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 metabolomic studies: Overview and workflow

LC-MS

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

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, Chemometric tools

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-offlight 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 [1 1]), 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

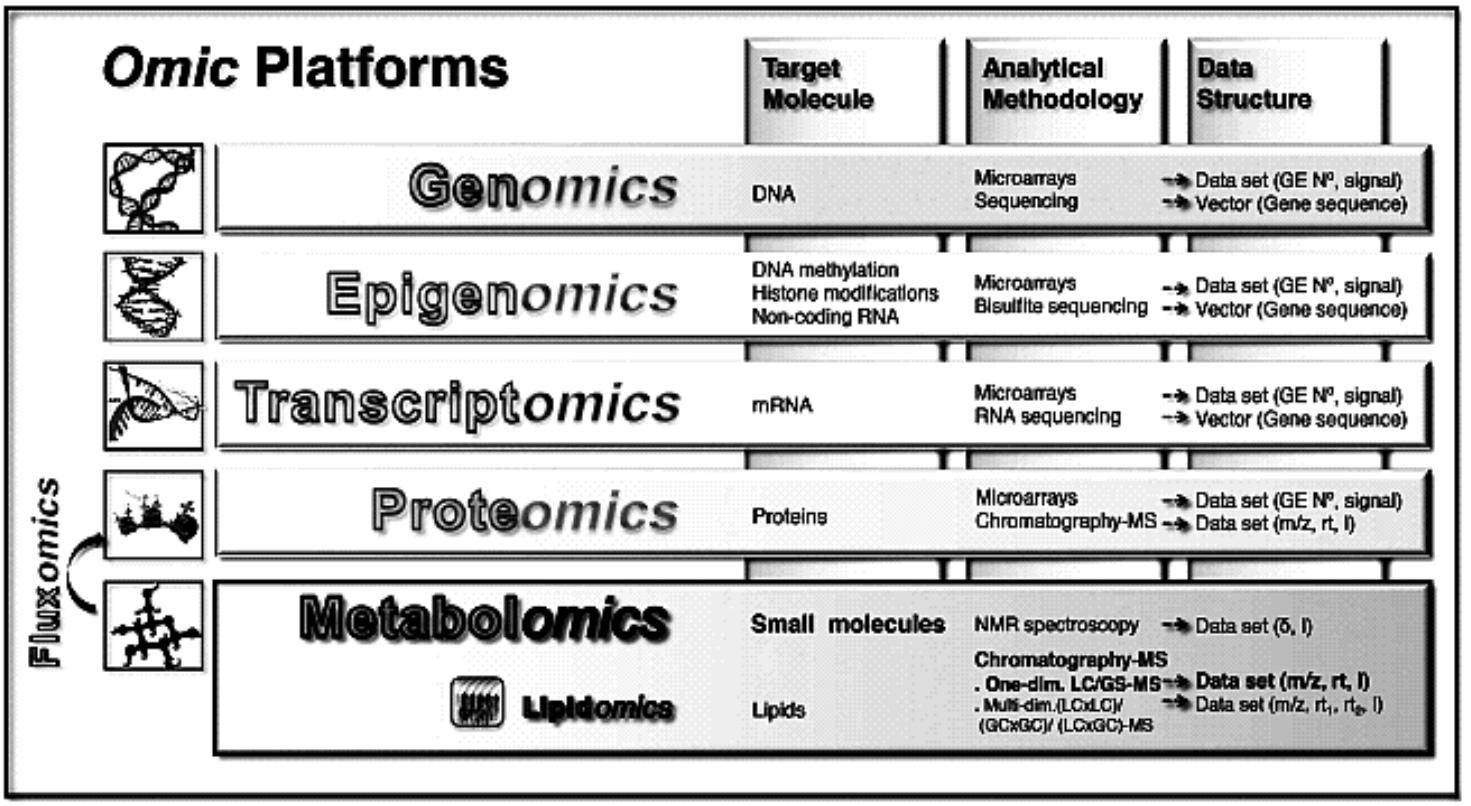


Fig. 1. Overview of OMIC platforms: target molecules, analytical methodologies used and structure of the generated mass-to-charge ratio, rt: retention time, I: intensity). **Data structure shown when considering only one sample.*

data (GE N°: number of genes, δ: chemical shift, m/z:

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 MSbased 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 (LCMS) 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 informationrich 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 highlyextensive 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 stateof-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.

