Nutrimetabolomics:
An Integrative Action for Metabolomic Analyses in Human Nutritional Studies

Appendix 4

Chapter 7 - Data Processing

Marynka M. Ulaszewska[1,†], Christoph H. Weinert[2,†], Alessia Trimigno[3,†], Reto Portmann[4,†], Cristina Andres Lacueva[5], René Badertscher[4], Lorraine Brennan[6], Carl Brunius[7], Achim Bub[8], Francesco Capozzi[3], Marta Cialiè Rosso[9], Chiara E. Cordero[9], Hannelore Daniel[10], Stéphanie Durand[11], Bjoern Egert[2], Paola G. Ferrario[8], Edith J.M. Feskens[12], Pietro Franceschi[13], Mar Garcia-Aloy[5], Franck Giacomoni[11], Pieter Giesbertz[14], Raúl González-Domínguez[5], Kati Hanhineva[15], Lieselot Y. Hemeryck[16], Joachim Kopka[17], Sabine Kulling[2], Rafael Llorach[5], Claudine Manach[18], Fulvio Mattivi[1,19] Carole Migné[11], Linda H. Münger[20], Beate Ott[21,22], Gianfranco Picone[3], Grégory Pimentel[20], Estelle Pujos-Guillot[11], Samantha Riccadonna[13], Manuela J. Rist[8], Caroline Rombouts[16], Josep Rubert[1], Thomas Skurk[21,22], Pedapati S. C. Sri Harsha[6], Lieven Van Meulebroek[16], Lynn Vanhaecke[16], Rosa Vázquez-Fresno[23], David Wishart[23], and Guy Vergères[20]

[1]Department of Food Quality and Nutrition, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy
[2]Department of Safety and Quality of Fruit and Vegetables, Max Rubner-Institut, Karlsruhe, Germany
[3]Department of Agricultural and Food Science, University of Bologna, Italy
[4]Method Development and Analytics Research Division, Agroscope, Federal Office for Agriculture, Berne, Switzerland
[5]Biomarkers & Nutrimetabolomics Laboratory, Department of Nutrition, Food Sciences and Gastronomy, XaRTA, INSA, Faculty of Pharmacy and Food Sciences, Campus Torribera, University of Barcelona, Barcelona, Spain. CIBER de Fragilidad y Envejecimiento Saludable (CIBERFES), Instituto de Salud Carlos III, Barcelona, Spain
[6]School of Agriculture and Food Science, Institute of Food and Health, University College Dublin, Dublin, Ireland
[7]Department of Biology and Biological Engineering, Food and Nutrition Science, Chalmers University of Technology, Gothenburg, Sweden
[8]Department of Physiology and Biochemistry of Nutrition, Max Rubner-Institut, Karlsruhe, Germany
[9]Dipartimento di Scienza e Tecnologia del Farmaco Università degli Studi di Torino, Turin, Italy
[10]Nutritional Physiology, Technische Universität München, Freising, Germany
[11]Plateforme d'Exploration du Métabolisme, MetaboHUB-Clermont, INRA, UNH, Université Clermont Auvergne, Clermont-Ferrand, France
[12]Division of Human Nutrition, Wageningen University, Wageningen, The Netherlands
[13]Computational Biology Unit, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy
[14]Molecular Nutrition Unit, Technische Universität München, Freising, Germany
[15]Institute of Public Health and Clinical Nutrition, Department of Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
[16]Laboratory of Chemical Analysis, Department of Veterinary Public Health and Food Safety, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium
[17]Department of Molecular Physiology, Applied Metabolome Analysis, Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany
[18]INRA, UMR 1019, Human Nutrition Unit, Université Clermont Auvergne, Clermont-Ferrand, France
[19]Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy
[20]Food Microbial Systems Research Division, Agroscope, Federal Office for Agriculture, Berne, Switzerland
[21]Else Kröner Fresenius Center for Nutritional Medicine, Technisal University of Munich, Munich, Germany
[22]ZIEL Institute for Food and Health, Core Facility Human Studies, Technical University of Munich, Freising, Germany
[23]Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, Canada
†First authors

Mol Nutr Food Res DOI: 10.1002/mnfr.201800384

7. Data Processing

From a data processing point of view, the main challenges to be faced in large scale untargeted metabolomic experiments are: i) the large amount of data to process; ii) the alignment of analytical profiles in large sample sets both in MS and NMR based technologies; and iii) the unambiguous annotation of the features obtained with the corresponding metabolites.

Feature extraction is the first step of any analysis workflow starting from the raw data. Usually it is a complex multi-step procedure that is technology-specific. Moreover, in untargeted experiments, the identified features usually are not the metabolites, but only analytical entities typical of each assay (ions, '$m/z/t_R$' couples, components, NMR bins...). Regardless of the nature of features, their extraction will result in the generation of a data matrix, containing the intensity of all experimental variables in all samples.

At the end of the feature extraction, the data analyst is presented with "raw" data matrices that will undergo the subsequent statistical analysis after an appropriate pre-processing. It is important to stress that the impact of this step on the outcome of the data analysis can be dramatic, in particular for untargeted assays: if something is lost in this extraction phase, it will never come back regardless the sophistication of the data analysis approach. The general pre-processing workflow for MS- and NMR-based technologies will be discussed in the first part of this section. The second part focuses on available software tools and specific approaches for GCxGC alignment.

7.1 Feature Extraction in MS-based Metabolomics

Automated data extraction is a complex multi-step process that can include feature alignment, peak detection and integration, background and blank subtraction, de-isotoping, deconvolution, and normalization. In all cases, this step can be performed either with commercial softwares, often provided by the instrument vendors, or by more flexible open-source solutions. In this second case, the raw data have to be converted into "open" formats, so that a "data conversion" step can be optionally required. Typical software packages include MarkerLynxTM, Progenesis QITM (Waters Corporation),[1] MarkerViewTM (AB Sciex),[2, 3] SieveTM (Thermo Fisher Scientific Inc.),[4] Mass ProfilerTM (Agilent Technologies[5-7] and some open-source tools such as XCMS,[7] MetAlign,[7] MZmine,[8, 9] and OpenMS (http://open-ms.sourceforge.net/).

