum MS measured intensity. Since the maximum MS measured intensity of the evaluated sample was $3.5118 \cdot 10^5$ a.u., the recommended threshold value would be between 351.18 and 3511.8 a.u. In particular, we selected an intermediate value of 750 as the optimum value. The higher and the lower values tested (7500 and 75 a.u., respectively) were chosen to clearly demonstrate that a decrease in the threshold value produces increasing numbers of ROI, together with a major computation time (in seconds), while an increase in threshold results in the opposite effects. Hence, the threshold value requires to be adjusted with prudence since it can enhance data quality by eliminating noise, but immoderate threshold values may cause information loss. In fact, this parameter should be better visually evaluated from the graphical output, to make sure that it results in noise diminution without signal deformation.

In the second case, the study of the admissible mass deviation on an ROI search, the three options tested corresponded to mzerror values of 0.5, 0.05 and 0.005 Da/e. The optimum mass deviation value should be halfway between an excessive and an insufficient mass accuracy.

Table 1. Number of ROIs and computation time resulting from ROI searches performed with three different values of the input parameters (signal threshold, mass error tolerance and minimum occurrences). In cursive are indicated the optimum values of the parameters. The results shown are obtained considering the variation of one parameter while the other two remain fixed in their optimum value.

Parameters of the ROI search		Number of ROI	Computational time[a] (s)
Signal threshold (a.u.)	7500	55	0.8
	750	*300*	*1.8*
	75	1357	8.8
Mass error tolerance (*Da/e*)	0.5	267	1.8
	0.05	*300*	*2.0*
	0.005	356	2.0
Minimum occurrences	100	23	1.7
	10	*300*	*1.9*
	1	449	1.9

[a] Computational time using a 64-bit Windows Intel(R) Core™ i5-3470 CPU computer of 8GB and version 8.2.0 (R2013b) of MATLAB.

In this example case it was observed that with an mzerror value of 0.005 *Da/e*, peaks corresponding to the same ion were divided into distinct parts, whereas for a value higher than 0.5 *Da/e*, the opposite situation occurred, and peaks corresponding to distinct ions were collapsed into the same chromatographic signal. Thus, the optimum mzerror value was set to 0.05 *Da/e*. The higher and lower values tested (0.5 and 0.005 *Da/e*, respectively) were again selected to easily visualize their effect on final ROI selection. As occurred with the threshold parameter, a decrease in mzerror value evolved in an increase in the number of ROIs. In this case, however, the growth in ROI number was not as spectacular as for the threshold parameter, and the elapsed computation time remained almost constant for all calculations (see **Table 1**). In the third case, evaluation of the minimum occurrences on an ROI search, the three values tested corresponded to 100, 10 and 1. The minimum number of occurrences is directly related to peak width range and detector speed, which varies among high performance liquid chromatography (HPLC) (20-50 seconds) and ultra-high performance liquid chromatography (UHPLC) (5-12 seconds) systems.

In the current showing case, the system used to analyze the sample was an Acquity UHPLC system, and thus, the optimum number of occurrences should correspond to a peak with a range of 5-12 seconds. In particular, we observed that with this instrumentation, the interval between each occurrence was 0.63 seconds, and thus, we selected 10 occurrences (*i.e.*, 6.3 seconds) as the optimum value. When considering results obtained for the three values tested, the same tendency observed for the other parameters was again evidenced, obtaining higher numbers of ROI when decreasing values of the minimum number of occurrences and lower numbers of ROI when increasing it. As for mzerror parameter, an increase in ROI number at lower minimum number of occurrences was less considerable than for the threshold parameter, and the elapsed computational time was practically the same in the three calculations (see **Table 1**).

The hereby presented example clearly illustrates the importance of proper optimization of ROI parameters before the application of the method. It also highlights the influence of the particular instrumental specifications (*e.g.*, mass accuracy) on these parameters.

5.1.2. Evaluation of ROI profiles

After the ROI search in individual matrices, their profiles were evaluated in order to see whether they fit the chromatographic shape of the original data or not.
In **Fig. 4** the two distinct graphical representations of three ROIs obtained after an ROI compression in Control 1 sample are shown. The three selected ROI correspond to the *m/z* values of 703.5740 *Da/e* (**Fig. 4a**), 271.1875 *Da/e* (**Fig. 4b**) and 391.2841 *Da/e* (**Fig. 4c**). The selected ROI demonstrate three completely distinct elution profiles and related mass distributions. In the first case (**Fig. 4a**), the elution profile of the ROI with *m/z* 703.5740 *Da/e* describes a single-peak curve and the corresponding mass distribution is appreciably regular among time.

The second case (**Fig. 4b**) corresponding to ROI with *m/z* 271.1875 *Da/e* is particularly interesting since it describes a double-peak curve. As observed in the mass spectrum of this ROI, three slightly distinguishable regions of mass measurements are presented, corresponding to the initial measurements of the profile curve, first peak and second peak. This case would correspond for instance to different isomeric chemical compounds resolved by the chromatographic column, but having equal *m/z* value at the considered mass deviation.
Finally, in the third case (**Fig. 4c**), the elution profile of ROI with *m/z* 391.2841 *Da/e* distinguishes two clusters of MS points. The first cluster, located around 200 seconds is associated with the chromatographic peak whereas the second cluster, located between 600 and 1200 seconds is related to the background noise. The representations of mass traces provide valuable information of the nature of experimental MS measurements. In general, such information is unknown by MS users and can be crucial for a better analysis and interpretation of LC-MS data. Once selected the optimum parameters for the ROI search, the augmentation was performed and a final augmented ROI matrix was generated. The dimensions of that matrix were (11394 x 481), the x-dimension corresponding to six times the number of retention times of one sample (*i.e.*, 1899) and the y-dimension corresponding to the total number of common and uncommon ROIs among the six samples.

5.2. Data resolution through MCR-ALS analysis

Once the augmented data matrix of compressed data of the six samples has been constructed the next required step is the analysis by MCR-ALS.

In MCR-ALS analysis, the selection of the number of pure components is the first crucial step. The optimum number of MCR-ALS components should be high enough to explain all the chromatographic peaks but also background (*e.g.*, solvent), noise and other unknown signal contributions. Moreover, an increase in the number of components should produce a diminution of the lack of fit and the corresponding increase of the explained variance. Otherwise, no more components should be added to the calculation. In the presented example, the selected number of components was proposed to be 50 for the MCR-ALS analysis of the augmented matrix, resulting in a percentage of lack of fit lower than the 7% and an explained variance of 96.5%.
It is important to mention however, that due to the sparse condition of MS spectra, their resolution has little ambiguity[21,22] and the underestimation of the number of MCR-ALS components will not cause misinterpretation of the results but only a loss of information. In that case, final interpretation will only be given for the finally resolved components.

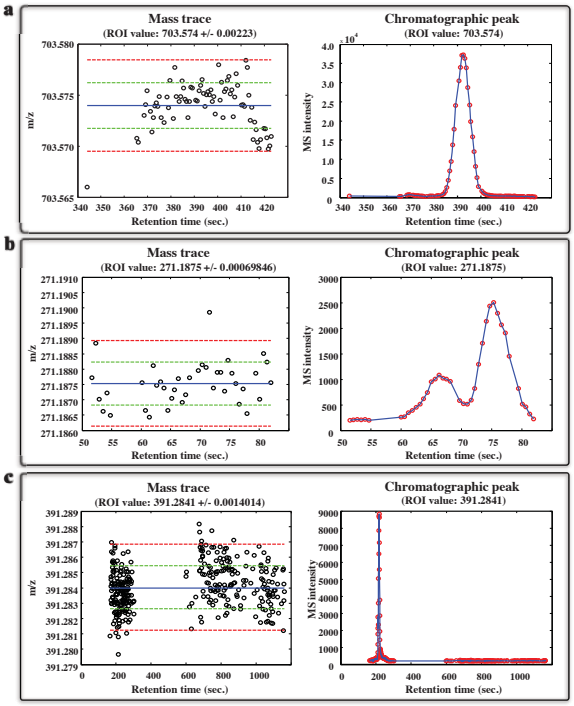


Fig. 4. Representation of the chromatographic elution profiles and mass traces of three ROI values: a) one ROI with a single-peak elution profile, b) one ROI with a double-peak elution profile and c) one ROI with an elution profile with two clearly distinguished regions, corresponding to a defined peak and the baseline/background noise.

3. Biomarker discovery

Information of concentration and spectra profiles of the resolved MCR-ALS components is finally used for biomarker assessment. However, in order to find most relevant MCR-ALS components (*i.e.*, the ones that significantly vary among control and stressed samples), further statistical evaluation is required. Distinct statistical tests can be used for this evaluation, such as the classical *Student's t*-test, used in this study. This test, together with other statistical tests may be performed using the functions and protocol[18] available at http://cidtransfer.cid.csic.es/descarga.php?enlace1=3adcd456c13cef5c017b66269651912c.