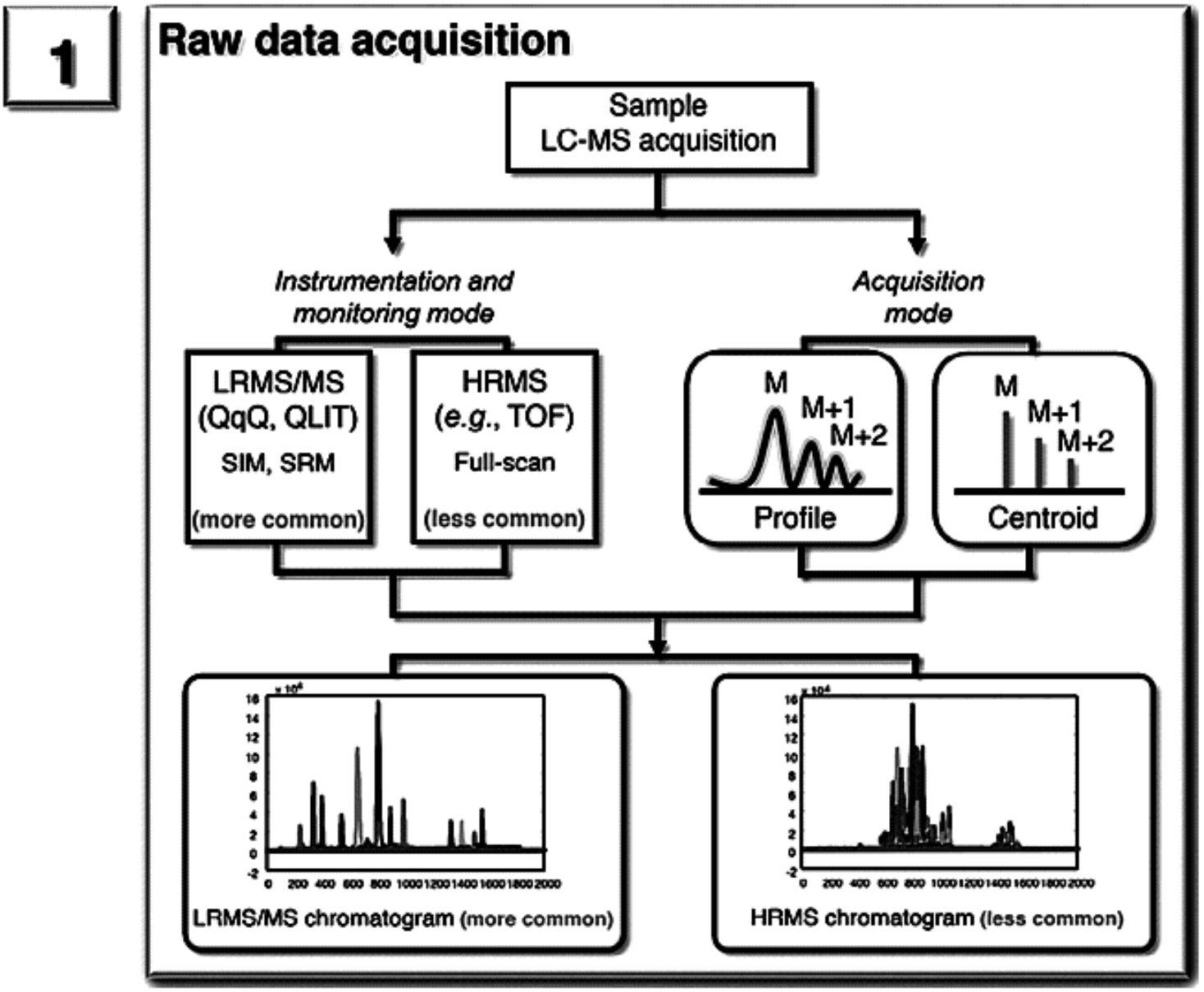
2. General overview of the data analysis approaches

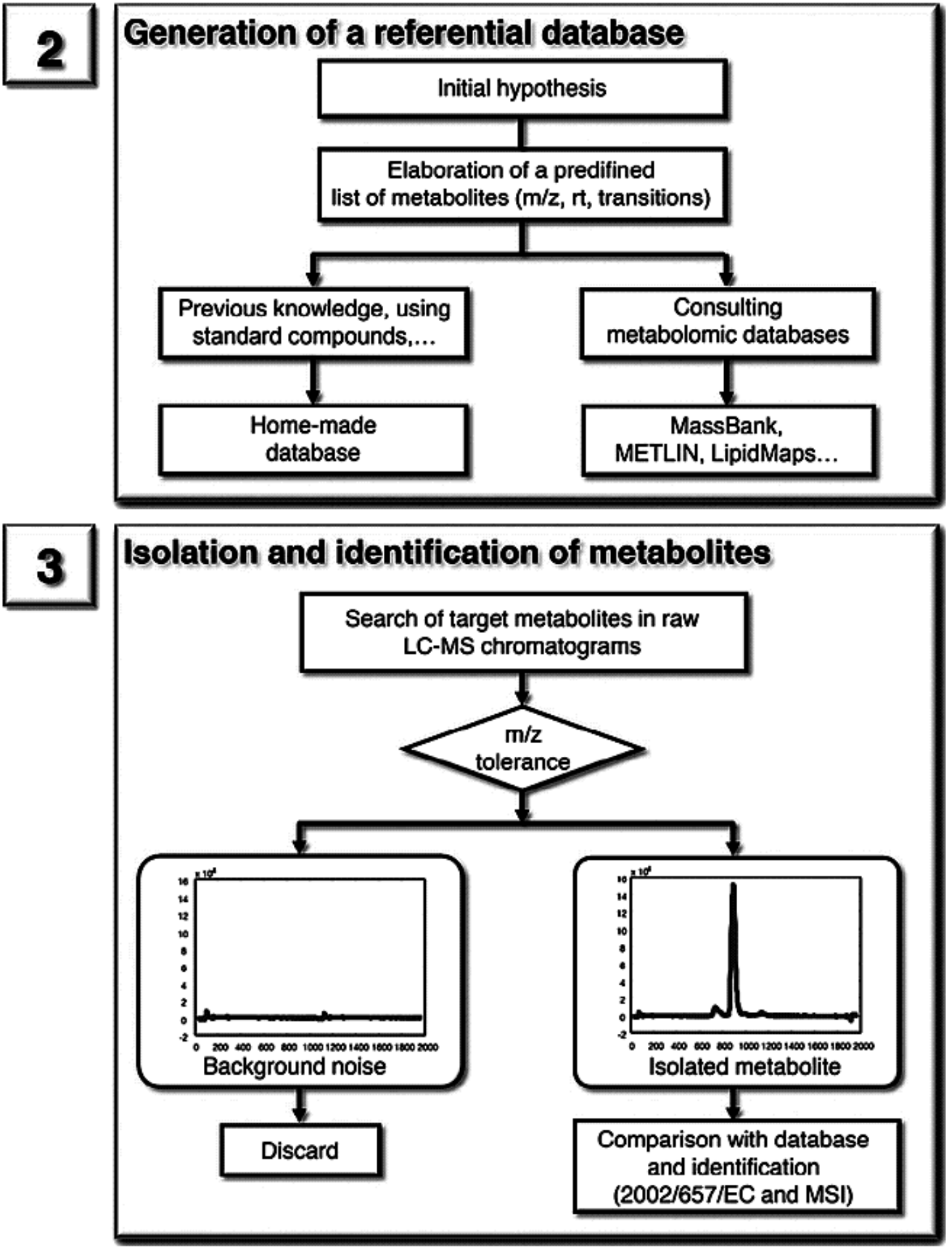
LC-MS metabolomic data analysis strategies are primarily designed for targeted and untargeted studies. However, future advances in LC-MS metabolomics may lead to a merging of targeted and untargeted analyses; with the targeted approach providing more sensitive and accurate detection of predetermined 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 timeconsuming 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