In all cases, the objective of preprocessing is to take into account unavoidable instrumental drifts ensuring that the intensity of the features can be reliably compared across the samples. This step is a prerequisite to avoid biases during statistical analysis. Thus, in many cases, the alignment of samples is conducted, even if this step can be skipped with some deconvolution procedures.

7.1.1. Data Conversion

All mass spectrometers save raw data in proprietary formats. This lack of standardization is a strong limitation for further data processing steps. For the storage of raw data in open formats, several open-source standards are available. Among them, the common data format (CDF) is quite popular. Unfortunately, CDF files are not suitable to store multi-event MS experiments in a single file and are not designed to store spectral metadata (e.g. collision energy, precursor, etc.). More recently, XML-based solutions have been implemented, such as mzXML,[10] mzML[11] in which multi-event experiments are supported. Of note, the ProteoWizard software[12, 13] can also be used for conversion in batch mode.[14, 15]

7.1.2. Peak Picking

As discussed in the introduction, in many cases, the feature extraction requires to identify the ionic species showing a chromatographic peak in the t_R domain distinguishing them from noise. This process is defined as "peak picking" because each detected metabolite will produce at least a peak in the '$m/z/t_R$' plane.

Different tools, either commercial or open-source, are actually available to perform this task, among them we can mention MZmine,[16] MetAlign,[17] and XCMS.[18] It is important to point out that all these approaches do not take into consideration compound spectra but work on a "feature based" approach and, therefore,generate a lot of redundancy in the resulting datasets. Several studies focused on the comparison of automated pre-processing tools for LC- and GC-MS data. Coble and Fraga[19] evaluated four freeware tools, MetAlign, MZmine, XCMS, and SpectConnect on a signature discovery task. The comparative study was performed on accurate mass from LC-MS and nominal mass from GC-MS data. MetAlign was the best performer to detect components in low and high-resolution data (more than 80%). However, the authors recommend combining pre-processing tools for untargeted detection of compounds to improve the performances. **Table S7.1** summarizes the main peak picking parameters to look at.

7.1.3. Deconvolution

Objective and Principle

Deconvolution can be considered as an alternative to peak picking because, with this approach, the features are identified on the basis of a multivariate deconvolution process that extracts and constructs pure compound spectra from raw data. Analytes that coelute in the chromatographic domain, i.e. two (or more) peak apices that cannot be unambiguously resolved, are candidates for so called 'peak deconvolution'.

Deconvolution - although this term as such is inadequately used for the context of GC/LC-MS - aims to separate the coeluting compounds mathematically, based on their chromatographic peak properties and their mass spectral channels. When background subtraction as a means to purify the mass spectrum of an analyte is not a relevant option, spectral deconvolution is one way to detect and extract analytes properly. Spectral skewing (spectral tilting) is related to the change amongst the ratios of the analyte mass fragment intensities during the analyte's chromatographic elution, i.e. intensity changes in the ion source during the consecutive scans. Due to the lower acquisition rates, spectral skewing is inherently common to quadrupole instruments (scanning mass acquisition, e.g. from low to high m/z) as opposed to ToF-MS (simultaneous mass acquisition) instruments.

Table S7.1. Parameters for peak picking in mass spectrometry.

Parameter	Mzmine	MetAlign	XCMS	metaMS
Type of data	Low and high resolution	Low and high resolution	Low and high resolution	Low and high resolution
Input	mzML, mzXML, mzdata, NetCDF, Thermo RAW, Waters RAW	mzXML, mzdata, NetCDF, Thermo RAW, Waters RAW, Agilent RAW	mzXML, mzdata, NetCDF, Agilent RAW	mzML, mzXML, mzdata, NetCDF
Peak width	No option		FWHM	FWHM
Minimum signal-to-noise			snthresh	No option
m/z resolution			mzdiff	No option
RI alignment	No	No	No	Yes

FWHM: full width at half maximum of chromatographic peaks; *m/z*diff: minimum difference in *m/z* for peaks with overlapping t_R; RI: retention index; snthresh: signal to noise threshold.

Accordingly, spectral skewing can produce so called interfering ions and will affect the (pure) analyte's apex position and, as a consequence, has to be taken into account by the available software module, being either proprietary, stand-alone or an open source solution.

The approaches addressing the mathematical purification of coeluting compounds can be subdivided in two divergent categories. The first category encompasses algorithms that investigate the chromatographic peak shapes of the available mass channels individually, while the second category includes a suite of multivariate approaches considering all mass channels at once such as orthogonal signal correction (OSD),[20] non-negative matrix factorization (NNMF)[21] or multivariate curve resolution (MCR) for GC×GC-MS[22] and liquid chromatography/high resolution mass spectrometry (LC-HRMS).[23] In the second approach, the features are not isolated ions anymore, but component spectra that could be related to the ionization patterns of unique metabolites. It is noteworthy that the deconvolution algorithms belonging to the first category are heavily interweaving with the peak detection step. Indeed, peaks not being captured by a peak detection algorithm cannot be subjected to the deconvolution process.

