Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography A, 1158 (2007) 318–328
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Review

Data processing for mass spectrometry-based metabolomics

Mikko Katajamaa [a,*], Matej Orešič [b,*]

[a] *Turku Centre for Biotechnology, Tykistökatu 6, FIN-20521 Turku, Finland*
[b] *VTT Technical Research Centre of Finland, Tietotie 2, P.O. Box 1500, FIN-02044 VTT, Espoo, Finland*

Available online 19 April 2007

Abstract

Modern analytical technologies afford comprehensive and quantitative investigation of a multitude of different metabolites. Typical metabolomic experiments can therefore produce large amounts of data. Handling such complex datasets is an important step that has big impact on extent and quality at which the metabolite identification and quantification can be made, and thus on the ultimate biological interpretation of results. Increasing interest in metabolomics thus led to resurgence of interest in related data processing. A wide variety of methods and software tools have been developed for metabolomics during recent years, and this trend is likely to continue. In this paper we overview the key steps of metabolomic data processing and focus on reviewing recent literature related to this topic, particularly on methods for handling data from liquid chromatography mass spectrometry (LC–MS) experiments.

Keywords: Metabolomics; Lipidomics; Proteomics; Normalization; Alignment; Liquid chromatography; Mass spectrometry; Feature extraction; Peak detection; Deconvolution

Contents

1. Introduction

Metabolomics is a discipline dedicated to the global study of metabolites, their dynamics, composition, interactions, and responses to interventions or to changes in their environment, in cells, tissues, and biofluids. Concentration changes of specific groups of metabolites may be descriptive of systems responses to environmental or genetic interventions, and their study may therefore be a powerful tool for characterization of complex phenotypes [1–3] as well as for development of biomarkers for specific physiological responses [4,5].

Multiple experimental platforms are commonly applied in the studies of metabolites, including NMR, LC–MS, GC–MS, CE–MS and infrared spectroscopy [6–8]. Technologies used in metabolomics produce large amounts of data, and handling such complex metabolomic datasets is an important step that has big impact on extent and quality at which the metabolite identification and quantification can be made. Since in metabolomics we are primarily interested in biological systems

* Corresponding authors. Tel.: +358 20 722 4491; fax +358 20 722 7071.
E-mail addresses: mikko.katajamaa@btk.fi (M. Katajamaa), matej.oresic@vtt.fi (M. Orešič).

0021-9673/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2007.04.021

responses resulting in metabolite level regulation related to genetic variation or multitude of environmental changes, it is important to separate *interesting* biological variation from *obscuring* sources of variability introduced in studies of metabolites, including at various stages of data processing. The quality of data processing is therefore an essential step for our ability to properly analyze and interpret metabolomic data.

Data handling tasks in metabolomics can be roughly divided into two steps: *data processing* and *data analysis*. The data processing step consists of low-level processing of raw data with signal processing methods and combining data between measurements. These tasks transform the raw data into format that is easy to use in the subsequent data analysis steps. The data analysis stage includes tasks for analysis and interpretation of processed data. This typically includes multivariate analyses such as clustering of metabolic profiles or discovering important differences between groups of samples. In proteomics, a similar classification has been proposed, while further dividing the data processing stage into low-level and mid-level [9].

Since metabolomic platforms are being increasingly utilized to characterize biological systems, increasing amounts of metabolomic data are being accumulated. In order to keep the data accessible and comparable between laboratories, there is need for standards for reporting of the experiment details. Minimum information about a metabolomics experiment (MIAMET) [10] defines the minimum required information that should be stored from metabolomic experiment along with the measured data. MIAMET description contains requirements on storing details of metabolomic data processing and analysis as well as other experiment steps. Recently presented formal models, like MeMo [11] and ArMet [12], implement the requirements of MIAMET and will help in designing MIAMET compatible software tools for metabolomics.

A broad picture of metabolomics has already been presented in numerous review articles [3,6,7,13–18]. In this review we will focus on the recent literature describing new methods for metabolomic data processing, especially for LC–MS type of data. From the view point of data processing, metabolomics and unlabeled proteomic profiling using LC–MS require the same processing steps, while the differences are mainly in sample complexity and availability of MS/MS spectra for peptide identification. For this reason, few methods for proteomics data processing of relevance to metabolomics are also covered in this review.

2. Data processing

2.1. Overview

In mass spectrometry-based metabolomics, the starting point for data processing is a set of raw data files, each file corresponding to a single biological sample. A single LC–MS data file is a collection of successively recorded histograms, each representing hits of ionized molecules on the detector during a small time frame [19]. A histogram consists of a number of *m*/*z* and intensity data points.

The basic aim of data processing is to transform raw data files into representation that facilitates easy access to characteristics of each observed ion. These characteristics include *m*/*z* and retention time of the ion and an ion intensity measurement from each raw data file. In addition to these basic features, data processing can extract additional information like isotope distribution of the ion.

Since different instrument vendors utilize different proprietary data formats, a preliminary step for data processing required in software that supports metabolomic data from multiple vendors, is conversion of such raw proprietary data into common raw data format such as netCDF (ASTM E2078-00, Standard Guide for Analytical Data Interchange Protocol for Mass Spectrometric Data) or mzXML [20]. Vendor software packages usually contain scripts that can perform data conversion to netCDF or ASCII formats. Converters to more recent mzXML format have been developed both by research groups and companies.

Typical data processing pipeline usually proceeds through multiple stages, including *filtering*, *feature detection*, *alignment* and *normalization* (Fig. 1). Filtering methods process the raw measurement signal with aim of removing effects like measurement noise or baseline. Feature detection is used to detect representations of measured ions from the raw signal. Alignment methods cluster measurements across different samples and normalization removes unwanted systematic variation between samples.

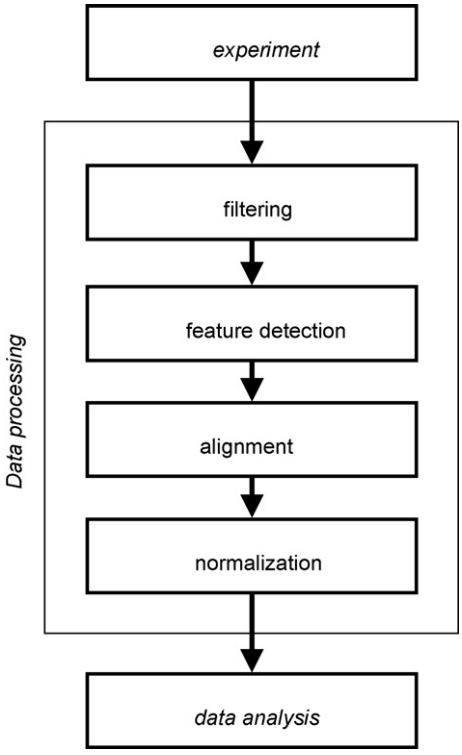

Fig. 1. Summary of metabolomic data processing workflow.