1
Raw data acquisition
Sample
LC-MS acquisition
Instrumentation and monitoring mode
Acquisition mode
LRMS/MS
(QqQ, QLIT)
SIM, SRM
(more common)
HRMS
(e.g., TOF)
Full-scan
(less common)
M
M+1
M+2
Profile
M
M+1
M+2
Centroid
LRMS/MS chromatogram (more common)
HRMS chromatogram (less common)


2
Generation of a referential database
Initial hypothesis
Elaboration of a predifined list of metabolites (m/z, rt, transitions)
Previous knowledge, using standard compounds,…
Consulting metabolomic databases
Home-made database
MassBank, METLIN, LipidMaps…
3
Isolation and identification of metabolites
Search of target metabolites in raw LC-MS chromatograms
m/z tolerance
Background noise
Isolated metabolite
Discard
Comparison with database and identification (2002/657/EC and MSI)

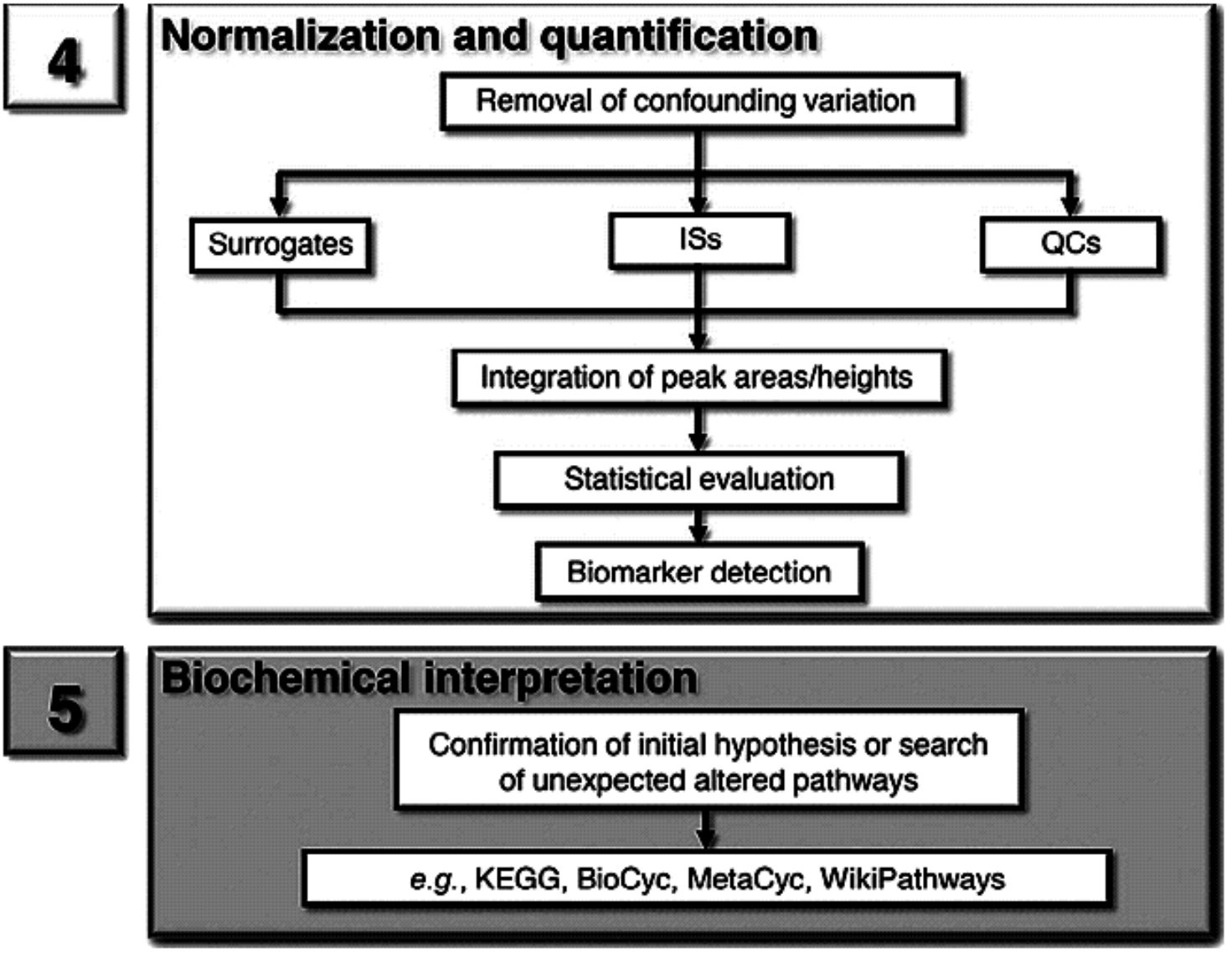