In **Fig. 5**, a representation of the elution and spectra profiles of three representative MCR-ALS resolved components is shown. As it can be observed in the elution profiles of these components (**Fig. 5a**), there is a noticeable difference in the areas and heights of the chromatographic peaks among control and exposed samples. Such difference indicates an up-regulation of these lipids after the treatment with TBT.

In order to evaluate the significance of such alteration a classical statistical *Student's t*-test was performed component by component, using as a criterion a *p*-value lower than 0.05. The results of the test reflect that the three components showed significant changes in their heights among the two groups (*i.e.*, controls and exposed), which made them potential biomarkers for TBT exposure. When needed, multiple comparisons procedures (MCPs)[23] can be applied to avoid the assignation of false positives. These statistical procedures are intended to consider and suitably manage the multiplicity effects through some shared or joint measure of mistaken inferences. Alternatively, ANOVA and its multivariate extensions for well-designed data can be applied[24,25] to better ascertain the reliability of the observed effects by TBT exposure. Also, the fold-changes for the three components were calculated (**Fig. 5a)** resulting in 3.5-fold, 4.5-fold and 4.0-fold for components A, B and C, respectively. In order to identify the lipid species corresponding to these MCR-ALS components, their MS spectra profiles were evaluated. As shown in **Fig. 5b**, the exact masses associated with components A, B and C were 872.7702, 874.7857 and 902.8171 *Da/e*, respectively. Further identification using MS databases such as Lipid Maps (http://www.lipidmaps.org) was possible. As shown in the same figure, components A, B and C corresponded to triacylglycerol species 52:4, 52:3 and 54:3, respectively. It is important to stand out that such identification was possible, in a higher extent, thanks to the fact that no loss of mass spectral information occurred after ROI compression.

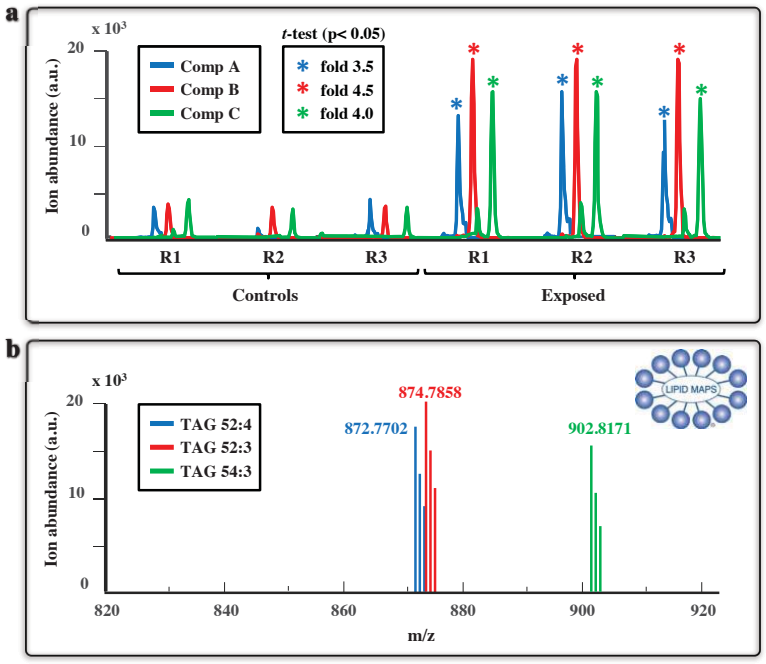


Fig. 5. Elution profile (a) and mass spectra (b) of three resolved MCR-ALS components. * Statistical significant differences respect to the control (p-value < 0.05). Lipid Maps was the MS reference database used for the identification of lipids.

6. Conclusions

The chemometric LC-MS data analysis strategy proposed in this study based on the ROI compression followed by MCR-ALS analysis has been shown to be a powerful approach to analyze LC-MS metabolomic data sets. On the one hand, the principal benefit of performing an ROI compression is the capacity to minimize the primary dimensions of the data (gigabytes of storage) whilst escaping from any loss of spectral accuracy. On the other hand, the main advantages attributed to MCR-ALS analysis include: i) the possibility of immediate chemical identification of the metabolites thanks to the MS information provided in the analysis; ii) the high degree of interpretability of the results; iii) the flexibility in the structure and nature of the data sets that can be potentially analyzed and iv) the added value as a preprocessing method, that does not require peak shaping nor chromatographic alignment for the simultaneous analysis of multiple samples.

Abbreviations

CWT: Continuous wavelet transformations; **HPLC**: High performance liquid chromatography; **LC-MS**: Liquid chromatography coupled to mass spectrometry; **MCR-ALS**: Multivariate Curve Resolution-Alternating Least Squares; **MS**: Mass spectrometry; **ROI**: Regions of interest; **TBT**: Tributyltin; **UHPLC**: Ultra-high performance liquid chromatography.

Declarations

Funding

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 320737. The first author acknowledges the Spanish Government (Ministerio de Educación, Cultura y Deporte) for a predoctoral FPU scholarship. Finally, all the authors acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

Availability of data and materials

ROIMCR has been implemented as a MATLAB package. ROIMCR MATLAB functions are available at http://cidtransfer.cid.csic.es/descarga.php?enlace1=3adcd456c13cef5c017b66269651912c. Example data to test ROIMCR procedure are available at http://cidtransfer.cid.csic.es/descarga.php?enlace1=d5e1de55b1d9b83b1668fc81e151e2ea.

Authors' contributions

EG wrote the manuscript, acquired the data used to test the methodology and participated in the evaluation of the efficacy of the strategy. JJ made substantial contributions in the development of the algorithm and revised the manuscript critically for important intellectual content. RT designed the ROIMCR algorithm and provided guidance on the implementation and the design of experiments. All the authors read, contributed to and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

(1) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends Anal. Chem.* **2016**, *82*, 425–442.

(2) Wen, B.; Mei, Z.; Zeng, C.; Liu, S. metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformatics* **2017**, *18* (1), 183.

(3) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **2012**, *84* (11), 5035–5039.

(4) Http://kwanjeeraw.github.io/metabox/. Metabox by kwanjeeraw http://kwanjeeraw.github.io/metabox/ (accessed Jun 7, 2017).

(5) Http://metax.genomics.cn/. Welcome to metaX homepage! http://metax.genomics.cn/ (accessed Jun 7, 2017).

(6) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **2012**, *84* (1), 283–289.

(7) Fernández-Albert, F.; Llorach, R.; Andrés-Lacueva, C.; Perera, A. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics* **2014**, *30* (13), 1937–1939.

(8) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W652-60.

(9) Giacomoni, F.; Le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015**, *31* (9), 1493–1495.

(10) Katajamaa, M.; Miettinen, J.; Oresic, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22* (5), 634–636.

(11) Http://www.metalign.nl. No Title http://www.metalign.nl.

(12) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **2008**, *9* (1), 504.

(13) Stolt, R.; Torgrip, R. J. O.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. Second-order peak detection for multicomponent high-resolution LC/MS data. *Anal. Chem.* **2006**, *78* (4), 975–983.

(14) Tikunov, Y.; Lommen, A.; de Vos, C. H. R.; Verhoeven, H. A.; Bino, R. J.; Hall, R. D.; Bovy, A. G. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **2005**, *139* (3), 1125–1137.

(15) Tauler, R. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.* **1995**, *30* (1), 133–146.

(16) Gorrochategui, E.; Casas, J.; Pérez-Albaladejo, E.; Jáuregui, O.; Porte, C.; Lacorte, S. Characterization of complex lipid mixtures in contaminant exposed JEG-3 cells using liquid chromatography and high-resolution mass spectrometry. *Environ. Sci. Pollut. Res. Int.* **2014**, *21* (20), 11907–11916.

(17) Gorrochategui, E.; Jaumot, J.; Tauler, R. A protocol for LC-MS metabolomic data processing using chemometric tools. *Protoc. Exch.* **2015**, doi:10.1038/protex.2015.102.

(18) Jaumot, J.; Gargallo, R.; de Juan, A.; Tauler, R. *A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB*; 2005; Vol. 76.

(19) de Juan, A.; Tauler, R. Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. *Crit. Rev. Anal. Chem.* **2006**, *36* (3–4), 163–176.

(20) de Juan, A.; Jaumot, J.; Tauler, R.; Tauler, R.; Neymeyr, K.; Werner-Washburne, M. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods* **2014**, *6* (14), 4964.

(21) Tauler, R.; Smilde, A.; Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.* **1995**, *9* (1), 31–58.

(22) Cook, D. W.; Rutan, S. C. Analysis of Liquid Chromatography–Mass Spectrometry Data with an Elastic Net Multivariate Curve Resolution Strategy for Sparse Spectral Recovery. *Anal. Chem.* **2017**, acs.analchem.7b01832.

(23) Hochberg, Y.; Tamhane, A. C. *Multiple Comparison Procedures*; John Wiley & Sons Ltd, 1987.

(24) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R.-J. A. N.; van der Greef, J.; Timmerman, M. E. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21* (13), 3043–3048.