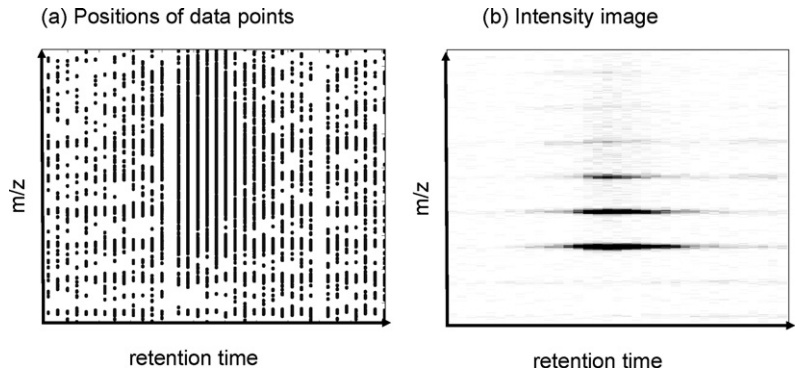

Fig. 2. (a) Positions of data points in a small *m/z* and retention time region of the whole profile mode raw data. (b) Corresponding two-dimensional intensity image created by binning *m/z* values to bitmap's resolution. Image is drawn over the same region as in (a) using MZmine software [36].

2.2. Raw data preprocessing and filtering

A frequently used first step in raw data processing is centroiding, which can be done already during data acquisition. Centroiding processes each histogram separately and combines multiple data points representing the same peak in the histogram into a single data point with a one *m/z* and intensity value. Using centroided data in the data processing has the advantage of making data files smaller and easier to manage, but it may complicate noise level estimation in LC–MS data [21] and limit available strategies for feature detection.

Handling LC–MS data in its raw form is difficult because histograms are typically non-uniformly sampled, and sampling intervals may not be same between histograms. To enable processing or visualization of data in its native two-dimensional form, all *m/z* values in histograms must be binned to fixed *m/z* values. Ion intensity for a fixed *m/z* value can be defined as sum of all intensities binned together or alternatively computed with interpolation from a continuous spectrum. As a result, the data is transformed into a two-dimensional matrix, with one index corresponding to the retention time scans, and another to fixed *m/z* values (Fig. 2). Matrix values represent the ion intensities. This matrix representation facilitates processing data using, for example, two-dimensional filter masks, with the possible drawback of losing resolution in *m/z* domain.

LC–MS data contains both chemical noise and random noise. Chemical noise is typically caused by molecules in buffers and solvents and can be especially strong at the beginning and end of

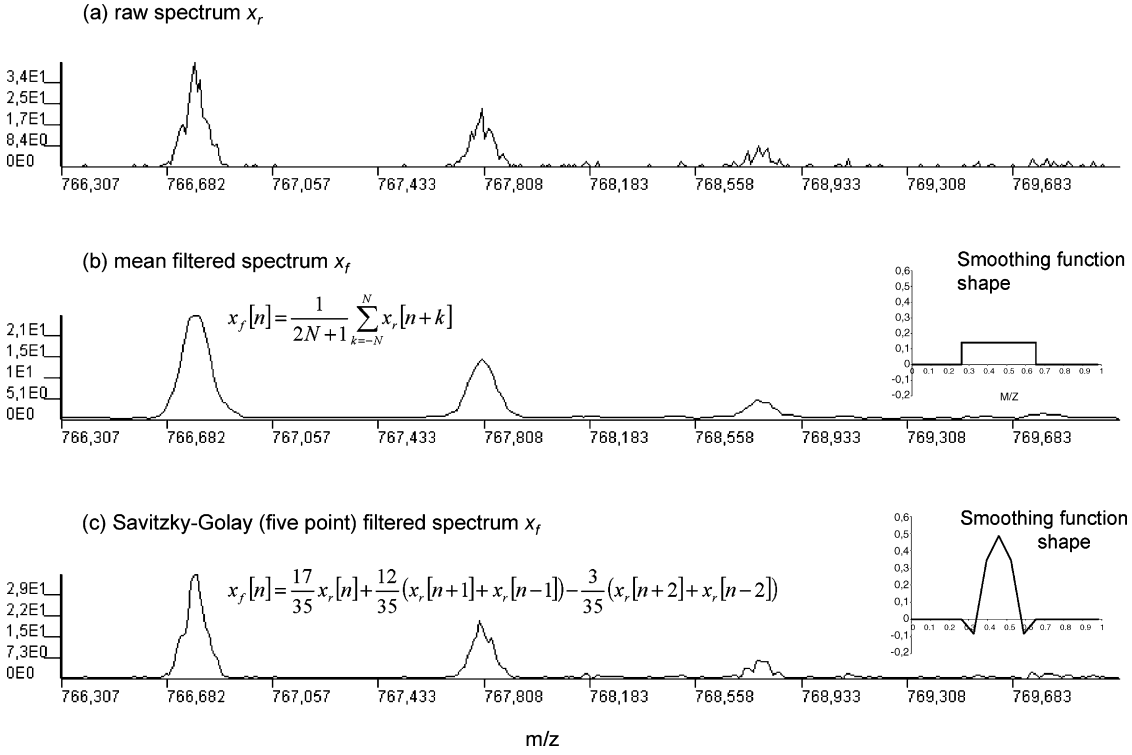

Fig. 3. Examples of raw data filtering in chromatographic direction for a specific *m/z* value. (a) Raw data, (b) mean filter with window width 0.1 Da, and (c) Savitzky-Golay filter with five data points.

the elution [19], while the random noise is mainly attributed to the detector. Noise reduction methods aim at removing random noise from the measurement signal. These methods are typically implemented using traditional signal processing techniques such as filtering with moving average window [22], median filter [21,22] in chromatographic direction and Savitzky-Golay type of local polynomial fitting [23] and wavelet transformation [24] in *m*/*z* direction. Illustrative examples of spectrum filtering are shown in Fig. 3.