The success of the spectral deconvolution process depends strongly on the quality of the raw data. It is commonly assumed that at least a partial chromatographic separation, a Gaussian peak shape, a sufficient data acquisition rate (or, more precisely, an adequate number of data points per peak) as well as consistent mass spectra (no skewing) are crucial.[24] The problem is that, concerning the aforementioned factors, the literature provides no generally valid thresholds above or below which deconvolution can be safely performed. As a rare example in the literature, Zushi et al.[25] optimized deconvolution parameters in more detail using GC×GC-HR-TOF-MS data. They confirmed that a higher data acquisition rate (i.e., 50 Hz instead of 25 Hz) helps deconvoluting especially closely eluting substances. At the same time, they pointed out that a very high data acquisition rate may cause higher detection limits and may inflate unnecessarily the data volume as well as the computational burden. Surprisingly, the use of high mass resolution did apparently not improve the deconvolution result.[25] Beyond that, other practically relevant questions remained unanswered so far, for example concerning the minimum number of data points per peak, the minimum distance between neighbouring peaks (in terms of data points between the peak apices) or the tolerability of a limited spectral skewing. As the result of spectral deconvolution depends not only on raw data quality but also on other factors like the algorithmic approach used,[26, 27] it seems to be difficult to define generally valid requirements for a successful spectral deconvolution.

Added Value

After deconvolution, the quantification of the purified compound is done by selection of a respective quantification ion (QI) for relative quantification, where care has to be taken to ensure that this QI is coherently determined for the deconvoluted compound spectra across all samples. Here, different quantification ions may be selected according to the software package under use. An evaluative juxtaposition of the currently available software solutions can be found in the literature.[26]

Computing deconvoluted spectra also improves the matching when trying to identify a compound against a provided reference spectrum library, such as NIST (in GC-MS) or an in-house developed library of pure compounds. Here, the purified (deconvoluted) spectra are likely to increase the similarity index (SI) value for the library spectrum match. Due to their inherent complexity, different software solutions provide different results in terms of false positive and false negative peak detection rates. Moreover, the careful adjustment of the respective software input parameters is of critical importance and requires an in depth understanding of both the instrumental apparatus as well of the underlying mathematical deconvolution procedure in order to exploit its full potential.

Available Tools

Deconvolution approaches are more common in GC-MS than in LC-MS experiments, even if their popularity is increasing also for LC-based assays. Alternative tools based on different deconvolution algorithms have been developed by either instrument manufacturers or academic teams. Mastrangelo et al. summarizes the main characteristics of these tools[28] and their field of application depending on the data type (high or low resolution). AMDIS (Automated Mass Spectral Deconvolution and Identification) is probably one of the earliest and widest known application for spectral deconvolution of GC-MS data. AMDIS has been developed at NIST: it analyzes mass channels independently without any spectral alignment and determines so called ‘model peaks’. It can be used through a Graphical User Interface (GUI) and is provided as a 32 bit application, which unfortunately is not able to adequately treat GC×GC-MS raw data files. BinBase aligns compounds across samples and provides compound quantification and identification based on self-constructed libraries.[29] ADAP (Automated Data Analysis Pipeline for untargeted metabolomics) is a freeware tool for the analysis of GC/LC-MS data[30, 31] that has been recently integrated into MZmine (since v. 2.24.). Similarly to AMDIS, ADAP investigates chromatographic mass channels separately and, based on a hierarchical clustering of fragment ions, it calculates individual chromatographic peak features (CPF). This tool is popular and often used in the version specific for GC-based data, namely ADAP-GC.[30] Another freely available R based deconvolution tool worth to be mentioned is eRah.[20] MetaboliteDetector can also be considered as a prominent tool for the analysis of GC-MS data and it is provided with a GUI.[32] MetaboliteDetector algorithm is inspired by the algorithms implemented in AMDIS[33] with further improvements, including, for instance, the analysis of highly resolved profile mass data.

For GC-MS instruments reporting nominal masses, we recommend excluding mass fragments with *m/z*-values of (73, 74, 75, 147, 148, and 149) as they are not specific for the differentiation of compounds. These mass fragments predominantly result from the derivatization reagent (TMS) and their intensities should be set to zero across all analytes for all sample runs, so as to increase the selectivity for spectral library matching. It is also beneficial not to include the above fragments into any deconvolution algorithms.

Although the deconvolution process allows good separation of coeluting compounds, it may not be sufficient for a correct assignment if it only relies on spectral similarity. For this reason, the retention data are further employed. We suggest using both retention index (RI) and t_R values to increase the accuracy and reliability of the identification. For GC×GC-ToF MS data, two softwares are commonly used: the commercial ChromaTOF software (LECO Corporation, St Joseph, MI, USA) and parallel factor analysis.[34] However, due to company policies, the underlying algorithmic details can often not be revealed, and these commercial software solutions cannot be obtained without purchasing the corresponding instruments.

7.1.4. *Integrated Solutions*

More recently, alternative solutions were developed to combine both approaches (automatic extraction and deconvolution), especially to process GC-MS data, where analytical correlations in datasets are much more important. The Metab R package[35] was designed to automate the analysis of GC-MS data processed by AMDIS. GAVIN[36] and SpectConnect[37] also employ AMDIS result files as the input for the extraction program. Different open source solutions were also developed with the objective to propose full workflows that integrate all the necessary steps for data processing in GC-MS-based untargeted metabolomics. In particular, eRah,[20] an R package, incorporates a novel spectral deconvolution method using multivariate techniques based on orthogonal signal deconvolution (OSD), alignment of spectra across samples, quantification, and automated identification of metabolites by spectral library matching. As a 64 bit application it is capable of deconvoluting also larger GC×GC-MS raw data runs. More recently, the metaMS package was implemented under a web-based platform in a fully dedicated tool named metaMS.runGC within the Workflow4Metabolomics 3.0 Galaxyonline infrastructure, which allows building workflows including pre-processing, statistical analysis, and annotation steps, and also offers the first reference repository for metabolomic workflows.[38]

At this point, it is important to note the specificity of GC×GC-MS data, as this technology is a valuable instrumental setup also for nutritional metabolomics experiments. GC×GC-MS has indubitably sovereign chromatographic separation power when compared to traditional GC-MS systems, its separation being reported as 3 times higher.[39] On the other hand, the size of the experimental raw data (due to the required higher scan rates) and the resulting tables, due to the repetitive peak modulations after the raw data processing step (signal detection and spectra extraction), are drastically larger. One should keep in mind that decreasing the modulation time period PM from e.g. 9 seconds to 3 seconds will increase the size of the resulting peak and mass spectral tables by a factor of approximately 3. Of note, despite the increased separation power, coelution is still apparent, thus the deconvolution of chromatographic peaks is still necessary, although becoming less frequent.

