Generic Software Frameworks for GC-MS Based Metabolomics

Nils Hoffmann and Jens Stoye
Genome Informatics, Faculty of Technology, Bielefeld University
Germany

1. Introduction

Metabolomics has seen a rapid development of new technologies, methodologies, and data analysis procedures during the past decade. The development of fast gas- and liquid-chromatography devices coupled to sensitive mass-spectrometers, supplemented by the unprecedented precision of nuclear magnetic resonance for structure elucidation of small molecules, together with the public availability of database resources associated to metabolites and metabolic pathways, has enabled researchers to approach the metabolome of organisms in a high-throughput fashion. Other "omics" technologies have a longer history in high-throughput, such as next generation sequencing for genomics, RNA microarrays for transcriptomics, and mass spectrometry methods for proteomics. All of these together give researchers a unique opportunity to study and combine multi-omics aspects, forming the discipline of "Systems Biology" in order to study organisms at multiple scales simultaneously.

Like all other "omics" technologies, metabolomics data acquisition is becoming more reliable and less costly, while at the same time throughput is increased. Modern time-of-flight (TOF) mass spectrometers are capable of acquiring full scan mass spectra at a rate of 500Hz from 50 to 750 m/z and with a mass accuracy <5 ppm with external calibration (Neumann & Böcker, 2010). At the opposite extreme of machinery, Fourier-transform ion-cyclotron-resonance (FTICR) mass spectrometers coupled to liquid chromatography for sample separation reach an unprecedented mass accuracy of <1 ppm m/z and very high mass resolution (Miura et al., 2010). These features are key requirements for successful and unique identification of metabolites. Coupled to chromatographic separation devices, these machines create datasets ranging in size from a few hundred megabytes to several gigabytes per run. While this is not a severe limitation for small scale experiments, it may pose a significant burden on projects that aim at studying the metabolome or specific metabolites of a large number of specimens and replicates, for example in medical research studies or in routine diagnostics applications tailored to the metabolome of a specific species (Wishart et al., 2009).

Thus, there is a need for sophisticated methods that can treat these datasets efficiently in terms of computational resources and which are able to extract, process, and compare the relevant information from these datasets. Many such methods have been published, however there is a high degree of fragmentation concerning the availability and accessibility of these methods, which makes it hard to integrate them into a lab’s workflow.

The aim of this work is to discuss the necessary and desirable features of a software framework for metabolomics data preprocessing based on gas-chromatography (GC) and comprehensive two-dimensional gas-chromatography (GCxGC) coupled to single-dimension detectors (flame/photo ionization, FID/PID) or multi-dimension detectors (mass spectrometry, MS). We compare the features of publicly available Open Source frameworks that usually have a steep learning curve for end-users and bioinformaticians alike, owing to their inherent complexity. Many users will thus be appaled by the effort it takes to get used to a framework. Thus, the main audience of this work are bioinformaticians and users willing to invest some time in learning to use and/or program in these frameworks in order to set up a lab specific analytical platform. For a review of LC-MS based metabolomics data preprocessing consider (Castillo, Mattila, Miettinen, Orešič & Hyötyläinen, 2011).

Before we actually compare the capabilities of these different frameworks, we will first define a typical workflow for automatic data processing of metabolomics experiments and will discuss available methods within each of the workflow's steps.

We will concentrate on frameworks available under an Open Source license, thus allowing researchers to examine their actual implementation details. This distinguishes these frameworks from applications that are only provided on explicit request, under limited terms of use, or that are not published together with their source code (Lommen, 2009; Stein, 1999), which is still often the case in metabolomics and may hamper comparability and reuse of existing solutions. Additionally, all frameworks compared in this work are available for all major operating systems such as Microsoft Windows, Linux, and Apple Mac OSx as standalone applications or libraries.

Web-based methods are not compared within this work as they most often require a complex infrastructure to be set up and maintained. However, we will give a short overview of recent publications on this topic and provide short links to the parts of the metabolomics pipeline that we discuss in the following section. A survey of web-based methods is provided by Tohge & Fernie (2009). More recent web-based applications for metabolomics include the retention time alignment methods Warp2D (Ahmad et al., 2011) and ChromA (Hoffmann & Stoye, 2009), which are applicable to GC-MS or LC-MS data, and Chromaligner (Wang et al., 2010), which aligns GC and LC data with single-dimension detectors like FID.

Tools for statistical analysis of multiple sample groups and with different phenotypes have been reported by Kastenmüller et al. (2011). However, other tools aim to integrate a more complete metabolomics workflow including preprocessing, peakfinding, alignment and statistical analysis combined with pathway mapping information like MetaboAnalyst (Xia & Wishart, 2011), MetabolomeExpress (Carroll et al., 2010), or MeltDB (Neuweger et al., 2008). These larger web-based frameworks integrate other functionality for time-course analysis (Xia et al., 2011), pathway mapping (Neuweger et al., 2009; Xia & Wishart, 2010a) and metabolite set enrichment analysis (Kankainen et al., 2011; Xia & Wishart, 2010b).

In the Application section, we will exemplarily describe two pipelines for metabolomics analyses based on our own Open Source framework Maltcms: ChromA, which is applicable to GC-MS, and ChromA4D, which is applicable to data from comprehensive GCxGC-MS experiments. We show how to set up, configure and execute each pipeline using instructional datasets. These two workflows include the typical steps of raw-data preprocessing in metabolomics, including peak-finding and integration, peak-matching among multiple replicate groups and tentative identification using mass-spectral databases, as well as visualizations of raw and processed data. We will describe the individual steps of the workflows of the two application pipelines to give the reader a thorough understanding of the methods used by ChromA and ChromA4D.