In addition to sharp and random distortions, the quality of LC–MS data may also be affected by chemical noise, causing a shift in the baseline in the intermediate mass range in LC–MS spectra [19]. Baseline removal is typically a two-step process: (1) finding the baseline shape and (2) subtracting the shape from the raw signal. For example, Haimi et al. estimate the baseline by first segmenting a spectrum and then performing a linear regression through the lowest points of smoothed spectrum segments [25]. Other one-dimensional background estimation methods are low-order polynomial Savitzky-Golay filter [23] or iterative asymmetric least-squares estimation [26]. Baseline removal has also been approached by estimate background from a two-dimensional intensity image and then removing it with two orthogonal (retention time and *m*/*z*) one-dimensional passes [27].

2.3. Feature detection

The purpose of feature detection stage of data processing is to identify all signals caused by true ions and avoid detection of false positives. This step also aims to provide as accurate quantitative information about ion concentrations as possible. Feature detection is an essential step in the metabolomic data processing pipeline, yet in practice rarely performed perfectly. This is therefore an important area for further method development.

There are three main strategies for solving the feature detection problem (Fig. 4).

The first strategy performs detection in two directions by finding peaks independently in both *m*/*z* and retention time direction. Such *vectorized peak detection* method searches for data points with intensity above a threshold level in two directions, and data points that meet these criteria are defined as peaks [21,23]. The threshold level is determined using all intensity values along the vector in one direction. Similarly, feature detection can be done independently in two directions with additional constraints on allowed peak shapes in chromatographic direction [28]. Bellew et al. developed a feature detection method that works in three steps [27]: (1) identify local maxima within each scan using wavelet additive decomposition, (2) smooth peaks over time, and identify peaks that are sustained over multiple scans and (3) assemble all peaks into isotope groups that appear, maximize and disappear at the same time.

The second strategy is slicing data to extracted ion chromatograms (XIC), with each one covering a narrow *m*/*z* range, therefore avoiding the problem of searching for peaks in *m*/*z* direction. These chromatograms can be processed independently in time domain using second-order Gaussian filter to find peak inflection points for integration [29] or by calculating a threshold level based on mean or median of chromatogram and searching for areas in chromatogram above the threshold level [22].

The third strategy for feature extraction is model fitting against the original raw signal. One such approach is fitting a three-dimensional model of a generic isotope pattern to highest peak in raw signal and subtracting the fit from signal [30]. This process is repeated iteratively until highest remaining peak is near the background level. Also Leptos et al. use model fitting with two-dimensional data [31]. Detecting entire isotope patterns instead of individual peaks may improve detection results by reducing the number of detected false positive noise peaks.

Because of challenges in feature detection, direct comparison of raw data has been proposed as an alternative to the feature detection step. A simple approach is to compare data points in raw chromatograms directly [32]. Direct comparison of two-dimensional gas-chromatograms has been applied by Shellie et al. [33]. Another proposed alternative is to reverse the order of feature detection and alignment steps [34]. In such case, feature detection is performed on the merged raw data from pairs of samples. Peaks that match three simple conditions are searched: (1) peak must be above threshold, (2) corresponding peak must have high intensity also in nearby spectra, and (3) there must be another peak within the isotopic range of the peak. Also Baran et al. [35] have presented a method for direct comparison of hyphenated mass spectrometry data. This method visualizes differences between two or multiple datasets as a two-dimensional

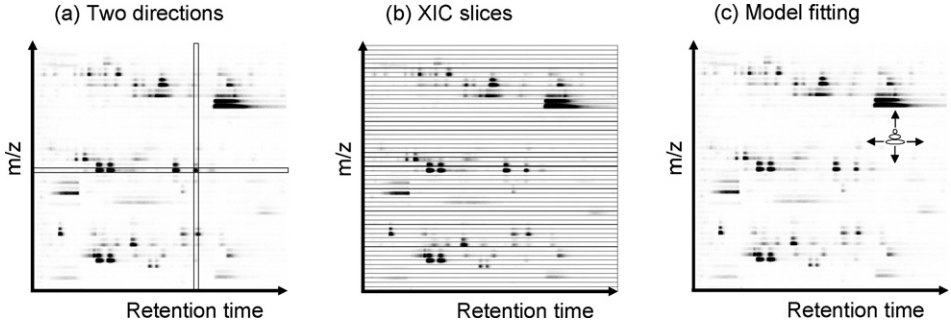

Fig. 4. Examples of peak detection strategies. (a) Peak detection performed separately in two dimensions (retention time and *m*/*z*), (b) by slicing the whole dataset to extracted ion chromatograms and processing them independently, or (c) fitting a model to the data.

plot and also uses statistical tests for finding differences between two groups of replicate samples.

Although the focus on single peaks as features is often an appropriate choice when soft ionization techniques such as electrospray ionization are used, this is generally not the case. Generally features should combine information from isotopic peaks of an ion, ion adducts, different charge states and also fragment ions of a compound. Isotope pattern detection can be done after feature detection by using pattern matching with raw data [23] or grouping detected features with suitable *m/z* differences [36]. Alternatively, isotope pattern detection can be included already in the feature detection step by fitting a model of a generic isotope pattern to raw signal [30]. Multiple ions may correspond to different fragments from the same molecule, making quantitative analysis a challenge. Deconvolution methods are therefore needed which can assign different ions to the same metabolite. Such methods have been, for example, widely used in GC–MS data processing [37–39]. Deconvolution algorithms commonly utilize the fact that different fragments from the same molecule have the same retention time as well as on assumption that their profiles across multiple samples are highly correlated as they are subject to the same biological variation and systematic error. The main challenge in the use of deconvolution methods is that metabolomic experiments on complex biological matrices lead to a large number of overlapping peaks, with similar retention times and overlapping isotope patterns. Additionally, several metabolites may be subject to the same regulatory mechanisms in a biological system, therefore their levels are highly correlated. The latter may lead to difficulties in ability to separate systematic error and biological variability, which is needed for successful deconvolution.

2.4. Alignment

Despite the advances in chromatographic techniques used for metabolomics, there is always some variation in retention times of a metabolite across different sample runs. Alignment is needed for correcting retention time differences between runs and combining data from different samples.