Fig. 2. Overview flowchart listing the five steps (grey shaded areas) involved in the data analysis approach for targeted studies: raw data acquisition, generation of a referential database, isolation and identification of metabolites, normalization and quantification, and biochemical interpretation. These steps are grouped in three major areas: data aquisition (light-grey), data processing and feature detection (mediumgrey) 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 (mediumgrey 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., lowresolution 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 1 890 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.

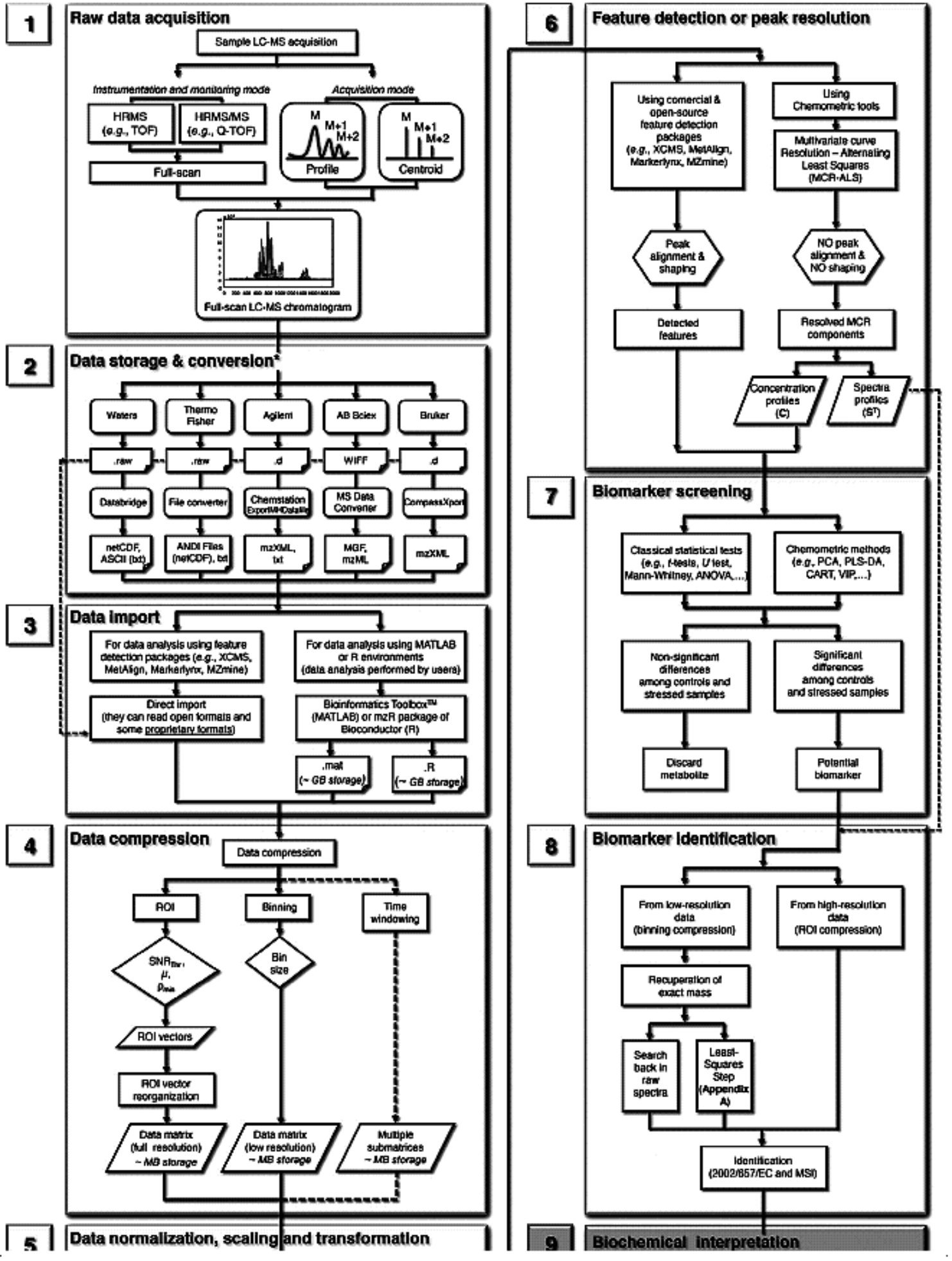
Regardless the instrumentation used for targeted metabolomics (i.e., LRMS/MS or HRMS), identification of metabolites can be enhanced when acquiring data in profile mode, as explained in Section 3.1.1. For instance, Erve et al. [63] and Amorisco et al. [64] used the advantages of acquiring in profile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