7.1.5. *Alignment*

Despite the advances in chromatographic techniques used for metabolomics, there is always some variation in t_R of a metabolite across different sample runs. Alignment is needed for correcting t_R differences between runs and combining data from different samples. Most alignment methods work in pairwise fashion by aligning either only pairs of samples or multiple samples against a selected reference sample or a template. In general, the choice of the reference sample has effect on the alignment results.

Alignment methods can be roughly divided into two categories: i) methods using raw data as input and generating a set of mappings that transform t_R axis of each run to a common t_R axis; ii) Methods that start by first detecting the features in each sample and then looking for a consensus list of “common” features to construct the final data matrix. Some alignment methods combine both approaches, e.g. by first performing a t_R mapping between runs and then clustering the detected features using corrected rts. Niu et al.[40] compared eight programs for the alignment of metabolomic GC-MS data. They collected low-resolution data and tested the performances of SpecConnect, MetaboliteDetector, MetAlign, MZMine, TagFinder, XCMS, MeltDB, and GAVIN. Some of these tools detected more than 80% of the compounds of interest. Even if the comparison cannot be generalized since limited to two datasets, the authors highlighted the difficulties in optimizing the different programs and identified a subset of tools showing good performances on both matrices. Reviews with full description of different alignment methods applied in mass spectrometry based metabolomics can be found in the literature.[41, 42]

7.1.6. Data Processing Approaches for GC×GC-MS Metabolomics Datasets

Compared to one-dimensional GC-MS or LC-MS data, GC×GC-MS raw data like other multi-dimensional chromatographic data exhibit a structural peculiarity: the existence of multiple chromatographic peaks per analyte due to the fractionation of the first-dimensional eluate by the process of modulation (see section 5.4). While the modulation is technically necessary to exploit the separation potential of the second chromatographic dimension, it amplifies the raw data volume drastically and, at the same time, creates the need to recombine the different peaks of the same analytes ("modulations") in a given run at a later stage in order to determine a final quantitative measure (i.e. area or height). As a consequence, specific algorithms and software are required to process GC×GC-MS metabolomics data. To date, there are mainly two different approaches which will be described briefly in the following.

The Peak Feature-based Approach: the Extension of the Traditional One-Dimensional Approach to GC×GC Data

This approach comprises two different stages. In the first stage, the instrument manufacturer's software (usually LECO's ChromaTOF™) is used to extract the GC×GC raw data by performing essential steps like baseline correction, S/N filtering, mass spectral deconvolution, peak detection and integration, library matching and the recombination of the peaks or "modulations" belonging to the same analyte (often referred to as "peak merging" or "demodulation"). By this automatic processing, one textual peak list per raw data file is created, of which size is considerably smaller than the original GC×GC raw data (reduction by a factor of about 5-10). In the second stage, the remaining steps like the correction of t_R shifts as well as the alignment are performed based on the peak lists using academic softwares, e.g. INCA,[43] Guineu,[44] DISCO2,[45] MetPP,[46] BiPACE2D,[47] or SquareDance.[48]

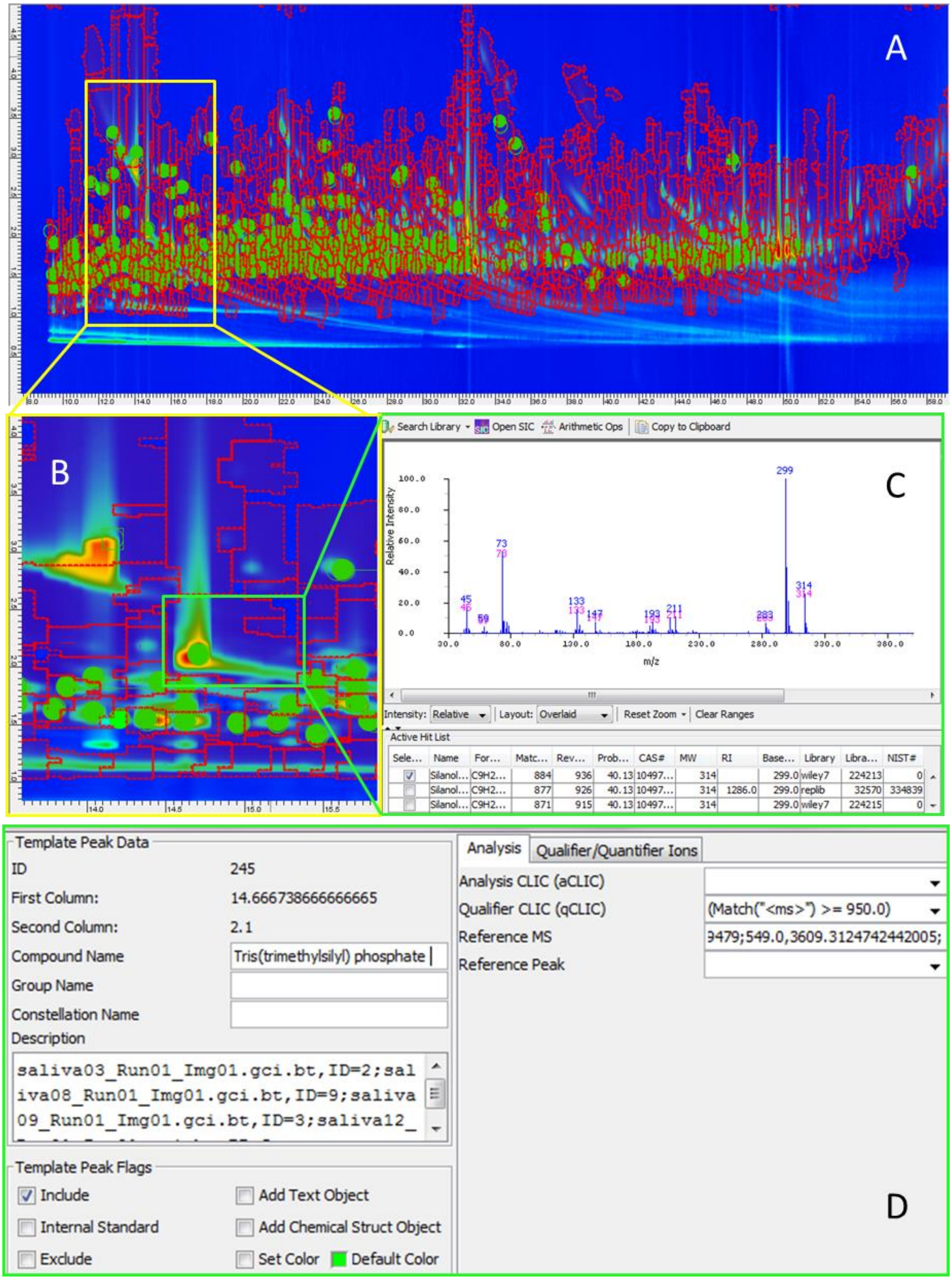
The Peak-region Feature Approach: Analysis of GC×GC Data by Pattern Matching