Most alignment methods work in pair-wise fashion by aligning either only pairs of samples or multiple samples against a selected reference sample or a template. In general, the choice of reference sample has effect on the alignment results. Alignment methods can be roughly divided to two categories: (1) methods which use raw data as input material and generate a set of mappings that transform retention time axis of each run to a common retention time axis (Fig. 5a and b), and (2) methods that cluster detected features and produce a matrix where each row corresponds to one cluster (i.e. an ion) and columns contain some measurement (e.g., peak area) for each sample (Fig. 5c). Some alignment methods combine both approaches, e.g., by first conducting a retention time mapping between runs

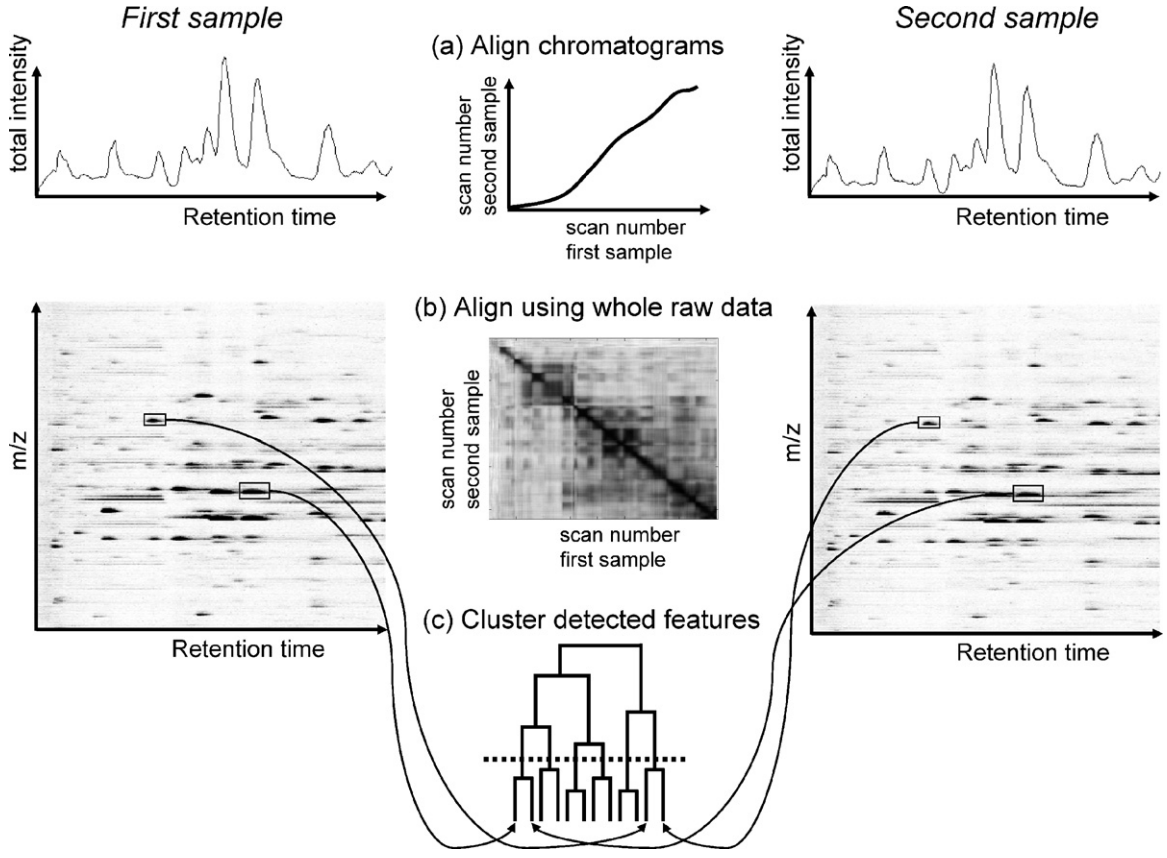

Fig. 5. Examples of two-step alignment strategies. First step: (a) alignment using information from chromatographic profiles or (b) alignment using correlation matrix calculated from entire raw data. (c) Clustering of detected features.

and then clustering detected features using corrected retention times.

Choice of alignment method usually dictates the type of required downstream data analysis. Alignment methods in the first category require comparing the aligned raw signals for finding differences between samples, while methods in the second category lead to multivariate data analysis, similar to microarray expression data analysis in transcriptomics.

A simple, much explored alignment strategy is mapping the retention time axis of one total ion chromatogram (TIC) to another. The correlation optimized warping (COW) method developed by Nielsen et al. is one well-known method for this task [40]. In addition to the COW method, various other types of warping algorithms for aligning chromatograms have been examined [26,41–43]. Also Fast Fourier Transform has been applied to alignment of chromatographic data [44]. Three common approaches for chromatographic alignment were recently compared [45].

Since total ion chromatogram represents a collapsed view of the raw data, using the full two-dimensional raw data can potentially lead to better alignment results. Examples of alignment methods that use the entire raw data include method by Prakash et al. [34] and ChromAlign method [46]. Both algorithms first compute a matrix of similarity scores between pairs of spectra in two runs. Dynamic programming is then used to find the optimal path through the matrix, also defining the mapping of spectra between the two runs. In the method by Pierce et al. a piecewise single dimension retention time alignment algorithm is adapted for aligning two-dimensional data [47]. The continuous profile model (CPM) method divides available two-dimensional data to four *m*/*z* bins instead of aligning only a single total ion chromatogram [48].

A simple approach for aligning detected features is clustering of chromatographic peaks without correction of retention time [49]. This method first clusters peaks within each group of replicates and then between two groups of replicates. Another straightforward method aligns detected features pair-wise between each peak list and a dynamically grown template [28]. The alignment is based on proximity in *m*/*z* and uncorrected retention time. Alignment methods that cluster detected features without implementing a retention time correction generally require high reproducibility of chromatography as they can only account for small variations in retention times.

Two-step alignment methods are required when it is necessary to correct retention times between runs before clustering detected features. An approach introduced by Bellew et al. requires user to choose a reference sample against which retention times are mapped non-linearly from every file [27]. To establish the mapping, a linear mapping is first created using highest intensity mass-matched features, followed by iterative application of smoothing-spline regression methods from the linear model residuals. Second step of the alignment method then matches features on basis of closeness of *m*/*z* and mapped retention time using divisive clustering separately on both properties.

Smith et al. presented an alignment method that seeks to align all samples simultaneously instead of pair-wise matching [29]. This avoids the problems related to selecting a good reference sample for pair-wise matching. In this method, detected features from all samples are first clustered into groups which can be used as temporary standards. Inside each cluster, a retention time shift can be computed for each sample as difference between a feature and cluster's median. Shifts obtained from different clusters can be used in creating a non-linear retention time deviation contour for each sample.