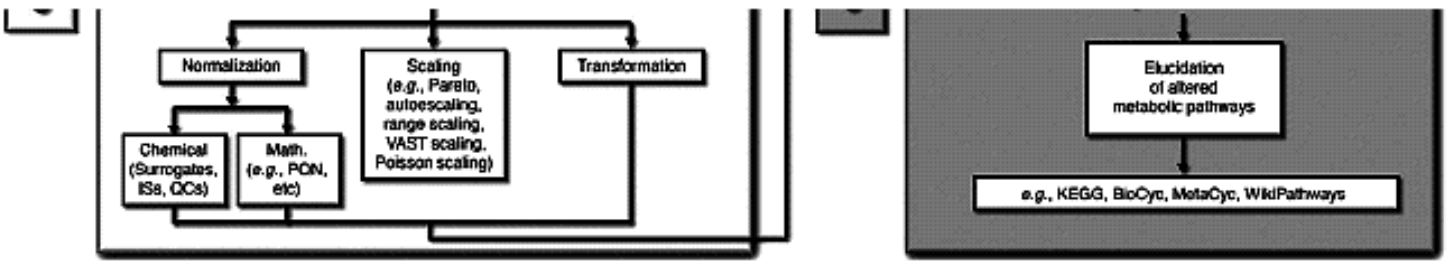

Normalization
Scaling
(e.g., Pareto,
autoscaling,
range scaling,
VAST scaling,
Poisson scaling)
Transformation
Chemical
(Surrogates,
ISs, QCs)
Math.
(e.g., PQN,
etc)
Elucidation
of altered
metabolic pathways
e.g., KEGG, BioCyc, MetaCyc, WikiPathways

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 KruskalWallis 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 highdimensional 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 highresolution 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 GCMS, 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 XMLbased 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

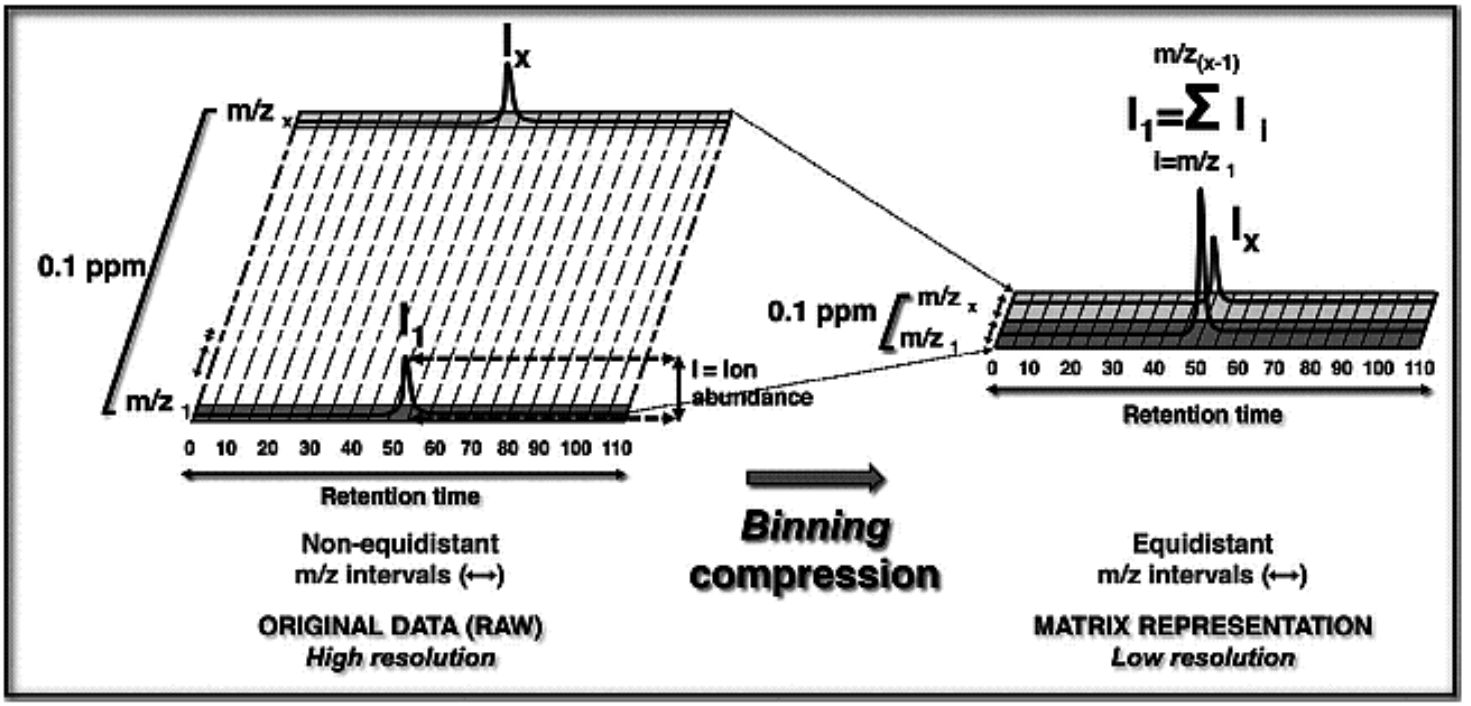


Fig. 4. Scheme of the steps involved in the compression of data when using binning. Example shown for a particular region of an LC-HRMS chromatogram, using a bin size of 0.1 ppm.