In simple terms, the peak-regions features based approach makes use of the fact that GC×GC data can be visualized as two-dimensional plots which can be processed with methods developed for pattern recognition. These approaches based on *peak-region* features have become popular because of their comprehensiveness and for the uniform treatment of information from each sample constituent, both knowns and unknowns.[49-52] These approaches are implemented in commercial software packages (GC Image, GC Image LCC, Lincoln, NE, US), are intuitive and have available tutorials for new users(http://www.gcimage.com/gcimage_tutorial/index.html). Of interest for complex samples analysis as in metabolomic profiling and fingerprinting investigations, are those approaches that combine *2D-peak* and *2D-chromatographic region* concepts. *Region* features characterize multiple data-points (e.g. collecting all MS data-point events in each chromatographic *region*) and are delineated by 2D-peaks contour as it is defined by the 2D-data integration algorithms. In ideal conditions, each single chemical entity is chromatographically resolved by its neighbours and can be univocally characterized by its chromatographic and spectrometric parameters (t_R, detector response, and mass spectral information - named peak metadata) and by its absolute and relative position within the *pattern* of all detectable constituents. This pattern can be treated as a template of known or unknown 2D peaks and/or *peak-regions* to be used for pattern matching, i.e. to find reliable correspondences between pattern peaks/regions with the peak pattern of a target chromatogram. Positive correspondences can be constrained by fixing a bi-dimensional retention search space (with a certain tolerance around the expected ^1t_R and ^2t_R) and MS spectral similarity for direct and reverse matching. Additionally, pattern shifts due to chromatographic inconsistencies or extra-chromatographic phenomena can be compensated by choosing suitable transformation algorithms (affine or polynomial functions) that smartly guide the template matching.[53]

Very complex samples, as those from nutrimetabolomics work-flows, pose several additional challenges due to coelution phenomenon and/or the wide dynamic range for metabolites that may be present as trace peaks in some samples and as very abundant peaks in some others. Peak detection inconsistencies can cause matching errors between the reference template and the target chromatogram. In these situations, the *peak-region* features overcome these challenges; the approach attempts to define one region (i.e. a small 2D t_R window) per peak over the chromatographic plane to achieve the one-feature-to-one-analyte selectivity of peak features methods but with the implicit matching of *region* features. *Regions* are graph objects with a defined contour that can be added to a reference template for pattern matching. Briefly the work-flow that creates *peak-regions* from a data set of chromatograms: (a) the peak detection algorithm detects and records the peak patterns in individual chromatograms of the set; (b) a few peaks (named registration peaks) that are reliably matched across all samples-but-one are fixed (matching constraints are fixed a priori by the analyst); (c) the algorithm aligns and combines sample chromatograms to create a composite chromatogram; (d) a pattern of *region* features from the peaks detected in the composite chromatogram is defined. Then, when a target chromatogram is analyzed, (e) the registration peaks are matched to target chromatogram pattern, the feature *regions* are aligned relative to those peaks, and the characteristics of those features are computed to create a feature vector for the target chromatogram, and finally (f) the feature vector is used for cross-sample analysis (e.g. classification, discriminant analysis, clustering, etc.).

Panel A in **Figure S7.1** below shows a 2D contour plot of a saliva sample where all detectable constituents are highlighted by *peak-regions* contours (red graphs) and single 2D-peaks are marked as green circles. In panel B, a zoomed area shows some individual peaks chromatographically resolved by neighbours, Panel C shows the MS fragmentation pattern of one known analyte (i.e. tris(trimethylsilyl) phosphate) while panel D illustrates all peak-region metadata collected (t_R, compound name, the list of chromatograms where the peak was positively matched, reference MS and the qualifier CLIC function that is the matching function recalling the NIST algorithm).