(25) Engel, J.; Blanchet, L.; Bloemen, B.; van den Heuvel, L. P.; Engelke, U. H. F.; Wevers, R. A.; Buydens, L. M. C. Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Anal. Chim. Acta* **2015**, *899*, 1–12.

2.3. DISCUSSION OF RESULTS: Setting the basis for untargeted omic data analysis

This section of the manuscript does not respond to a conventional results' discussion section since the research included in this Chapter is basically theoretical and methodological and aims to provide the data analysis strategy necessary to generate omic results (further discussed in Chapters III and IV). For this reason, the main purposes of this section of the manuscript are: first, to give extensive details of the fundamentals of the developed methodology (*e.g.*, crucial steps and some tips that can be useful for other researchers in the field) and secondly, to put it into context with what is found in the literature regarding chemometrics for untargeted omic data analysis.

Open-formats are required to analyse LC-MS data files outside vendor packages

As demonstrated in this Thesis, the analysis of LC-MS data sets outside the software of the MS vendors demands the conversion of the original formats of the data into open-formats. There is a variety of open-formats that can be used; some of the most accepted are shown in TABLE 2.1.

TABLE 2.1. Mass Spectrometry most popular open data formats

- **JCAMP-DX:** standardized ASCII based file format for data exchange in mass spectrometry initially developed for infrared spectroscopy. This format was found impractical for the large MS data sets generated nowadays, but it is still used for exchanging moderate numbers of spectra[402].
- **ANDI-MS or netCDF:** Analytical Data Interchange for Mass Spectrometry (ANDI-MS) is a format for exchanging data readable and writable for many MS software packages. ANDI files, initially developed for chromatography-MS data, are based on netCDF which is a software tool library for writing and reading data files[403].
- **mzData:** first attempt of standardized format for MS data now replaced by mzXML[404].
- **mzXML:** eXtensible Markup Language based common file format for proteomics MS data developed simultaneously to mzData format that is still in use by the proteomics community[405].
- **mzML:** unified standard format borrowing best aspects of mzData and mzXML formats intended to replace them[406].

Apart from the variety of open-formats, a large diversity of vendor-formats exists. As it can be observed in FIGURE 2.1, each LC-MS vendor generates its own particular data files (*e.g.*, *Waters* generates *Masslynx.raw* files whereas *AB SCIEX* generates *WIFF* files), that need to

be further transformed into open-formats. To do that, most of the commercial packages provide specific tools for the conversion, some of them are also mentioned in FIGURE 2.1.

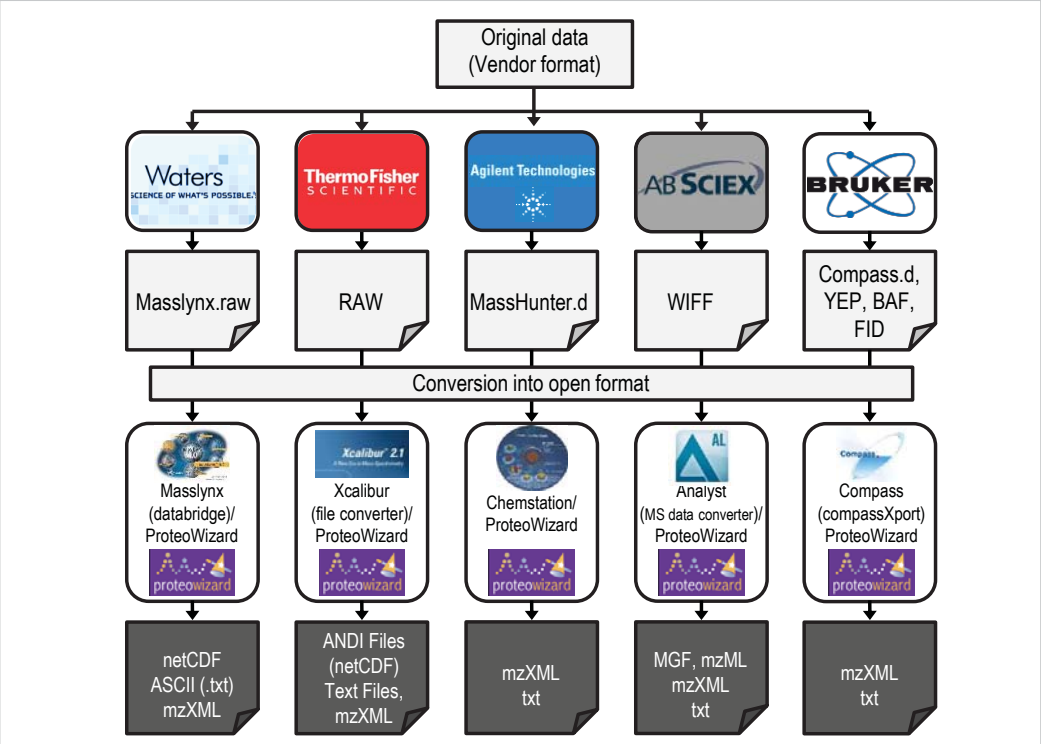


FIGURE 2.1

Schematic representation of the data conversion process for the most popular LC-MS vendors: Waters, Thermo Fisher, Agilent Technologies, AB SCIEX and Bruker Corporation.

In the protocol for LC-MS data analysis developed in this Thesis, the steps involved in the conversion of data generated using *Waters* and *Thermo Fisher* instrumentation are provided (Scientific article II, Section 2.2.1). *Waters* Corporation provides a specific tool named *Databridge* in the software *Masslynx* that allows for data conversion and *Thermo Fisher Scientific* provides a tool named *File converter*, which is included in the software *Xcalibur*. Furthermore, an alternative strategy that allows for data conversion, without the necessity of using the packages provided by LC-MS vendors, consisting on the use of an external software name *ProteoWizard*[321,322] is also presented. Few studies are found in the literature that address the issue related with data conversion, because the vast majority of scientific researchers that perform omic studies use external data analysis packages that already include an option for data conversion. One of the few studies that explored data conversion of

mass spectrometry data was the one of Holman, J.D. *et al*.[407] In that study the authors provided three protocols showing the details of the use of *ProteoWizard* software for data conversion, taking format features, coding options, and vendor particularities into account. First protocol called "*Transcoding MS data from raw format via MSCONVERT GUI*" was intended for first-time users of *ProteoWizard* who feel most comfortable with graphical user interfaces. Second protocol called "*Transcoding MS data from raw format via MSCONVERT*" was intended to assist researchers who are comfortable in a command-line environment. Third protocol called "*Converting mzML data to simple text formats for search engines*" was intended to assist researchers who need to convey their data to search engines requiring simpler text formats.

***R* and *MATLAB*: most popular computing platforms for metabolomics**

One of the most popular environments for computing omics data is *R*, due to the fact that it is an open source software, free to all users[408]. In fact, XCMS (various forms (X) of chromatography mass spectrometry), which is one of the most popular data analysis platforms for LC-MS metabolomic data sets, is an R-based software. XCMS, as with any R-based package, it is command line driven and demands some basic knowledge of the R programming language. Other open source tools that can be used to analyse omic data include Java (MZmine 2[409] software uses Java platform), C/C++ and MATLAB, the latter platform is worldwide used in the Chemometrics field and it is the one selected in this Thesis. Although less extensively used in omic studies than *R* language, MATLAB platform has also been used as the computer and visualization environment in some omic studies. For instance, the authors Arakelyan, A. *et al*.[410] used MATLAB to develop an algorithm for assessment of pathway activity changes and also created a KEGGParser tool for parsing, editing, and visualizing KEGG pathway maps. In that study the authors demonstrated the powerful computing environment and the huge variety of sophisticated toolboxes that MATLAB provides for performing complex bioinformatic calculations, as opposed to other computing languages such as C/C++ and Java, that are not flexible enough to allow for quick development and testing of new algorithms for omics. In the same study, the authors used the MATLAB Bioinformatics toolbox to develop their algorithms. Similarly, in this Thesis we utilized MATLAB Bioinformatics toolbox and its package *Mass Spectrometry and Bioanalytics* to

import LC-MS data (already in open-format). This package contains specific tools for data import, pre-processing and spectrum and signal analysis. In the protocol developed in this Thesis (Scientific article II, Section 2.2.1) the data import tool, containing three functions named *mzcdinfo*, *mzcdfread* and *mzcdf2peaks* was used. A description of these functions is provided in TABLE 2.2. Overall, the data generated after the use of these functions contain a cell array named **peaks**, that contains *m/z* and ion intensity values at each of the scans and a vector named **time**, providing information of the retention times associated with the LC-MS data set.