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.

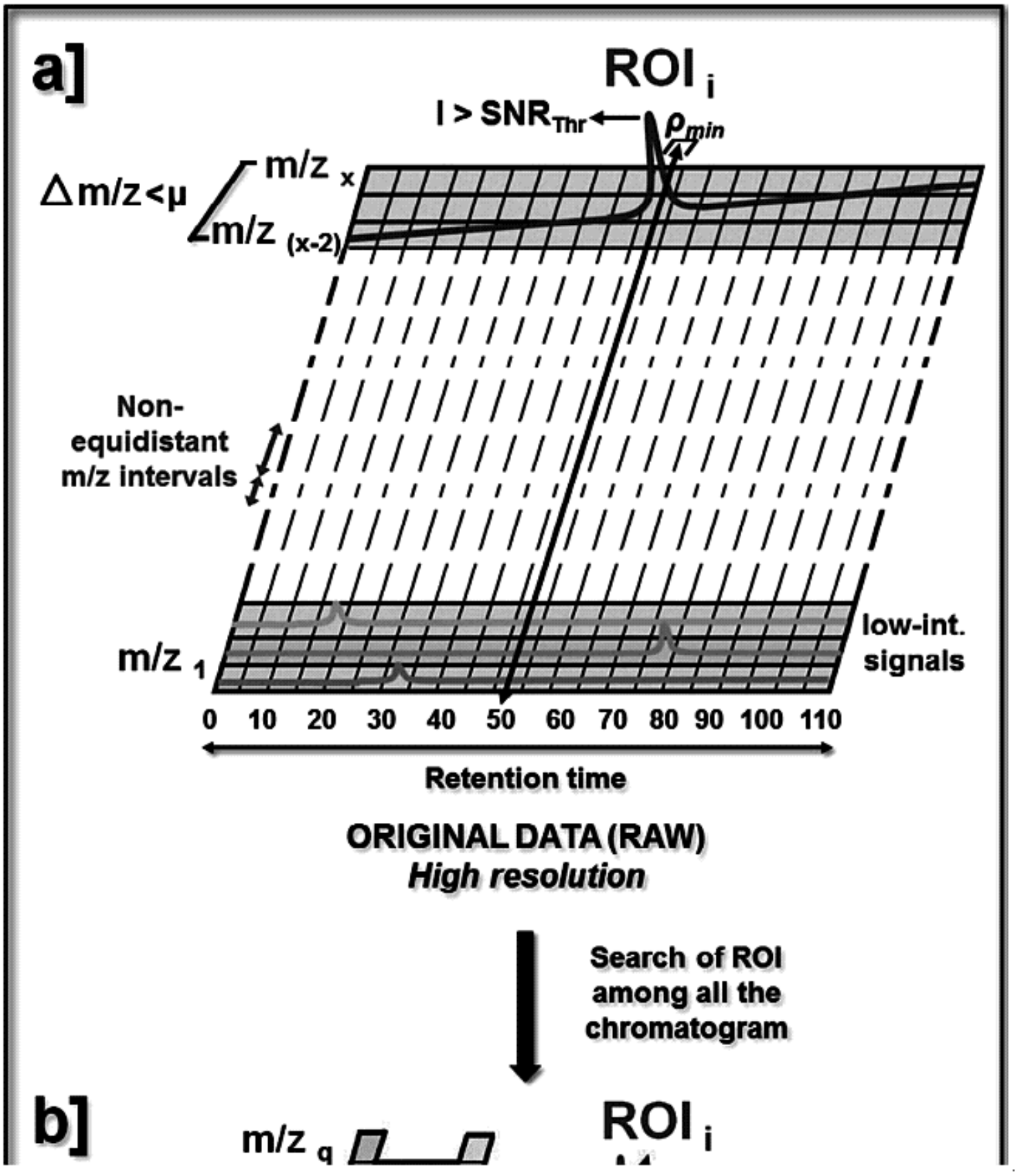
3.2.5. Data intensity normalization, scaling and transformation