Nordström et al. divide alignment methods into three main strategies: (1) alignment of chromatograms, (2) curve resolution (peak detection and matching between samples) and (3) summation or binning of chromatographic data [50]. Examples of methods in the two first categories have been presented in this review. Methods of the third category bin data to time windows in chromatographic direction, moving possible alignment errors to the borders of the windows. During the later data analysis stage only window segments are then studied for potentially interesting differences.

If detected features can be identified in each sample, the identifications can be used to help alignment or even to avoid the alignment stage. Such strategy is particularly appealing in proteomics and lipidomics, because automatic identification for peptides and lipids is generally easier than for other metabolites. Higgs et al. identify a set of landmark peptides from each sample which are used to calculate the retention time shift in the local neighborhood of each landmark [51]. Hermansson et al. describe a lipidomics data processing approach without alignment, which includes a method for automatically assigning lipid classes to detected features [30]. Automatic assignments require a set of manually assigned reference compounds on a reference chromatogram and a polynomial function relating retention time and fatty acid chain length, which is used to predict retention time shifts inside each lipid class.

2.5. Normalization

The goal of normalization is to remove the unwanted systematic bias in ion intensities between measurements, while retaining the interesting biological variation. Chemical diversity of metabolites, leading, for example, to different recoveries during extraction or responses during ionization in mass spectrometer, makes separation of interesting biological variation and unwanted systematic bias a difficult task.

Strategies for normalization of metabolic profile data can be divided into two major categories:

(1) Statistical models used to derive optimal scaling factors for each sample based on complete dataset [52], such as normalization by unit norm [53] or median [23] of intensities, or the maximum likelihood method [2].
(2) Normalization by a single or multiple internal (i.e., added to sample prior to extraction) or external (i.e., added to sample after extraction) standard compounds based on empirical rules, such as specific regions of retention time [30,54].

The statistical approach suffers from the lack of an absolute concentration reference for different metabolites. In addition,

constraining the data to a specific norm based on total signal affects its covariance structure, therefore requiring special caution when pursuing multivariate analysis of such data [55]. Metabolites as physiological end-points, largely affected by the environment, do not posses the *self-averaging* property, i.e. concentration increase in a specific group of metabolites is generally not balanced by a decrease of another group.

Quantitative analytical methods have commonly relied on utilization of isotope labeled internal standard for each metabolite measured. However, in broad metabolic profiling approaches this is not practical. The number of metabolites is large, they are chemically too diverse to afford a common labeling approach, and many of them may not even be known. The availability of stable isotope labeled references is generally also very limited. Recent studies in *in vivo* labeling of microbial metabolome offer a promising solution to this bottleneck for microbial metabolomics. In such studies, the standardized extracts grown on stable isotope labeled medium and thus containing mainly isotope labeled metabolites, can provide standards for most of the measured metabolite [56,57].

Using a set of selected internal and external standards is an alternative when a full set of isotope labeled standard is not available. In such case, the *assignment* of the standards to normalize specific peaks remains unclear. One possible approach is to assign a specific standard to metabolite peaks based on similarity in specific chemical property such as retention time in liquid chromatography (LC) column. For example, Bijlsma et al. utilize three external standard references for lipid profiling, chosen as mono-, di-, and triacyl lipid species representing most abundant lipid classes in their respective region of retention time [54]. A related approach normalizes metabolites based on multiple internal standards, with the normalization factor based on distance to the metabolite peaks both in the retention time and *m*/*z* [28]. However, such approaches still suffer from the problem that retention time or mass-to-charge ratios are not necessarily descriptive of all matrix and chemical properties leading to obscuring variation. For example, in the lipid separation based on reversed-phase LC diverse lipid species such as ceramides, sphingomyelins, diacylglycerols, and several phospholipid classes, are overlapping in retention time, and it is not

Table 1
Commerical software tools for metabolomic data processing

Name	Vendor	Features	Main application field and examples
BlueFuse	BlueGnome, Cambridge, UK	Filtering, peak detection, and alignment. Univariate and multivariate methods for data analysis	Metabolomics with MS and NMR data
Chenomx NMR Suite	Chenomx, Edmonton, Canada	Data conversion, analyzing spectra to compounds and concentrations	Metabonomics with NMR data [63]
Genedata Expressionist	Genedata, Basel, Switzerland	Filtering, peak extraction, *m*/*z* and retention time alignment. Metabolite identification using third-party databases. Includes also analysis and interpretation modules and integrated database	Cross-omics platform for transcriptomics, proteomics and metabolomics. Metabolomics module works with MS data
LineUp	Infometrix	Alignment of chromatographic data. Alignment can be used also for spectroscopic data	Chromatographic alignment
MarkerLynx	Waters, Milford, MA, USA	Peak detection and alignment. Principal component analysis (PCA)	Metabonomics with LC–MS data [61,64,65]
MarkerView	Applied Biosystems, Foster City, CA, USA	Peak detection and alignment. PCA and *t*-test methods for data analysis. Visualization and reporting	Metabolomics with LC–MS data [66]
MassHunter Profiling software	Agilent Technologies, Santa Clara, CA, USA	Feature extraction and alignment	Proteomics with LC–MS data
Metabolic Profiler	Bruker Daltonic & Bruker BioSpin, Billerica, MA, USA	Bucket raw data into retention time, *m*/*z* table with intensities. Identification using libraries. PCA for data analysis	Metabolomics with MS and NMR
metAlign	PlanResearch International B.V., Wageningen, The Netherlands	Filtering, baseline correction, peak detection, alignment	Broad, LC–MS and GC–MS data [59,67–69]
MS Resolver	Pattern Recognition Systems, Bergen, Norway	Resolve multicomponent data from multidetection instrumentation into individual contributions	Broad, LC–MS and GC–MS data [61,70]
Profile	Phenomenome Discoveries, Saskatoon, Canada	File conversion, peak detection and alignment. Tools for statistical analysis and data mining	Metabolomics with MS data
Rosetta Elucidator	Rosetta Biosoftware, Seattle, WA, USA	Peak detection and alignment, statistical analysis and visualization	Proteomics with LC–MS data
Sieve	Thermo Fisher Scientific, Waltham, MA, USA	Direct comparison approach to comparing multiple LC–MS datasets. Uses ChromAlign [46] for chromatographic alignment	Proteomics with LC–MS data

reasonable to assume same normalization factor can be applied to all these species.