TABLE 2.2. MS data import functions of the MATLAB Bioinformatics Toolbox™

- ***mzcdfinfo or mzxmlinfo:*** this function extracts the information of the *netCDF* or *mzXML* files, returning a MATLAB structure, named *InfoStruct*, containing the name of the file, the date time stamp of the file, the size of the file in bytes, the number of scans, the run start and end times, the units for time, the minimum and maximum *m/z* values in all scans, the minimum and maximum intensity values in all scans, and the nature of the MS data (*i.e.,* profile or centroid).
- ***mzcdfread or mzxmlread*:** this function reads MS data from the *netCDF* or *mzXML* files and gives as an output argument a MATLAB structure containing the MS information. If a *netCDF* or *mzXML* variable contains local attributes, an additional field is created, with the name of the field being the variable name appended with the attributes string. The number and names of the fields will vary, depending on the mass spectrometer software, but typically there are mass values and intensity values fields.
- ***mzcdf2peaks or mzxml2peaks:*** this function extracts peak information from the MATLAB structure *mzCDFStruct or mzXMLStruct* created by the *mzcdfread* or *mzxmlread* function. An array of matrices containing *m/z* values and ion intensity values is created, named *peaks*, together with a scalar or vector of retention times associated with a LC-MS data set, named *time*.

Regions of interest: best choice for LC-MS data compression

As extensively stated in this Thesis, next step that needs to be covered in the overall data analysis strategy is data compression. Among the different data compression strategies that can be used (*e.g.*, binning, time windowing, wavelets) in this Thesis data compression based on the search of regions of interest (ROI) was selected due to its capacity to reduce data size without the loss of spectral accuracy. This type of compression was first used in a feature detection algorithm named *centWave*[411] of the data analysis platform XCMS. In this Thesis,

the fundamentals of ROI compression were taken from that algorithm and implemented into MATLAB but with some modifications. In particular, in contrast to the *centWave* algorithm, no continuous wavelet transformation (CWT) neither Gauss-fitting were applied for LC-MS peak modelling steps. The developed function to perform ROI compression was named *ROIpeaks* (Supplementary function provided in Scientific article II, Section 2.2.1). Resulting from the ROI compression using *ROIpeaks* function, three output variables were obtained in MATLAB workspace: a vector containing final mean *m/z* values of ROIs (**mzroi**), a new data matrix including the MS spectra of every scan in its rows and the chromatograms of every ROI in its columns (**MSROI**), and a cell array (**roicell**), providing for each ROI, information of their founding *m/z* values, retention times, MS intensities, scan numbers and the calculated *m/z* value. In FIGURE 2.2, a schematic representation of the steps involved in the import of the data and in their compression and data matrix construction is shown. In this figure, the Bioinformatics functions used and the function developed in this Thesis for ROI compression are mentioned, together with the corresponding input and output variables.

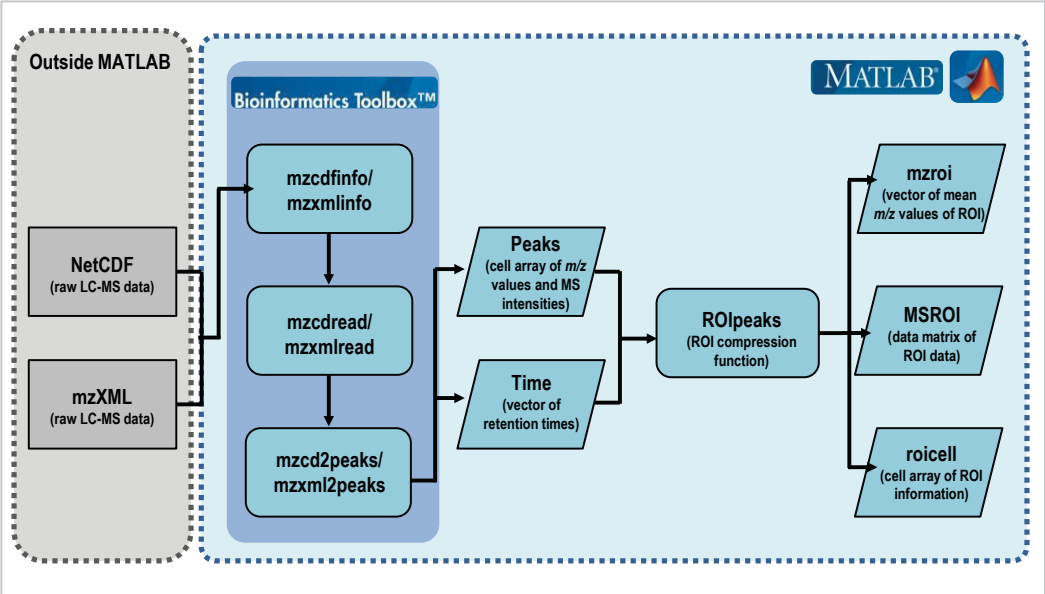


FIGURE 2.2

Overview flowchart listing the steps involved in the import of data to MATLAB environment and their further ROI compression. In this figure, rectangles indicate LC-MS files in open-format, rounded rectangles indicate MATLAB functions and parallelograms indicate data matrices, cell arrays or vectors. Grey-shaded area corresponds to the steps involved outside MATLAB platform, whereas blue-shaded area involves MATLAB environment.

ROI compression requires initial optimization of parameter settings

A crucial step regarding ROI search is the optimization of the parameter settings. In this Thesis, the three parameters used to define a ROI included: i) threshold (chromatographic signal-to-noise threshold, to filter out undesired background noise and other non-desired contributions such as those of the solvent), ii) mzerror (admissible mass deviation) and iii) minimum number of occurrences (*i.e.*, retention times) to define a chromatographic peak. These parameters were optimized (Scientific article III, Section 2.2.2) for an LC-TOF-MS data set generated in a lipidomic study and the optimum parameters suggested a threshold between 0.1-1% maximum MS intensity, an mzerror of 0.05 *Da/e* when working with a TOF mass spectrometer (an instrument with a resolution of 11,500 FWHM at *m/z* 556) and a minimum number of 10 consecutive measurements to define a chromatographic UHPLC peak. In that optimization, it was observed that the number of ROIs increased with decreasing values of threshold, mzerror and minimum number of occurrences. Resulting from the optimization of the threshold, it was evidenced that the selection of a proper threshold value resulted in enhanced quality of the data by the elimination of noise, whereas the selection of excessively high threshold produced information loss (Scientific article III, Section 2.2.2). These two situations are exemplified in **FIGURE 2.3**, in which a small window of an UHPLC-TOF-MS chromatogram initially containing chromatographic peaks and background noise is compressed, by selecting two different threshold values. Selection of the optimum threshold improves the condition of LC-MS data. In fact, as shown in **FIGURE 2.3b**, some low-intensity signals initially hidden behind the background noise, stand out after filtering out the solvent and noise contributions. However, when an excessive threshold filter is applied (**FIGURE 2.3c**), these low-intensity signals and other meaningful chromatographic peaks are lost. Such loss is particularly evident in the dotted areas represented in **FIGURES 2.3 b** and **c**.Subsequently, the optimization of the mzerror evidenced that this parameter needs to be optimized for a particular mass spectrometer whereas the optimization of the minimum number of occurrences was strongly related with the type of chromatography used (*i.e.*, UHPLC vs. HPLC) (Scientific article III, Section 2.2.2). The importance of optimizing the input parameters to perform a ROI compression was evidenced in our study and it is also a matter of concern of researchers that use automated data analysis platforms. This is because, as evidenced in this Thesis, the suboptimal setting of the parameters in a ROI compression can easily lead to

biased results. In fact, in order to deal with the difficulty of selecting optimum ROI parameters, the inventors of XCMS software developed a tool for automated optimization of peak picking parameters named IPO (Isotopologue Parameter Optimization)[412], which allowed the optimization of retention time correction and grouping parameters. Retention time correction was optimized by minimizing the relative retention time variances within peak groups. Also, grouping parameters were optimized by maximizing the number of combinations of peaks (*i.e.*, groups of peaks) from different samples showing similar masses and retention times.

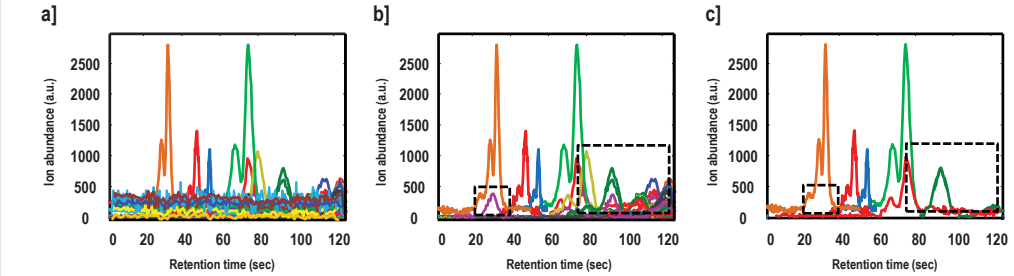


FIGURE 2.3

Chromatographic window containing information of a) non-compressed data, b) compressed ROI data when using the optimum threshold value and c) compressed ROI data when using excessive threshold value. Dotted squares indicate the regions of the chromatogram showing higher differences between conditions b) and c).