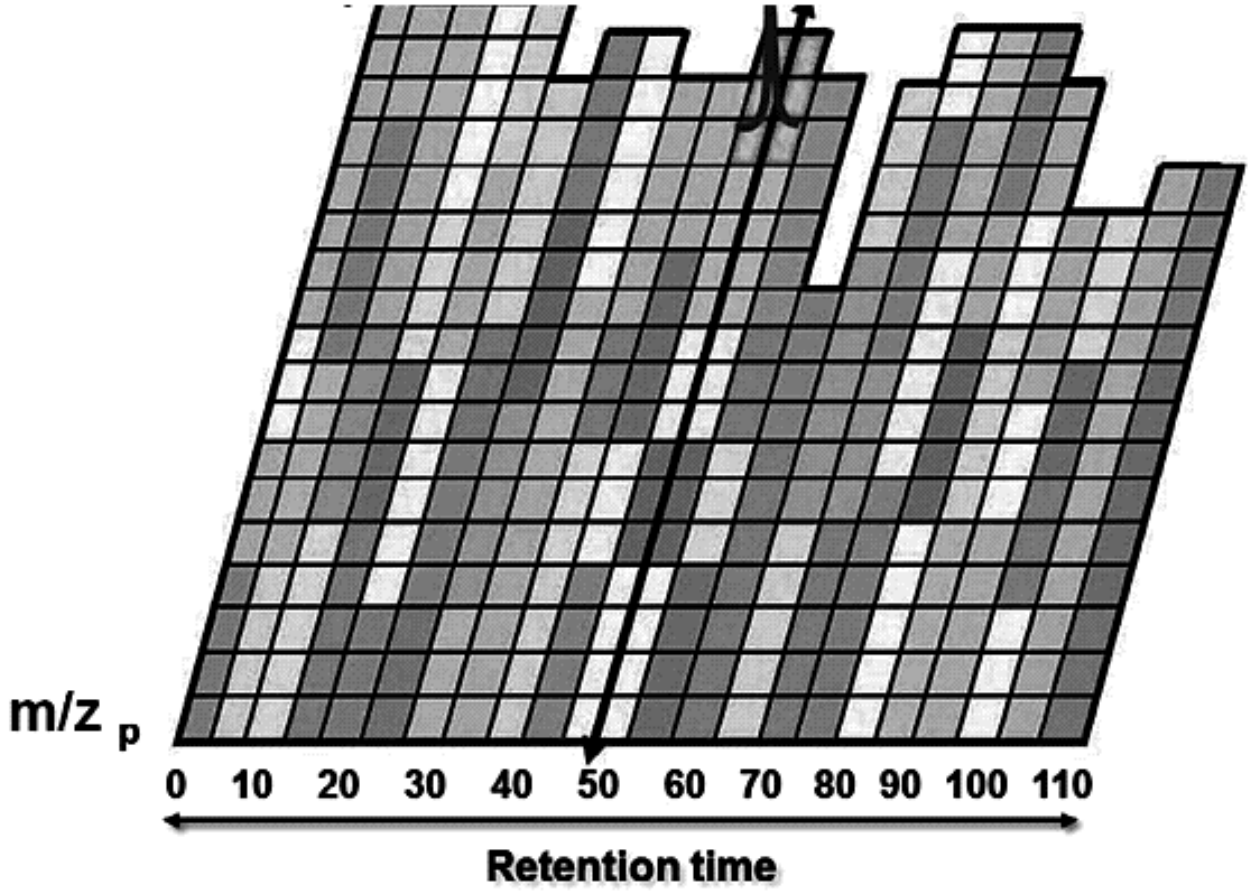
In untargeted approaches, three strategies can be used for removing the unwanted systematic bias in the measurements: sample normalization, data scaling and data transformation. Sample normalization is necessary to adjust the differences among samples whereas data scaling and transformation allow the 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) [1 15]. This method scales all the intensities in a spectrum using the most probable multiplicative factor calculated as the median of the quotients of the


a]
ROI i
I > SNRThr
ρmin
Δm/z<μ
m/z x
m/z (x-2)
Non-equidistant m/z intervals
m/z 1
low-int. signals
0 10 20 30 40 50 60 70 80 90 100 110
Retention time
ORIGINAL DATA (RAW)
High resolution
Search of ROI among all the chromatogram
b]
m/z q
ROI i

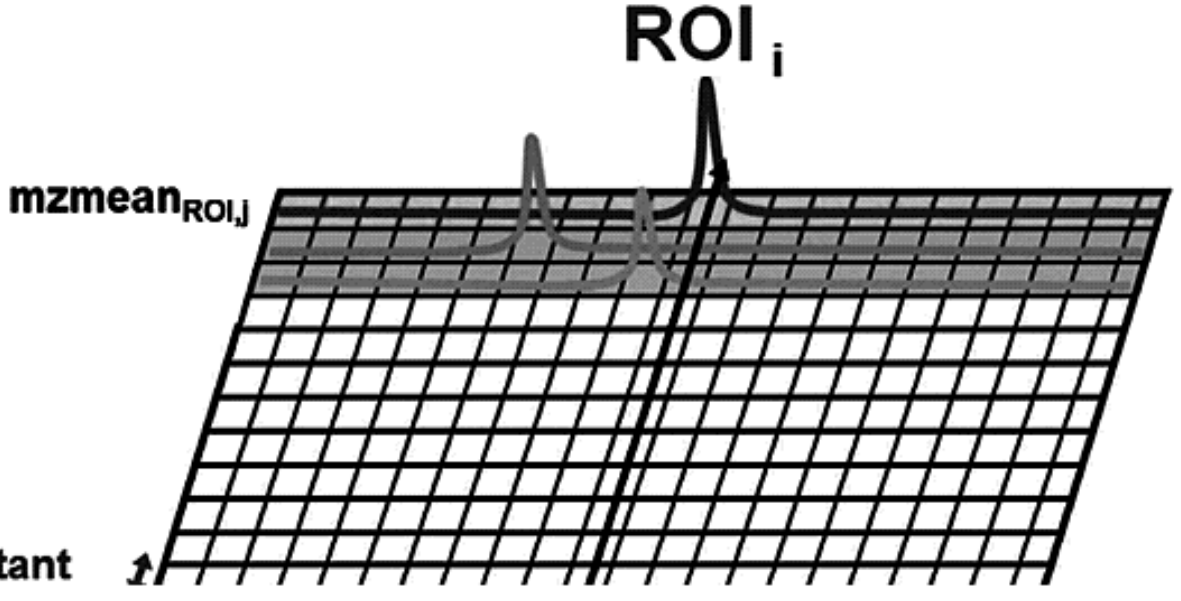


VECTORS containing ROI
High resolution

- **Agrupation of ROI**
- **Total number of ROI: j**
- **Calculation of $mzmean_{ROI}$**