Recently a promising new method was introduced: normalization using optimal selection of multiple internal standards (NOMIS), that learns the optimal assignment of internal or external standard peaks for each other detected peak in the sample from a repeatability study [58]. The basic premise of the NOMIS approach is that monitoring of multiple standard compounds across multiple sample runs may help determine how the standards are correlated, which variation is specific to a particular standard, and which patterns of variation are shared between the measured metabolites and the standards so they can be removed. Based on this premise, a statistical model was developed that models the systematic variation of metabolites as a function of variation of standard compounds.

3. Software tools for data processing

Increasing demand for better metabolomic data processing methods led to a number of software packages to meet the challenge. New tools are still being released and also some of the existing tools are still under further development. Available solutions can be roughly divided to two categories: commercial (Table 1) and freely available (Table 2) tools.

One difference between the groups of commercial and free software can be seen in the transparency of the implementation.

Table 2
Freely available software for metabolomic data processing

Name	Features	Main application field and examples	License type	Platform
Chrompare [71]	Comparison of chromatographic peak lists and raw chromatograms, automatic and manual normalization	Metabolomics with GC-FID data	Licence type unknown, available for download	Microsoft Excel Visual Basic
COMSPARI [72]	Visualization to aid searching for differences between pair of runs	Metabolomics with LC–MS and GC–MS data [73]	GNU General Public License	Implemented in C, for any recent platform including Linux and Windows
Continuous profile models [48]	Alignment and normalization of time series data	Proteomics with LC–MS data	Free for educational and research use, source code available	Toolbox for Matlab
HiRes [74]	Processing and analysis spectral data	Metabolomics with NMR data	Free for research and clinical purposes	Implemented in C++, for Windows platform
LCMSWARP [75]	Retention time alignment and feature clustering	Proteomics with LC–MS and LC–MS/MS data [76]	Unknown	Implemented in C++
MapQuant [31]	Noise filtering, peak detection and visualization.	Proteomics with LC–MS data	Harvard University open-source compatible license	Implemented in C, for Windows and Linux platforms
MathDAMP [35]	Direct comparison of raw data sets without peak picking. Includes methods for preprocessing (binning, baseline subtraction, smoothing) and normalization	Metabolomics with LC–MS, GC–MS and CE–MS data	Free	Package to Mathematica
MET-IDEA [77]	Extracts ion intensity data for listed ion/retention time values from multiple runs	Metabolomics with LC–MS, GC–MS and CE–MS data	Freely available to academic users upon request	Windows, .NET platform
MSFACTs [8]	Alignment and comparison of raw chromatograms or peak lists generated with a third-party software	Metabolomics with GC–MS and LC–MS data	Freely available upon request for academic and non-commercial use	Implemented in Java
MSight [78]	Visualization and visual analysis and comparison of multiple runs	Proteomics with LC–MS data	Free of charge	Windows platform
msInspect [27]	Peak detection, alignment, normalization and visualization	Proteomics with LC–MS data	Free software available under Apache 2.0 License	Implemented in Java. Requires R statistical language
MZmine [36]	Noise filtering, peak detection, alignment, normalization and visualization. Distributed computing	Metabolomics with LC–MS and GC–MS data [62,79–81]	GNU General Public License	Implemented in Java
SpecArray [24]	Noise filtering, centroiding, peak detection, alignment and visualization	Proteomics with LC–MS data	GNU General Public License	Implemented in C, for Linux platform
SuperHirn (Mueller, submitted)	Peak detection, alignment and normalization. Includes also analysis capabilities	Proteomics with LC–MS data	Not yet available	Implemented in C++, for Unix platforms
Xalign [82]	Peak detection, alignment between samples and quality control	Proteomics with LC–MS data	Available upon request from the author	Implemented in C++
XCMS [29]	Noise filtering, peak detection and alignment	Metabolomics with LC–MS and GC–MS data [62,83,84]	GNU General Public License	Implemented in R statistical language

Freely available tools are frequently released as open-source, which means that the details of the used algorithms are available for review and also for further development. Commercial solutions typically do not reveal at least some aspects of their implementation, although they cannot be generally categorized as black boxes. For example, Sieve software (Thermo Fisher Scientific, Waltham, MA, USA) includes published ChromAlign alignment method [46]. Several commercial tools, for example Rosetta's Elucidator (Rosetta Biosoftware, Seattle, WA, USA), contain scripting capabilities for incorporating new analysis methods and automation.

The applicability range of available software tools is broad: the smallest tools are tailor-made for a specific task like chromatographic alignment, while the biggest software suites combine everything from instrument control to data analysis into a single package. The data processing approach shared by the majority of the software tools includes the standard steps for filtering, feature detection, alignment and normalization. Some tools such as MathDAMP [35] and Sieve support the analysis approach using direct raw data comparison without a preceding feature detection step.

The software tools usually produce their output as a matrix containing peak intensities for all detected ions in different samples, which can be processed with various statistical tools. Especially the commercial solutions also include some in-built statistical methods for the downstream data analysis. The most commonly available methods for data analysis are the principal component analysis (PCA) used for projecting multivariate data to a low-dimensional plot, and standard statistical tests such as *t*-test. Availability of elementary data analysis options inside the software gives possibility to have an overview of the processed data and also helps in going back from analysis results to raw data, which is more laborious by using external statistical tools.

Some aspects to consider in choosing the software for metabolomic data processing are quality of processing, ease of use, performance and overall cost of the software. Quality of processing can be difficult to assess since there there is not any generally accepted benchmark [59]. Programs with a graphical user interface are easier to start using, but in the long run, also scripting and batch processing capabilities are usually needed. Data processing of a huge set of raw data measurements can be time consuming. For speeding up the processing, MZmine tool includes capabilites for distributing the computation to multiple processors or computers [36]. This can be useful feature especially when running data processing on a multi-processor workstation. However, on a large, shared distributed computing environment a command-line based software consisting of independent modules for each processing step is likely going to be easier solution to set up that an integrated software package with build-in distributed computing capabilities.

For proteomics software tools, there exists a recent review article that covers some of the tools listed here [60]. There are also a few articles comparing software tools by applying them on the same dataset: MarkerLynx and MS Resolver [61], as well as MZmine and XCMS [62]. In the Internet, listings of available software tools can be found at MS-Utils website (http://www.ms-utils.org/) and the Fiehn laboratory website (http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/).

4. Conclusions

Development of new methods and software tools for metabolomics has been an active area of research during recent years. Several new methods for different stages of metabolomic data processing were examined in this review. Since improvements are still needed at each stage of data processing, the method development trend is likely going to continue also in the coming years. One path leading to future innovative solutions can be seen in combining multiple processing stages together, for example metabolite identification with feature extraction and alignment.