Evaluation of ROI profiles is necessary to optimize parameter settings and to point out potential outliers

In this Thesis, in order to evaluate the generated compressed data to see whether they properly reproduce the patterns of the original data or not, a function named *ROIplot* was also elaborated (Supplementary function provided in Scientific article II, Section 2.2.1). Incorrect compression of original data may include halving peaks or signals not corresponding to a real chromatographic peak, or including more than one distinct chromatographic peak in the same ROI. In FIGURE 2.4 five distinct ROI profiles (and their mass traces) obtained when acquiring data with a TOF mass spectrometer are represented. The first of them (FIGURE 2.4a) corresponds to a chromatographic peak practically having a symmetrical shape. The second ROI profile (FIGURE 2.4b) also corresponds to a chromatographic peak, but in this case with a less symmetrical shape. This example of a ROI profile is particularly interesting due to the fact

that the non-symmetrical shape of the peak would lead to "peak shaping" to adjust it to a Gaussian curve when performing data analysis by most of data analysis software. Again, this is the case of the *centWave* algorithm of XCMS[411], that uses CWT to perform the modelling of the peak shapes (shaping), by detecting chromatographic peaks of differing width and by fitting a theoretical model (*e.g.*, Gaussian or Lorentzian). A search in the literature shows that a lot of empirical shape models have been developed for the correction of asymmetric chromatographic peaks. Most of them are summarized by Di Marco and Bombi[413] and among them, the most important chromatographic correction functions include the Exponentially Modified Gaussian, the Poisson, the Log-normal, the Edgeworth/Cramér series and the Gram/Charlier series. The reason why most LC-MS data analysis software require peak shaping is related to the fact that they do not perform resolution of the chromatographic profile. Therefore, the calculation of the areas of the chromatographic peaks is achieved in a univariate way (*i.e.*, to each single *m/z* measurement), and peak shape modelling is necessary to facilitate peak alignment among samples and further calculation of their areas. Contrarily, when using the methodology developed in this Thesis, no shaping correction is applied (since it is not needed, see below). The third ROI profile shown in **FIGURE 2.4c** corresponds to a halving peak whereas the fourth ROI profile (**FIGURE 2.4d**) evidence an example of two close peaks considered within the same ROI due to their close *m/z* values. Finally, the ROI profile shown in **FIGURE 2.4e** evidences an example of a profiles not corresponding to real chromatographic peak. When only a small minority of the total amount of generated ROIs has a non-chromatographic shape, the ROIs that not fit chromatographic peaks may be considered as "outliers" and may not be contemplated (*i.e.*, they are supressed) for further steps of the analysis (*i.e.*, peak resolution). If however they are included, they can be filtered out afterwards during the ROIMCR procedure (see below). In case the number of ROIs not showing chromatographic shape is considerably high, the ROI search can be performed again, changing the settings of the input parameters (*i.e.*, snthreshold, mzerror and minimum number of occurrences). Finally, another problem regarding the analysis of LC-MS metabolomics (and lipidomic) data sets frequently encountered is the strong overlap in retention time that may occur with those chromatographic peaks that share *m/z* values. Such problem necessitates the development of multivariate resolution procedures (see below) procedures.

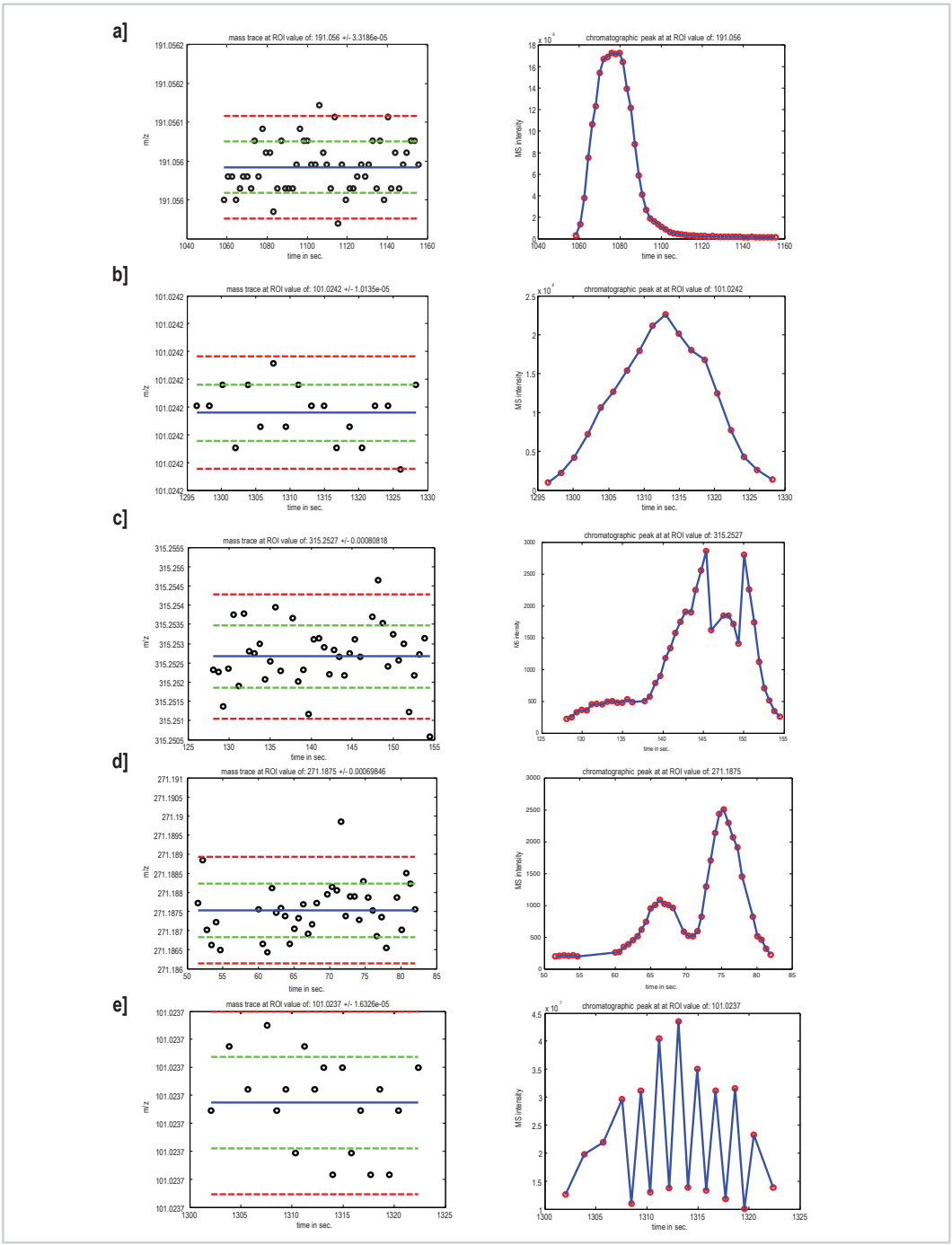


FIGURE 2.4

Mass traces and chromatographic profiles of four distinct ROIs, corresponding to: a) symmetrical chromatographic peak, b) non-symmetrical chromatographic peak, c) halving peak, d) double peak and e) strange feature not corresponding to a chromatographic peak.

Augmented ROI matrices provide compressed information of more than one sample useful to perform further comparative analyses

In this Thesis, however, previous to the resolution of LC-MS data, a function called *MSroiaug* was elaborated in order to compress more than one data file at a time and to generate compressed augmented data matrices (Supplementary function provided in Scientific article II, Section 2.2.1), which were later on resolved. The output variables of *MSroiaug* function consist of a vector containing final mean *m/z* values of common and uncommon ROIs (**mzroi_aug**), the final augmented ROI matrix containing compressed data of all the input files (**MSroi_aug**) and a vector containing the total number of the scans (**time_aug**). In FIGURE 2.5, a schematic representation of the input and output variables involved in a ROI search in two samples is shown.

ROIMCR: MCR-ALS applied to MSROI data, a chemometric tool that enables successful peak and spectra resolution in LC-MS omics studies

In this Thesis, multivariate curve resolution with the MCR-ALS procedure[414] allowed obtaining the purest elution and spectra profiles of the different constituents present in the LC-MS analysed samples. One of the main advantages of performing peak resolution through MCR-ALS analysis was related to the no need of peak alignment, since samples are aligned in the spectral dimension, no matter if they have differential time dimensions. In fact, this is one of the characteristics that mostly distinguish this data analysis strategy from the others existing in the literature, since most of them require peak alignment. A large number of metabolomic studies that require peak alignment methods to deal with within- and between-experiment variation are found in the literature. Moreover, the existing alignment methods can be classified in two major types: profile-based and feature-based alignment methods[415–417]. Profile-based methods perform alignment before peak detection and use the raw eluting chromatograms[418]. The feature-based methods perform alignment after peak detection. XCMS includes these two types of alignment methods, *OBI-warp (Ordered bijective interpolated warping)* and *peakgroups*, for alignment of profile matrix and features (peak groups), respectively[419,420]. Moreover, most of the existing feature-based alignment methods use reference variables, such as exogeneous internal standards and endogeneous metabolites[421–423], as landmarks for further retention time correction.