This comprehensive investigation, recently defined as UT fingerprinting (Untargeted and Targeted fingerprinting) has been successful in different application fields when large and complex dataset have to be effectively explored for reliable cross-sample analysis: fructose induced changes in mice metabolome,[54, 55] for breast cancer metabolomics,[49] in extra-virgin olive oil volatiles analysis[52] and in cocoa processing signaturing.[56, 57]



Intensity: Relative Layout: Overlaid

Active Hit List

Sele...	Name	For...	Matc...	Rev...	Prob...	CAS#	MW	RI	Base...	Library	Libra...	NIST#
☑	Silanol...	C9H2...	884	936	40.13	10497...	314		299.0	wiley7	224213	0
□	Silanol...	C9H2...	877	926	40.13	10497...	314	1286.0	299.0	replib	32570	334839
□	Silanol...	C9H2...	871	915	40.13	10497...	314		299.0	wiley7	224215	0

Template Peak Data

ID	245
First Column:	14.666738666666665
Second Column:	2.1
Compound Name	Tris(trimethylsilyl) phosphate
Group Name	
Constellation Name	

Description

saliva03_Run01_Img01.gci.bt,ID=2;saliva08_Run01_Img01.gci.bt,ID=9;saliva09_Run01_Img01.gci.bt,ID=3;saliva12_

Template Peak Flags

☑ Include □ Add Text Object

□ Internal Standard □ Add Chemical Struct Object

□ Exclude □ Set Color 🟩 Default Color

Analysis CLIC (aCLIC)	
Qualifier CLIC (qCLIC)	(Match("<ms>") >= 950.0)
Reference MS	9479;549.0,3609.3124742442005;
Reference Peak	

Figure S7.1. Picture-based processing of GC×GC-MS metabolomics data using the pattern matching approach. Panel A of Figure S7.1 shows a 2D contour plot of a saliva sample where all detectable constituents are highlighted by peak-regions contours (red graphs) and single 2D-peaks are marked as green circles. In panel B, a zoomed area shows some individual peaks chromatographically resolved by neighbours. Panel C shows the MS fragmentation pattern of one known analyte (i.e. tris(trimethylsilyl) phosphate) while panel D illustrates all peak-region metadata collected (tR, compound name, the list of chromatograms where the peak was positively matched, reference MS and the qualifier CLIC function that is the matching function recalling the NIST algorithm)

7.2. Feature Extraction in NMR-based Metabolomics

7.2.1. Fourier Transform (FT).

The first operation to be carried out for the processing of the acquired NMR spectra is to transform the raw data or Free Induction Decay (FID) from time-domain to frequency-domain. In fact, raw data consist of a convolution of many signals, each one with its specific oscillating intensity over the acquisition time, whilst it is more functional to the interpretation and analysis of data to determine the frequency and the area of each signal (originated by each proton). To do so, the mathematical FT operation is performed. This operation looks at the sine wave of a signal and analyzes it in order to determine its frequency to generate a new plot in the frequency domain. In other words, the FT operation transforms the original data recorded as signals intensities, oscillating and decaying over the time (time-domain), into a new spectrum expressed as signals intensities as a function of their frequency (frequency-domain) If a signal decays quickly, the new spectrum will appear as a broad peak, whilst if it has a slower rate of decay, it will appear as a sharp peak. Enhancement of the spectral quality could be carried by the addition of zeroes at the end of the FID data. This operation, called “zero-filling”, will not change any characteristics of the peaks or spectrum, but will help improving the spectral digital resolution. Zero-filling is carried out by defining the size of the dataset before the FT operation.

7.2.2. Baseline Correction

One of the first steps needed in spectral processing after FT is the correction of the baseline. Ideally, the baseline should be a straight horizontal line of zero intensity. However, since it consists of the average noise of the spectrum, the baseline can greatly diverge from the zero intensity. Baseline shifts occur because some incorrect data is usually collected at the beginning of the FID. Modern instruments have built-in automatic baseline correction algorithms, which is currently the best option. This method usually requires the definition of a region in which signals are clearly distinguishable from noise, so that the noise regions can be fitted and projected in the signal region to be subtracted from the spectrum.

7.2.3. Phase Correction

In order to attribute the correct peak shape to spectral peaks after the FT operation, it is necessary to perform a phasing step. This operation corrects for instrumental errors that necessarily occur during the acquisition of the raw spectrum. In particular the absorptive and dispersive modes of the complex spectrum (i.e. real and imaginary parts) are mixed in varying percentages in the acquired spectrum. In order to correct for this effect, a linear combination of the real and imaginary parts of the spectra is calculated and used for phasing. The phasing angle, which needs to be calculated, is a linear function of the chemical shift, this function being defined by an intercept (zero-order phase correction) and a slope (first-order phase correction). The first step of the phasing is to define a so-called “pivot peak” at a chemical shift of 0 and to optimize the value of the intercept on this peak. Then a peak at the opposite side of the spectrum is chosen to adjust the value of the slope. Nowadays, the best option for spectral phasing is to use automatic phase routines built in the instrument software (i.e. apk0.noe for NOESY Bruker spectra, aph for Varian spectra). These routines will generally provide a good phasing and avoid operator-based biases. Thus, the spectra should be phased manually, as described above, only when the quality of the automatic phasing is insufficient.

7.2.4. Frequency Calibration

For serum samples prepared with d6-4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), spectra are referenced to d6-DSS at δ = 0.00 ppm. When trimethylsilylpropanoic acid (TSP) is used spectra are calibrated to TSP at δ = 0.00 ppm.

7.2.5. Binning

Spectral data can be affected by numerous external variations, which could invalidate the results of statistical analysis. One of these variations is the inter-sample peak position, even after frequency calibration. This effect can be caused by many factors. Generally, this happens due to a difference in the pH of samples, since the chemical shift depends on the ionization state of the molecules in the sample Moreover, the concentration of metal ions, the chemical exchange, or the interactions occurring between proteins and metabolites can also cause variations in the chemical shift of molecules such as citrate.[58] These differences can result in the occurrence of specific groupings in multivariate data analysis, especially in urine. Urine generally has a pH value ranging between 5.50 and 6.50 but can also reach values ranging from 4.60 to 8.00. This shift can be caused by drug treatments, particular diets, or particular health conditions. If the pH is maintained between 7.10 and 7.70, shifts are reduced to less than 0.02 ppm (if the salt concentration is lower than 0.15 M) and the samples should consequently be buffered. However, as shifts might still occur in the pH 7 range, the impact of these variations should be further reduced by integrating signals into bins (or buckets): we recommend binning intervals with a width of 0.04 ppm. The binning operation also helps in reducing the size of the data matrix, while still keeping important information on the NMR signals. Importantly, binning is not an alignment operation, as for example the icoshift method[59]; due to the high number of peaks present in the spectra, alignment is not recommended for the NMR analysis of urine and serum samples except for specific regions where signals are clearly resolved.