Since each new data processing algorithm needs to be evaluated as part of the whole data processing chain, software packages should ideally support incorporation of new algorithms developed by researchers. Open source strategy is one way to address this need, and softwares such as MapQuant [31], msInspect [27], MZmine [28,36], SpecArray [24] and XCMS [29] will likely play an important role in facilitating progress in metabolomic data processing.

As data processing methods and software implementations for metabolomics mature, there will be increasing need for generation of reference datasets so that the methods could be objectively compared. Current standardization efforts such as the Metabolomics Standards Initiative supported by the Metabolomics Society (http://www.metabolomicssociety.org/) aim to provide such datasets and standards.

References

[1] L.M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M.C. Walsh, J.A. Berden, K.M. Brindle, D.B. Kell, J.J. Rowland, H.V. Westerhoff, K. van Dam, S.G. Oliver, Nat. Biotechnol. 19 (2001) 45.
[2] M. Oresic, C.B. Clish, E.J. Davidov, E. Verheij, J. Vogels, L.M. Havekes, E. Neumann, A. Adourian, S. Naylor, J. van der Greef, T. Plasterer, Appl. Bioinform. 3 (2004) 205.
[3] M. Oresic, A. Vidal-Puig, V. Hanninen, Expert Rev. Mol. Diagn. 6 (2006) 575.
[4] L. Pauling, A.B. Robinson, R. Teranishi, P. Cary, Proc. Natl. Acad. Sci. U.S.A. 68 (1971) 2374.
[5] J. van der Greef, E. Davidov, E. Verheij, J.T.W.E. Vogels, R. van der Heijden, A.S. Adourian, M. Oresic, E.W. Marple, S. Naylor, in: G.G. Harrigan, R. Goodacre (Eds.), The Role of Metabolomics in Systems Biology, Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis, Kluwer, Boston, MA, 2003, p. 171.
[6] J. van den Greef, P. Stroobant, R. van der Heijden, Curr. Opin. Chem. Biol. 8 (2004) 559.
[7] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Trends Biotechnol. 22 (2004) 245.
[8] A.L. Duran, J. Yang, L. Wang, L.W. Sumner, Bioinformatics 19 (2003) 2283.
[9] J. Listgarten, A. Emili, Mol. Cell. Proteomics 4 (2005) 419.
[10] R.J. Bino, R.D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B.J. Nikolau, P. Mendes, U. Roessner-Tunali, M.H. Beale, R.N. Trethewey, B.M. Lange, E.S. Wurtele, L.W. Sumner, Trends Plant Sci. 9 (2004) 418.
[11] I. Spasic, W.B. Dunn, G. Velarde, A. Tseng, H. Jenkins, N. Hardy, S.G. Oliver, D.B. Kell, BMC Bioinformatics 7 (2006) 281.