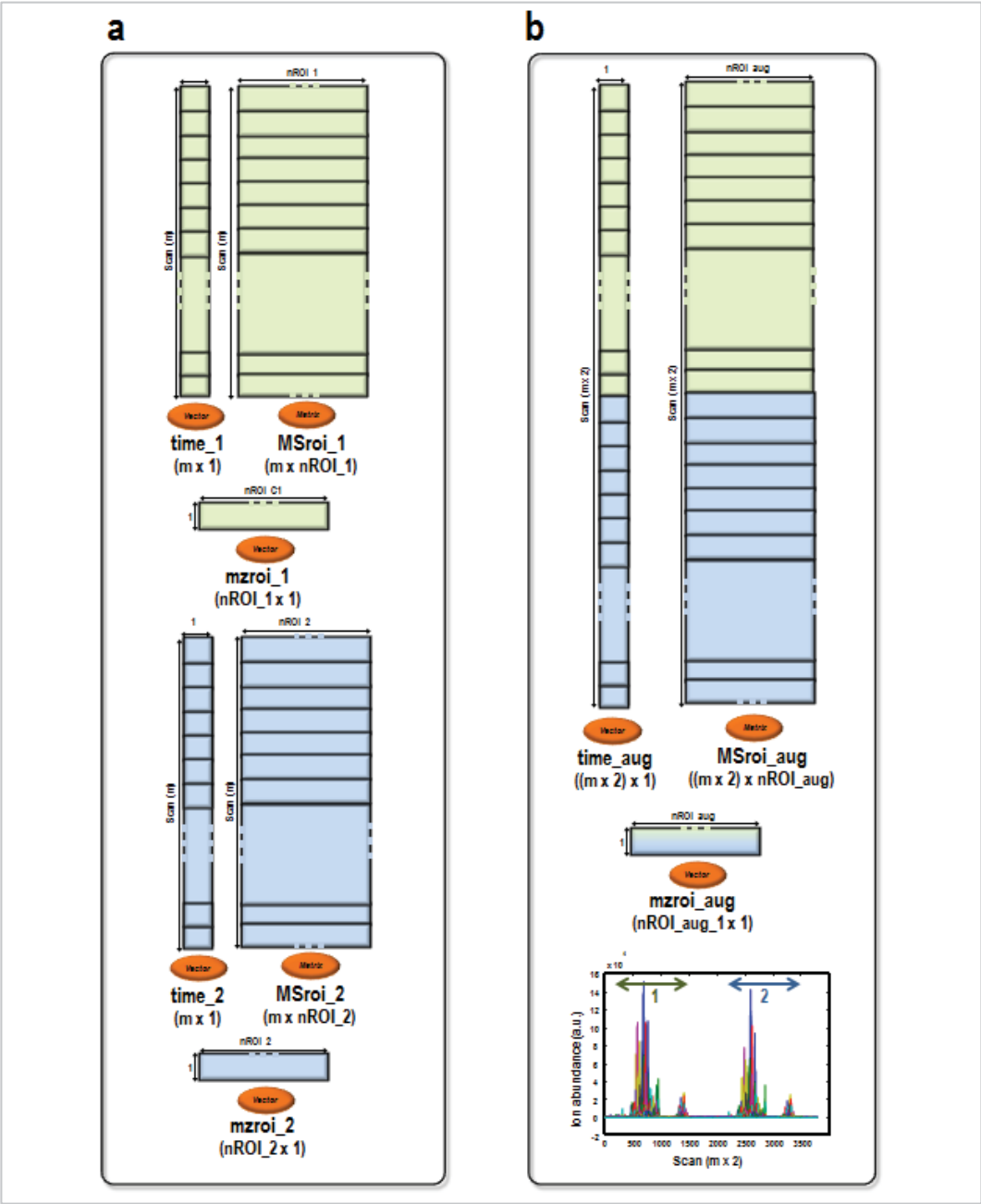


FIGURE 2.5

Schematic illustration of input (a) and output (b) variables of a ROI search in two data matrices, when using *MSroiaug* function. Output variables of ROI search of each individual data matrix (time_1, MSroi_1, mzroi_1, time_2, MSroi_2 and mzroi_2) are used as the input variables of this search, obtaining a column-wise augmented MS matrix (Msroi_aug), a time vector double-length its orifinal size (Msroi_aug) and a new mzroi vector containing *m/z* values of common and uncommon ROIs (mzroi_aug).

MCR-ALS spectra and elution profiles allow identification and relative quantification of metabolites (or lipids)

Last step included in the LC-MS data analysis protocol developed in this Thesis included the evaluation of both elution and spectra profiles of the MCR-ALS components. Evaluation of the concentration/elution profiles allowed the determination of the sample constituents showing significant alterations in exposed samples respect to controls (*i.e.*, potential biomarkers) together with their relative quantification (*i.e.*, degree of change/fold change respect to controls). In order to perform such statistical evaluation, a function named *plotprofilestable* was developed (Supplementary function provided in Scientific article II, Section 2.2.1), which allowed the performance of two types of t-test (*paired* and *two sample t-tests*) and a non-parametric Wilcoxon rank-sum test. On the other hand, the information provided in the spectra profiles allowed further identification of the metabolites/lipids.

Compressed MS data at low resolution can be transformed into high-resolution MS data to allow proper identification of metabolites (and lipids)

In the developed ROIMCR strategy, enhanced identification of metabolites (and lipids) is possible since the compressed data maintain the original spectral information at high resolution. However, when performing data compression with other approaches, such as binning, the loss of spectral information involved in the compression process difficult the final identification of the metabolites. For those cases, an alternative procedure to recover information of exact mass was developed in this Thesis, taking profit of the advantages of MCR-ALS analysis. In this sense, it was found that the estimation of exact mass could go through the determination of MCR-ALS resolved spectra ($\mathbf{S}^T$) at high resolution, in a two-step process. To demonstrate that, a first MCR-ALS analysis of low resolution compressed data ($\mathbf{D}_{LR}$) was performed, generating MCR-ALS concentration profiles ($\mathbf{C}$) and spectra profiles at low resolution ($\mathbf{S}^T_{LR}$, FIGURE 2.6). Then, it was assumed that concentration profiles ($\mathbf{C}^*$) resulting from the MCR-ALS analysis of non-compressed high resolution data ($\mathbf{D}_{HR}$) (FIGURE 2.6) would be practically the same than the obtained in the previous case ($\boldsymbol{C}=\boldsymbol{C}^*$). Under that assumption, the unknown spectra profiles at high resolution ($\mathbf{S}^T_{HR}$) could be extracted by a (least-squares) pseudo-inversion of the estimated concentration profiles matrix, C, ($\mathbf{S}^T_{HR}=\mathbf{C}^*\backslash\mathbf{D}_{HR}$), as again shown in FIGURE 2.6.