Reorganization into a matrix

c]



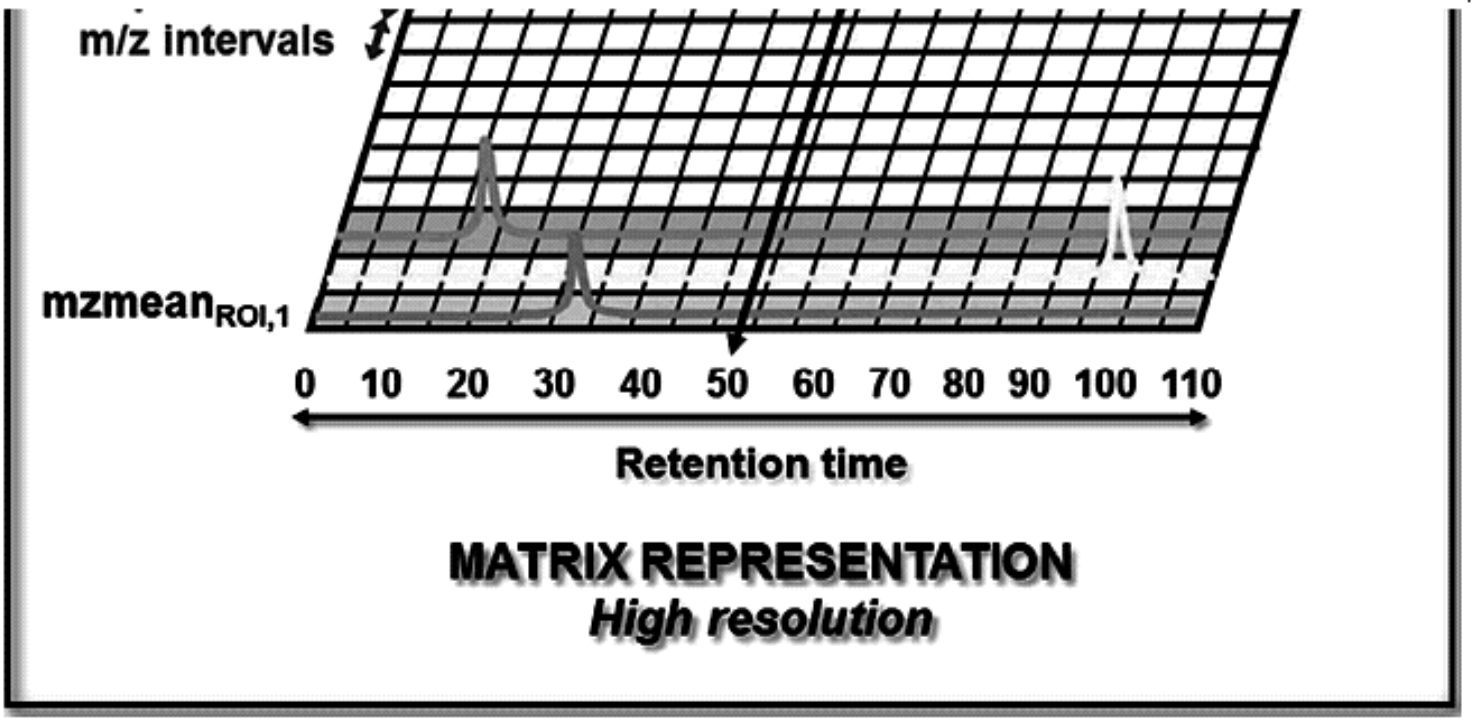


Fig. 5. Scheme of the steps involved in the compression of data by the search of ROI: a] original data with non-equidistant m/z intervals where a significant mass trace is represented as ROI_i (green) and distinguished from low-intensity signals (orange, pink and violet), b] vectors containing the distinct 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 [1 16]. 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 [1 19], 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 [1 18] 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 (MCRALS) [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

---

[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)].

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 $\mathbf{S}^{\mathbf{T}}$. 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 $\mathbf{S}^{\mathbf{T}}$ ($N \times J$) 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 $\mathbf{S}^{\mathbf{T}}$ 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/.

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 secondorder 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 MCRALS. 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 MCRALS 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 MSbased 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 crossvalidation). 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 crossvalidated. 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 jackknife 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 MCRALS) 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 longterm 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 leastsquares 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. [1 3 1] 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 simultaneouscomponent 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:10.111 1/1462-2920.12064.
[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:10.1 194/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) 21 1–221,
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:10.1366/000370288442873 4.
[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 1 1 (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 (201 1) 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 1 1 (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/29861 14?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 2 1 (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 (201 5) 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 101 1 (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) 1 187–1 197, 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) D51 1–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.