7.2.6. Data Reduction

During the analysis of NMR spectra in untargeted experiments, data reduction is achieved by removing uninformative regions of the spectra, such as signals from water (4.60 - 5.00 ppm) and from peripheral parts, which include almost exclusively noise signals (form -0.50 to 0.70 and > 9.00 ppm). Further data reduction is often achieved by “binning” or “bucketing” (see section 7.2.5). The bin is a cluster of adjacent data points. Of note, each bin defines a variable obtained by either the average, the maximum, or the integral of the intensity of all the data points contained in that bin. The most common option is to use the integral, i.e. the sum of intensities, of all data points. For example, a typical 64k NMR spectrum over 10 ppm of frequency scale, which is reduced using bin widths of 0.02 ppm, will result in ~500 bin integral values (i.e. 10/0.02).

7.2.7. Integration

The integration of spectral signals aims at quantifying the absolute or relative concentration of metabolites. The integrated intensity of a signal in a 1H-NMR spectrum is, in fact, directly correlated to the concentration and the number of hydrogens, which produce the signal. The NMR integrations are always relative, therefore an internal standard is needed to determine the absolute concentration.

References

[1] E. Pujos-Guillot, J. Hubert, J. F. Martin, B. Lyan, M. Quintana, S. Claude, B. Chabanas, J. A. Rothwell, C. Bennetau-Pelissero, A. Scalbert, B. Comte, S. Hercberg, C. Morand, P. Galan, C. Manach, *J. Proteome Res.* **2013**, 12, 1645.

[2] R. Llorach, S. Medina, C. Garcia-Viguera, P. Zafrilla, J. Abellan, O. Jauregui, F. A. Tomas-Barberan, A. Gil-Izquierdo, C. Andres-Lacueva, *Electrophoresis* **2014**, 35, 1599.

[3] X. Mora-Cubillos, S. Tulipani, M. Garcia-Aloy, M. Bullo, F. J. Tinahones, C. Andres-Lacueva, *Mol. Nutr. Food Res.* **2015**, 59, 2480.

[4] L. Y. Hemeryck, C. Rombouts, T. Van Hecke, L. Van Meulebroek, J. V. Bussche, S. De Smet, L. Vanhaecke, *Toxicol. Res.* **2016**, 5, 1346.

[5] M. J. Cichon, K. M. Riedl, L. Wan, J. M. Thomas-Ahner, D. M. Francis, S. K. Clinton, S. J. Schwartz, *Mol. Nutr. Food Res.* **2017**, 61, 1700241.

[6] K. Hanhineva, C. Brunius, A. Andersson, M. Marklund, R. Juvonen, P. Keski-Rahkonen, S. Auriola, R. Landberg, *Mol. Nutr. Food Res.* **2015**, 59, 2315.

[7] J. A. Rothwell, Y. Fillatre, J. F. Martin, B. Lyan, E. Pujos-Guillot, L. Fezeu, S. Hercberg, B. Comte, P. Galan, M. Touvier, C. Manach, *PloS One* **2014**, 9, e93474.

[8] H. Liu, T. J. Garrett, Z. Su, C. Khoo, L. Gu, *J. Nutr. Biochem.* **2017**, 45, 67.

[9] J. L. Preau, Jr., L. Y. Wong, M. J. Silva, L. L. Needham, A. M. Calafat, *Environ. Health Persp.* **2010**, 118, 1748.

[10] P. G. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. Mccomb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, R. Aebersold, *Nat. Biotechnol.* **2004**, 22, 1459.

[11] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz, E. W. Deutsch, *Mol. Cell. Proteomics* **2011**, 10, R110.000133.

[12] D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, *Bioinformatics* **2008**, 24, 2534.

[13] E. Chambers, D. M. Wagrowski-Diehl, Z. Lu, J. R. Mazzeo, *J. Chromatogr. B* **2007**, 852, 22.

[14] R. Smith, D. Ventura, J. T. Prince, *Brief. Bioinform.* **2015**, 16, 104.

[15] C. Brunius, L. Shi, R. Landberg, *Metabolomics* **2016**, 12, 173.

[16] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, *BMC Bioinformatics* **2010**, 11, 395.

[17] A. Lommen, in: Hardy, N. W., Hall, R. D. (Eds.), *Plant Metabolomics: Methods and Protocols*, Humana Press, Totowa, NJ 2012, pp. 229-253.

[18] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, *Anal. Chem.* **2006**, 78, 779.

[19] J. B. Coble, C. G. Fraga, *J. Chromatogr. A* **2014**, 1358, 155.

[20] X. Domingo-Almenara, J. Brezmes, M. Vinaixa, S. Samino, N. Ramirez, M. Ramon-Krauel, C. Lerin, M. Díaz, L. Ibáñez, X. Correig, A. Perera-Lluna, O. Yanes, *Anal. Chem.* **2016**, 88, 9821.

[21] Y. Zushi, S. Hashimoto, K. Tanabe, *Anal. Chem.* **2015**, 87, 1829.

[22] L. W. Hantao, H. G. Aleme, M. P. Pedroso, G. P. Sabin, R. J. Poppi, F. Augusto, *Anal. Chim. Acta* **2012**, 731, 11.

[23] M. M. Sinanian, D. W. Cook, S. C. Rutan, D. S. Wijesinghe, *Anal. Chem.* **2016**, 88, 11092.