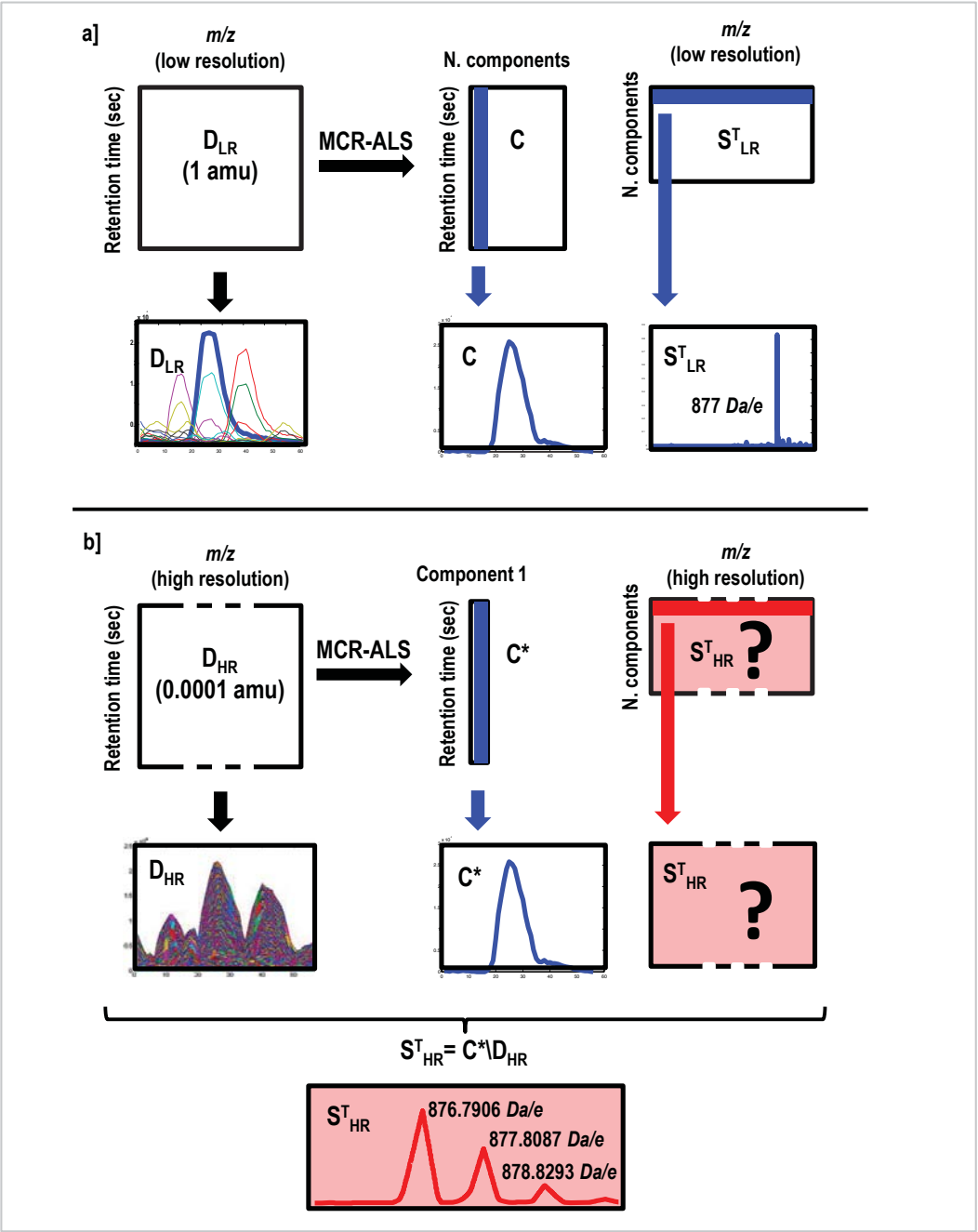


FIGURE 2.6

Representation of the matrices involved in MCR-ALS when performed with a) low resolution D matrix, obtaining C and S^T profiles at low resolution and b) high resolution D matrix and previously generated C profile, to obtain S^T at high resolution by an inversion of matrices. LR: low resolution MS data. HR: high resolution MS data. In this figure, the superindex (*) is used to distinguish whether **C** comes from MCR-ALS analysis of compressed or non-compressed data.

It is really important to stress out that the estimation of the new $\mathbf{S}^T$ at high resolution through the inversion of matrices is possible under the assumption that ***C*=*C**** is true Thus, for the correct fulfillment of the assumption, matrices $\mathbf{D}_{LR}$ and $\mathbf{D}_{HR}$ must contain data of the same LC-MS chromatogram, or the same region of the chromatogram (in the case of time windowing). In the example shown in FIGURE 2.6, the MCR-ALS analysis performed with low resolution data (generated after a binning compression) lead to obtaining a MCR-ALS spectra profile at a resolution of 1 amu, which corresponded to a *m/z* value of 877 *Da/e* for the first component. On the other hand, resulting from the MCR-ALS analysis performed with $\mathbf{D}_{HL}$ and **C***, the generated $\mathbf{S}_{HR}$ provided the exact mass of the same component, which was 876.7906 *Da/e*. Moreover, as seen in FIGURE 2.6b, not only information of the monoisotopic peak was provided but also information of the $[M+H]^+$ (*i.e.*, 877.8087 *Da/e*) and $[M+2H]^+$ (*i.e.*, 878.8293 *Da/e*) species at high resolution.

Data acquisition mode (*e.g.*, profile *vs.* centroid) and information of isotopic distribution are key features in final identification of metabolites

Another alternative to recover high resolution MS information, after a compression that has caused the loss of spectral accuracy (generally after binning compression) was tested in this Thesis. In this case, the information of MCR-ALS spectra profiles obtained at low resolution (1 amu) was used to find out the accurate mass (0.0001 amu resolution) when searching in the raw chromatogram. To do that, retention time was used to confirm the correspondence between the isolated chromatographic peak and the MCR-ALS resolved component. Then, a list of formula candidates was generated using formula determination tools. In this study, since the data were generated with an UHPLC system coupled to a Waters/LCT Premier XE TOF analyzer, controlled with Waters/Micromass MassLynx 4.1 software, the formula determination tool of this software was used. In addition, in order to evaluate the advantages and disadvantages of acquiring data in profile or centroid mode, two differential searches were performed in raw LC-MS chromatograms acquired in these two acquisition modes. Apart from this search using MassLynx software, another software for identification of MS data acquired under profile mode was used, named MassWorks (Cerno Bioscience). When using the latter software, a function named sCLIPS™ (self Calibrating Line-shape Isotope Profile Search)[424] was employed. That function enables users of accurate

mass instruments including TOF, high resolution quadrupoles, Orbitrap, magnetic sector, and FT-ICR MS to dramatically enhance formula identification through spectral accuracy without the requirement to run calibration standards. In FIGURE 2.7, a representation of the steps required to extract exact mass when searching in the raw chromatogram is shown. As it can be seen in this figure, the same information of the MCR-ALS spectra profile at low resolution ($\mathbf{S}^{T}_{LR}$) used in the previous strategy of inversion of matrices was used (*i.e.*, *m/z*= 877 *Da/e*). The search of mass 877 *Da/e* in the raw chromatogram (with an admissible mass error of 0.5 Da) resulted in an isolated chromatographic peak eluting at 16.27 minutes, a retention time that coincided with that obtained in the MCR-ALS concentration profile (**C**). Moreover, the exact mass associated to that isolated peak was searched separately for centroid and profile MS data. The obtained exact masses for $[M]^+$, $[M+1H]^+$ and $[M+2H]^+$ ions were 876.8015, 877.8041 and 878.8099 *Da/e*, respectively, for centroid data and 876.8022, 877.8207 and 878.8083 *Da/e*, respectively, for profile data.

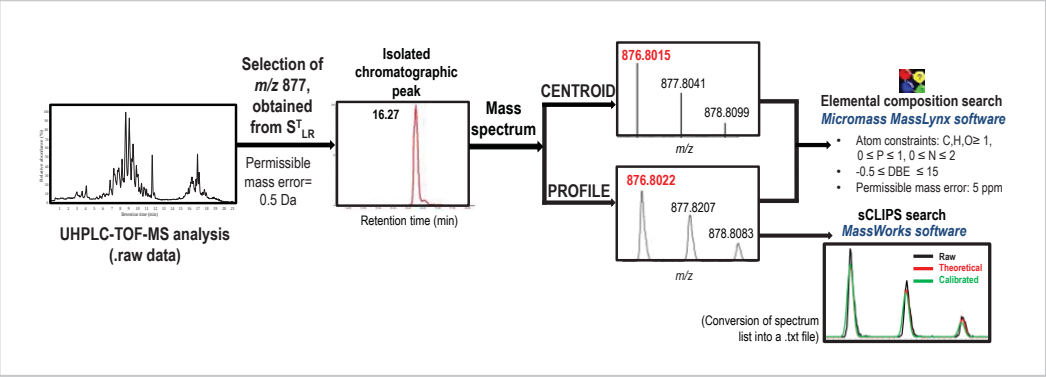


FIGURE 2.7

Representation of the steps involved in the recovery of the exact mass, when searching in the raw UHPLC-TOF-MS chromatograms, acquired in centroid and profile modes. *Not-published results*.

Moreover, in TABLE 2.3, the results of the formula identification when using MassLynx software (both for centroid and profile data) and MassWorks software (only for profile data) are shown. As it can be seen, when performed with centroid data, the formula identification tool of MassLynx was based on isotope information (i-FIT scores). In the case of MassWorks software, information of isotope distribution was also used and the percentage of coincidence between the measured mass and the theoretical mass of the formula candidate was expressed as a percentage of fitting of spectral accuracy.