[12] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A.R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R.J. Bino, R. Hall, J. Kopka, G.A. Lane, B.M. Lange, J.R. Liu, P. Mendes, B.J. Nikolau, S.G. Oliver, N.W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L.W. Sumner, T. Wang, S. Walsh, E.S. Wurtele, D.B. Kell, Nat. Biotechnol. 22 (2004) 1601.
[13] O. Fiehn, Plant Mol. Biol. 48 (2002) 155.
[14] K. Hollywood, D.R. Brison, R. Goodacre, Proteomics 6 (2006) 4716.
[15] R.D. Hall, New Phytol. 169 (2006) 453.
[16] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass Spectrom. Rev. (2006).
[17] I. Nobeli, J.M. Thornton, Bioessays 28 (2006) 534.
[18] V. Shulaev, Brief Bioinform. 7 (2006) 128.
[19] M. Hilario, A. Kalousis, C. Pellegrini, M. Muller, Mass Spectrom. Rev. 25 (2006) 409.
[20] 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. 22 (2004) 1459.
[21] C.A. Hastings, S.M. Norton, S. Roy, Rapid Commun. Mass Spectrom. 16 (2002) 462.
[22] D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, A. Emili, Mol. Cell. Proteomics 3 (2004) 984.
[23] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Anal. Chem. 75 (2003) 4818.
[24] X.J. Li, E.C. Yi, C.J. Kemp, H. Zhang, R. Aebersold, Mol. Cell. Proteomics 4 (2005) 1328.
[25] P. Haimi, A. Uphoff, M. Hermansson, P. Somerharju, Anal. Chem. 78 (2006) 8324.
[26] P.H. Eilers, Anal. Chem. 76 (2004) 404.
[27] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J.K. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, M. McIntosh, Bioinformatics 22 (2006) 1902.
[28] M. Katajamaa, M. Oresic, BMC Bioinformatics 6 (2005) 179.
[29] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.
[30] M. Hermansson, A. Uphoff, R. Kakela, P. Somerharju, Anal. Chem. 77 (2005) 2166.
[31] K.C. Leptos, D.A. Sarracino, J.D. Jaffe, B. Krastins, G.M. Church, Proteomics 6 (2006) 1770.
[32] T. Bamba, E. Fukusaki, J. Pestic. Sci. 31 (2006) 300.
[33] R.A. Shellie, W. Welthagen, J. Zrostlikova, J. Spranger, M. Ristow, O. Fiehn, R. Zimmermann, J. Chromatogr. A 1086 (2005) 83.
[34] A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, B. Schwikowski, Mol. Cell. Proteomics 5 (2006) 423.
[35] R. Baran, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert, M. Tomita, BMC Bioinformatics 7 (2006) 530.
[36] M. Katajamaa, J. Miettinen, M. Oresic, Bioinformatics 22 (2006) 634.
[37] S.E. Stein, J. Am. Soc. Mass Spectrom. 12 (1999) 770.
[38] J.M. Halket, A. Przyborowska, S.E. Stein, W.G. Mallard, S. Down, R.A. Chalmers, Rapid Commun. Mass Spectrom. 13 (1999) 279.
[39] A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, J. Chromatogr. A 1058 (2004) 209.
[40] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[41] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, J. Chromatogr. A 805 (2002) 17.
[42] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[43] J.T. Prince, E.M. Marcotte, Anal. Chem. 78 (2006) 6140.
[44] J.W. Wong, C. Durante, H.M. Cartwright, Anal. Chem. 77 (2005) 5655.
[45] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y. Vander Heyden, J. Chromatogr. A 1118 (2006) 199.
[46] R.G. Sadygov, F.M. Maroto, A.F. Huhmer, Anal. Chem. 78 (2006) 8207.
[47] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, Anal. Chem. 77 (2005) 7735.
[48] J. Listgarten, R.M. Neal, S.T. Roweis, P. Wong, A. Emili, Bioinformatics 23 (2007) e198.
[49] D.P. De Souza, E.C. Saunders, M.J. McConville, V.A. Likic, Bioinformatics 22 (2006) 1391.
[50] A. Nordström, G. O'Maille, C. Qin, G. Siuzdak, Anal. Chem. 78 (2006) 3289.
[51] R.E. Higgs, M.D. Knierman, V. Gelfanova, J.P. Butler, J.E. Hale, J. Proteome Res. 4 (2005) 1442.
[52] L.R. Crawford, J.D. Morrison, Anal. Chem. 40 (1968) 1464.
[53] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Bioinformatics 20 (2004) 2447.
[54] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Anal. Chem. 78 (2006) 567.
[55] J. Aitchison, The Statistical Analysis of Compositional Data, Blackburn Press, Caldwell, NJ, 2003, p. 416.
[56] L. Wu, M.R. Mashego, J.C. van Dam, A.M. Proell, J.L. Vinke, C. Ras, W.A. van Winden, W.M. van Gulik, J.J. Heijnen, Anal. Biochem. 336 (2005) 164.
[57] C. Birkemeyer, A. Luedemann, C. Wagner, A. Erban, J. Kopka, Trends Biotechnol. 23 (2005) 28.
[58] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, M. Oresic, BMC Bioinformatics 8 (2007) 93.
[59] O. Vorst, C.H.R. de Vos, A. Lommen, R.V. Staps, R.G.F. Visser, R.J. Bino, R.D. Hall, Metabolomics 1 (2005) 169.
[60] P.M. Palagi, P. Hernandez, D. Walther, R.D. Appel, Proteomics 6 (2006) 5435.
[61] H. Idborg, L. Zamani, P.O. Edlund, I. Schuppe-Koistinen, S.P. Jacobsson, J. Chromatogr. B 828 (2005) 14.
[62] T. Kind, V. Tolstikov, O. Fiehn, R.H. Weiss, Anal. Biochem. 363 (2007) 185.
[63] E.J. Saude, C.M. Slupsky, B.D. Sykes, Metabolomics 2 (2006) 113.
[64] M.E. Lenz, J. Bright, R. Knight, R.F. Westwood, D. Davies, H. Major, D.I. Wilson, Biomarkers 10 (2005) 173.
[65] P.D. Whitfield, P.J. Mantyla Nolbe, H. Major, R.J. Beynon, R. Burrow, A.I. Freeman, A.J. German, Metabolomics 1 (2005) 215.
[66] S. Wagner, K. Scholz, M. Sieber, M. Kellert, W. Voelkel, Anal. Chem. 79 (2007) 2918.
[67] Y. Tikunov, A. Lommen, C.H.R. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, Plant Physiol. 139 (2005) 1125.
[68] J.J.B. Keurentjes, J. Fu, C.H.R. de Vos, A. Lommen, R.D. Hall, R.J. Bino, L.H.W. van der Plas, R.C. Jansen, D. Vreugdenhil, M. Koornneef, Nat. Genet. 38 (2006) 842.
[69] R.J. Bino, C.H. Ric de Vos, M. Lieberman, R.D. Hall, A. Bovy, H.H. Jonker, Y. Tikunov, A. Lommen, S. Moco, I. Levin, New Phytol. 166 (2005) 427.
[70] I. Eide, G. Neverdal, B. Thorvaldsen, B. Grung, O.M. Kvalheim, Environ. Health Perspect. 110 (Suppl. 6) (2002) 985.
[71] T. Frenzel, A. Miller, K. Engel, Eur. Food Res. Technol. 216 (2003) 335.
[72] J.E. Katz, D.S. Dumlao, S. Clarke, J. Hau, J. Am. Soc. Mass Spectrom. 15 (2004) 580.
[73] J.E. Katz, D.S. Dumlao, J.I. Wasserman, M.G. Lansdown, M.E. Jung, K.F. Faull, S. Clarke, Biochemistry 43 (2004) 5976.
[74] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, T.R. Brown, Bioinformatics 22 (2006) 2562.
[75] N. Jaitly, M.E. Monroe, V.A. Petyuk, T.R. Clauss, J.N. Adkins, R.D. Smith, Anal. Chem. 78 (2006) 7397.
[76] A. Umar, T.M. Luider, J.A. Foekens, P.T. Ljiljana, Proteomics 7 (2007) 323.
[77] C.D. Broeckling, I.R. Reddy, A.L. Duran, X. Zhao, L.W. Sumner, Anal. Chem. 78 (2006) 4334.
[78] P.M. Palagi, D. Walther, M. Quadroni, S. Catherinet, J. Burgess, C.G. Zimmermann-Ivol, J.C. Sanchez, P.A. Binz, D.F. Hochstrasser, R.D. Appel, Proteomics 5 (2005) 2381.
[79] R. Laaksonen, M. Katajamaa, H. Päivä, M. Sysi-Aho, L. Saarinen, P. Junni, D. Lütjohann, J. Smet, R. Van Coster, T. Seppänen-Laakso, T. Lehtimäki, J. Soini, M. Oresic, PLoS ONE 1 (2006) e97.
[80] K.H. Pietiläinen, M. Sysi-Aho, A. Rissanen, T. Seppänen-Laakso, H. Yki-Järvinen, J. Kaprio, M. Oresic, PLoS ONE 2 (2007) e218.
[81] H. Rischer, M. Oresic, T. Seppanen-Laakso, M. Katajamaa, F. Lammertyn, W. Ardiles-Diaz, M.C. Van Montagu, D. Inze, K.M. Oksman-Caldentey, A. Goossens, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 5614.

[82] X. Zhang, J.M. Asara, J. Adamec, M. Ouzzani, A.K. Elmagarmid, Bioinformatics 21 (2005) 4054.
[83] E.P. Go, W.R. Wikoff, Z. Shen, G. O'Maille, H. Morita, T.P. Conrads, A. Nordstrom, S.A. Trauger, W. Uritboonthai, D.A. Lucas, K.C. Chan, T.D. Veenstra, H. Lewicki, M.B. Oldstone, A. Schneemann, G. Siuzdak, J. Proteome Res. 5 (2006) 2405.
[84] E.J. Want, C.A. Smith, C. Qin, K.C. Van, G. Siuzdak, Metabolomics 2 (2006) 145.