[24] D. Turner, *Chromatography Today* **2016**, February/March, 10.

[25] Y. Zushi, S. Hashimoto, K. Tanabe, *Anal. Chem.* **2015**, 87, 1829.

[26] H. Lu, Y. Liang, W. B. Dunn, H. Shen, D. B. Kell, TrAC *Trends Anal. Chem.* **2008**, 27, 215.

[27] X. Du, S. H. Zeisel, *Comput. Struct. Biotechnol. J.* **2013**, 4, e201301013.

[28] A. Mastrangelo, A. Ferrarini, F. Rey-Stolle, A. Garcia, C. Barbas, *Anal. Chim. Acta* **2015**, 900, 21.

[29] K. Skogerson, G. Wohlgemuth, D. K. Barupal, O. Fiehn, *BMC Bioinformatics* **2011**, 12, 321.

[30] Y. Ni, M. Su, Y. Qiu, W. Jia, X. Du, *Anal. Chem.* **2016**, 88, 8802.

[31] O. D. Myers, S. J. Sumner, S. Li, S. Barnes, X. Du, *Anal. Chem.* **2017**, 89, 8696.

[32] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, *Anal. Chem.* **2009**, 81, 3429.

[33] S. E. Stein, *J. Am. Soc. Mass Spectrom.* **1999**, 10, 770.

[34] R. A. Harshman, *UCLA Working Papers in Phonetics* **1970**, 16, 1.

[35] R. Aggio, S. G. Villas–Bôas, K. Ruggiero, *Bioinformatics* **2011**, 27, 2316.

[36] V. Behrends, G. D. Tredwell, J. G. Bundy, *Anal. Biochem.* **2011**, 415, 206.

[37] M. P. Styczynski, J. F. Moxley, L. V. Tong, J. L. Walther, K. L. Jensen, G. N. Stephanopoulos, *Anal. Chem.* **2007**, 79, 966.

[38] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E. A. Thévenot, *Int. J. Biochem. Cell Biol.* **2017**, 93, 89.

[39] J. H. Winnike, X. Wei, K. J. Knagge, S. D. Colman, S. G. Gregory, X. Zhang, *J. Proteome Res.* **2015**, 14, 1810.

[40] W. Niu, E. Knight, Q. Xia, B. D. McGarvey, *J. Chromatogr. A* **2014**, 1374, 199.

[41] M. Katajamaa, M. Oresic, *J. Chromatogr. A* **2007**, 1158, 318.

[42] A. Nordstrom, G. O'Maille, C. Qin, G. Siuzdak, *Anal. Chem.* **2006**, 78, 3289.

[43] M. F. Almstetter, I. J. Appel, M. A. Gruber, C. Lottaz, B. Timischl, R. Spang, K. Dettmer, P. J. Oefner, *Anal. Chem.* **2009**, 81, 5731.

[44] S. Castillo, I. Mattila, J. Miettinen, M. Orešič, T. Hyötyläinen, *Anal. Chem.* **2011**, 83, 3058.

[45] B. Wang, A. Fang, X. Shi, S. H. Kim, X. Zhang, *Bio-Inspired Computing and Applications*, Springer, Berlin 2012, pp. 486-491.

[46] X. Wei, X. Shi, I. Koo, S. Kim, R. H. Schmidt, G. E. Arteel, W. H. Watson, C. McClain, X. Zhang, *Bioinformatics* **2013**, 29, 1786.

[47] N. Hoffmann, M. Wilhelm, A. Doebbe, K. Niehaus, J. Stoye, *Bioinformatics* **2014**, 30, 988.

[48] B. Egert, C. H. Weinert, S. E. Kulling, *Journal of Chromatography A* **2015**, 1405, 168.

[49] S. E. Reichenbach, X. Tian, Q. Tao, E. B. Ledford, Jr., Z. Wu, O. Fiehn, *Talanta* **2011**, 83, 1279.

[50] S. E. Reichenbach, X. Tian, C. Cordero, Q. Tao, *J. Chromatogr. A* **2012**, 1226, 140.

[51] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, S. E. Reichenbach, X. Tian, Q. Tao, *J. Chromatogr. Sci.* **2010**, 48, 251.

[52] F. Magagna, L. Valverde-Som, C. Ruiz-Samblas, L. Cuadros-Rodriguez, S. E. Reichenbach, C. Bicchi, C. Cordero, *Anal. Chim. Acta* **2016**, 936, 245.

[53] S. E. Reichenbach, D. W. Rempe, Q. Tao, D. Bressanello, E. Liberto, C. Bicchi, S. Balducci, C. Cordero, *Anal. Chem.* **2015**, 87, 10056.

[54] D. Bressanello, E. Liberto, M. Collino, S. E. Reichenbach, E. Benetti, F. Chiazza, C. Bicchi, C. Cordero, *J. Chromatogr. A* **2014**, 1361, 265.

[55] D. Bressanello, E. Liberto, M. Collino, F. Chiazza, R. Mastrocola, S. E. Reichenbach, C. Bicchi, C. Cordero, *Anal. Bioanal. Chem.* **2018**, 410, 2723.

[56] F. Magagna, A. Guglielmetti, E. Liberto, S. E. Reichenbach, E. Allegrucci, G. Gobino, C. Bicchi, C. Cordero, *J. Agric. Food Chem.* **2017**, 65, 6329.

[57] F. Magagna, E. Liberto, S. E. Reichenbach, Q. Tao, A. Carretta, L. Cobelli, M. Giardina, C. Bicchi, C. Cordero, *J. Chromatogr. A* **2018**, 1536, 122.

[58] O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson, E. Holmes, *Anal. Chem.* **2005**, 77, 517.

[59] F. Savorani, G. Tomasi, S. B. Engelsen, *J. Magn. Reson* **2010**, 202, 190.