TABLE 2.3. Formula identification parameters when using MassLynx and MassWorks software under two distinct strategies. *Not-published results.*

Strategy	Software	Nature of the data	Measured mass	Calculated mass	Elemental composition	Mass error (ppm)	i-FIT	Spectral accuracy (%)
Search in raw LC-MS chromatogram	MassLynx formula ID	Centroid	876.8015	876. 8015	$C_{55}H_{106}NO_6$	0.0	3.3	-
		Profile	876.8022	876.8015	$C_{55}H_{106}NO_6$	0.8	-	-
	MassWorks (Sclips)	Profile	876.8022	876.8015	$C_{55}H_{106}NO_6$	0.8	-	92.2
Inversion of MCR-ALS matrices	MassWorks (Sclips)	Profile	876.7906	876.7915	$C_{55}H_{106}NO_6$	-12.4	-	76.2

Moreover, sCLIPS[424] function of MassWorks software was also used to perform formula identification of the exact mass generated in the strategy of the inversion of matrices ($\mathbf{S}^T_{HR}$). The results evidenced that in all cases, the first candidate was a lipid with an elemental composition of $C_{55}H_{106}NO_6$. However, the strategy that lead to the lowest error was the search in the raw chromatogram containing centroid data using MassLynx software (0.0 ppm of error), followed by the search in the raw chromatogram containing profile data using MassLynx software and the search in the raw chromatogram containing profile data using MassWorks software (0.8 ppm of error both), and finally the search performed after the inversion of matrices (15 ppm). In fact, the error associated to the latter search was by far the largest error obtained among the four searches. Thus, it was evidenced that the two strategies tested in this Thesis for the recovery of the exact mass lead to the same results, although a higher error was associated when exact mass was obtained after an inversion of matrices and not after searching in the raw chromatogram. Concerning the capacity of MassLynx versus MassWorks software to perform elemental composition calculation, no significant differences were found, since both formula identification tools allowed proper identification of the formula candidate (with errors ≤ 0.8 ppm). Moreover, it was found that both software performed elemental composition determination in a similar way (*i.e.*, by searching the adequacy between the isotopic distribution of the measured ion and the one of a theoretical candidate). In the case of the Waters elemental composition calculator, the isotopic fitting is expressed with i-FIT parameter (the lower i-FIT, the higher isotopic adequacy). In fact, the use of i-FIT

parameter (*i.e.*, fit of the experimental data to the theoretical isotope distribution) was a solution to perform elemental composition analysis with a wide range and number of elements, because in these cases, the search generally leads to a list of hundreds or even thousands of proposed combinations within the exact mass tolerance of the instrument[425]. A search in the literature shows that MassLynx i-FIT algorithm is used to allow enhanced identification of metabolites in omic studies[426,427]. For instance, Zhao *et al.* used this algorithm to identify seven endogeneous metabolites in a metabonomic study of adenine-induced changes in metabolic profiles of rat faeces[428]. Differently to MassLynx software, MassWorks software expresses the isotopic fitting as a percentage of the spectral accuracy (using 100% to indicate equal isotopic distribution between the theoretical and the measured ion). Some omic studies have benefited from MassWorks software to perform metabolite identification in omic studies. One example of them is an study of Ho H.P. *et al.*, in which new minor metabolites of penicillin G in human serum were identified by multiple-stage tandem MS and the use of MassWorks calibration software[429].

Overall, in this Thesis it was evidenced that the best approach to perform LC-MS metabolomic (and lipidomic) data analysis consists on data compression based on the search of ROIs, since there is no loss of spectral accuracy, and data resolution through MCR-ALS analysis (ROIMCR procedure). Moreover, two alternative strategies to still dispose of MS information at high resolution despite previous loss of it in the early step of compression (when using binning strategy for instance) have been proved to lead to successful results also. These two strategies are based on: i) the recovery of MCR-ALS spectra profiles at high resolution ($\mathbf{S}^{T}_{HR}$) following a pseudo-inversion of matrices, and ii) the search in the raw LC-MS chromatogram through the isolation of the ion of interest and further elemental composition calculation.

Finally, an additional advantage of using the ROIMCR procedure is related to the fact that it allows for a throughtout simultaneous resolution of most of the constituents of a set of multiple correlated samples (control and stressed samples) in a single data analysis step. This provides a very efficient and reliable tool for the investigation of the changes produced at a molecular level of metabolite and lipid concentrations of the investigated biological samples in exposure experiments, as the ones performed in this Thesis, for the purpose of environmental risk assessment studies.

2.4. CONCLUSIONS

From the scientific research included in this chapter, the following specific conclusions can be extracted:

Concerning the adequacy of *Chemometrics* to analyse metabolomic (and lipidomic) data generated by different analytical techniques,

- The use of multivariate data analysis methods has proved to enable the comprehensive analysis of the large complex megavariate data sets (often with incomplete, noisy and non-linear and collinear data structures) generated in metabolomic (and lipidomic) studies.
- Chemometric tools have shown to facilitate the shift from the concept of studying one chemical compound or process at a time to the more comprehensive concept of characterizing the whole biological systems in a single experiment.
- Chemometric tools have proved to be adequate to cover the distinct steps in data analysis, mainly consisting on data pre-processing and pre-treatment, exploratory data analysis by projection methods (*e.g.*, PCA), feature/biomarker detection (*e.g.* PLS-DA), data profiling and resolution by MCR-ALS methods and variance source exploration through the combination of classical ANOVA with multivariate methodologies (*i.e.*, ASCA).

Concerning targeted *vs* untargeted metabolomic (and lipidomic) LC-MS analytical approaches,

- Targeted metabolomic (and lipidomic) approaches only allow the study of a predefined group of metabolites (and lipids) contained in a referential database, which needs to be previously elaborated, whereas untargeted approaches enable extensive analysis of entire metabolomic (and lipidomic) profiles.
- Untargeted approaches involve complex data analysis (including data compression and resolution) since entire data sets (containing massive amount of MS-rich information) need to be processed. However, such analyses can be facilitated with the use of

multivariate chemometric tools. Targeted approaches are generally performed using classical statistical tools and they are usually more tedious and time-consuming.

Concerning the untargeted chemometric methodology for the analysis of LC-MS data sets developed in this Thesis,

- Data conversion of LC-MS files from vendor formats to open data formats (*e.g.*, text or mzXML formats) is required to analyse the data outside the software of the vendor. Among the different ways that exist for data conversion, the external software *ProteoWizard* has showed to enable the conversion of any type of vendor format.
- Compression of LC-MS data sets is necessary to reduce the vast amount of information into more computationally manageable formats and avoid problems related to the limited memory capacity of the computers. Among the distinct data compression strategies (*e.g.*, binning, windowing and ROI), the search of ROIs is proved to be very suitable for the compression (together with data matrix construction) of LC-MS data sets, due to the fact that no loss of spectral accuracy is derived from a ROI compression.
- Three parameters are required to perform a ROI search: signal-to-noise ratio threshold, mass admissible error and minimum number of occurrences. It is proved that the optimum threshold value should be adjusted between 0.1- 1% maximum MS measured intensity. The optimum mass deviation value should be selected halfway between an excessive and an insufficient mass accuracy and should be specifically adjusted for the type of mass spectrometer used. The optimum minimum number of occurrences should be adjusted according to the type of chromatography (*i.e.*, HPLC *vs.* UHPLC) and the corresponding width of the chromatographic peak.
- The developed ROI compression strategy for more than one sample is adapted to augmented data matrices containing relevant information of compressed data of more than one sample (*e.g.*, control and treated samples). In this way the comparison of the peak areas of the resolved elution profiles of the same metabolite in different samples can be performed. Moreover, the pure spectra resolved for the components are more reliable.

- MCR-ALS has shown to be a powerful chemometric method to perform LC-MS data resolution, mainly providing four advantages: i) the possibility of immediate chemical identification of the metabolites thanks to the MS information provided in the analysis; ii) the high degree of direct interpretability of the results; iii) the flexibility in the structure and nature of the data sets that can be potentially analyzed and iv) the added value of not requiring peak shaping nor chromatographic alignment for the simultaneous analysis of multiple samples.
- Among the different feature detection tools, in this Thesis it is demonstrated the adequacy of one-way ANOVA followed by a multiple comparisons test and of PLS-DA analysis through the determination of variables importance for projection (VIPs) for the detection of potential biomarkers for metabolite (and lipid) disruption.

Concerning the identification of potential biomarkers,

- In this Thesis it is demonstrated that proper identification of potential biomarkers requires having precise information about *m/z* values at high resolution. Moreover, it is evidenced that such information is achieved when performing data compression based on the search of ROIs, but it is lost when performing other types of compression, such as binning.
- Recovery of mass at high resolution from low-resolution compressed data can follow two strategies. First strategy consists on extracting MCR-ALS spectra profiles at high resolution (*i.e.*, $\mathbf{S}^T_{HR}$) by a (least-squares) pseudo-inversion of estimated concentration profiles at low resolution ($\mathbf{C}^*$) and LC-MS data at high resolution ($\mathbf{S}^T_{HR}=\mathbf{C}^*\backslash\mathbf{D}_{HR}$). Second strategy consists on using information of MCR-ALS spectra profiles at low resolution to find out the accurate mass when searching in the raw experimental chromatogram (measured at the highest resolution of the instrument), obtaining the isolated ion of interest and further calculating the elemental composition.
- The mode of acquisition of MS data (*i.e.*, profile *versus* centroid) has been proved to have an influence on final identification of metabolites, providing the latter continuous and extensive information of MS spectra profiles. More work is needed to take advantage of this mode.

- Distinct software exist to perform elemental composition determination. Among them, the elemental composition tool provided in the Masslynx software of WATERS corporation, and the software specifically developed for the identification of profile MS data, named MassWorks, have proved to be adequate to perform metabolite (and lipid) identification. Both software (*i.e.*, Masslynx and MassWorks) tools perform the elemental composition search based on the comparative isotopic distribution of the measured ion and that of a theoretical candidate, using the parameters i-FIT and spectral accuracy, respectively. This has been proven in this Thesis in the identification of some lipid species of human placental choriocarcinoma JEG-3 cells, analysed with a Waters/LCT Premier XE TOF analyser.