Finally, we discuss the current state of the presented Open Source frameworks and give an outlook into the future of software frameworks and data standards for metabolomics.

2. A typical workflow for a metabolomics experiment

Metabolomics can be defined as the study of the metabolic state of an organism or its response to direct or indirect perturbation. In order to find differences between two or more states, for example before treatment with a drug and after, and among one or multiple specimens, the actual hypothesis for the experiment needs to be defined. Based on this hypothesis, a design for the structure of the experiments and their subsequent analysis can be derived. This involves, among many necessary biological or medical considerations, the choice of sample extraction procedures and preparation methods, as well as the choice of the analytical methods used for downstream sample analysis.

Preprocessing of the data from those experiments begins after the samples have been acquired using the chosen analytical method, such as GC-MS or LC-MS. Owing to the increasing amount of data produced by high-throughput metabolomics experiments, with large sample numbers and high-accuracy/high-speed analytical devices, it is a key requirement that the resulting data is processed with very high level of automation. It is then that the following typical workflow is applied in some variation, as illustrated in Figure 1.

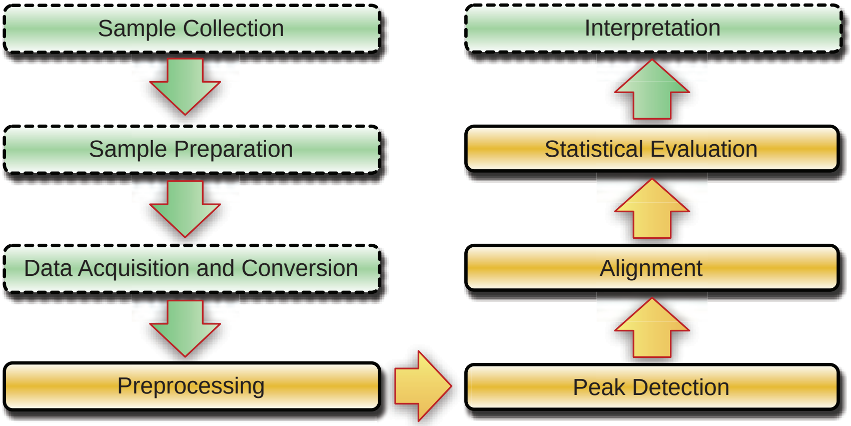


Fig. 1. A typical workflow for a metabolomics experiment. Steps shown in orange (solid border) are usually handled within the bioinformatics domain, while the steps shown in green (dashed border) often involve co-work with scientists from other disciplines.

2.1 Data acquisition and conversion

The most common formats exported from GC-MS and LC-MS machines today are NetCDF (Rew & Davis, 1990), based on the specifications in the ASTM/AIA standard ANDI-MS (Matthews, 2000), mzXML (Oliver et al., 2004), mzData (Orchard et al., 2005), and more recently as the successor to the latter two, mzML (Deutsch, 2008; Martens et al., 2010). All of these formats include well-defined data structures for meta-information necessary to interpret data in the right context, such as detector type, chromatographic protocol, detector potential and other details about the separation and acquisition of the data. Furthermore, they explicitly model chromatograms and mass spectra, with varying degrees of detail.

NetCDF is the oldest and probably most widely used format today. It is routinely exported even by older machinery, which offers backwards compatibility to those. It is a general-purpose binary format, with a header that describes the structure of the data contained in the file, grouped into variables and indexed by dimensions. In recent years, efforts were made to establish open formats for data exchange based on a defined grammar in extensible markup language (*XML*) with extendable controlled vocabularies, to allow new technologies to be easily incorporated into the file format without breaking backwards compatibility. Additionally, XML formats are human readable which narrows the technology gap. mzXML was the first approach to establish such a format. It has been superseded by mzData and, more recently, mzML was designed as a super-set of both, incorporating extensibility through the use of an indexed controlled vocabulary. This allows mzML to be adapted to technologies like GCxGC-MS without having to change its definition, although its origins are in the proteomics domain. One drawback of XML-based formats is often claimed to be their considerably larger space requirements when compared to the supposedly more compact binary data representations. Recent advances in mzML approach this issue by compressing spectral data using gzip compression.

The data is continuously stored in a vendor-dependent native format during sample processing on a GC-MS machine. Along with the mass spectral information, like ion mass (or equivalents) and abundance, the acquisition time of each mass spectrum is recorded. Usually, the vendor software includes methods for data conversion into one of the aforementioned formats. However, especially when a high degreee of automation is desired, it may be beneficial to directly access the data in their native format. This avoids the need to run the vendor's proprietary software manually for every data conversion task. Both the ProteoWizard framework (Kessner et al., 2008) and the Trans Proteomic Pipeline (Deutsch et al., 2010) include multiple vendor-specific libraries for that use case.

2.2 Preprocessing

Raw mass specrometry data is usually represented in sparse formats, only recording those masses whose intensities exceed a user-defined threshold. This thresholding is usually applied within the vendor's proprietary software and may lead to artificial *gaps* within the data. Thus, the first step in preprocessing involves the binning of mass spectra over time into bins of defined size in the m/z dimension, followed by interpolation of missing values. After binning, the data is stored as a rectangular array of values, with the first dimension representing time, the second dimension representing the approximate bin mass values, and the third dimension representing the intensity corresponding to each measured ion. This process is also often described as resampling (Lange et al., 2007).

Depending on various instrumental parameters, the raw exported data may require additional processing. The most commonly reported methods for smoothing are the Savitzky-Golay filter (Savitzky & Golay, 1964), LOESS regression (Smith et al., 2006) and variants of local averaging, for example by a windowed moving average filter. These methods can also be applied to interpolate values where gaps are present in the original data. The top-hat filter (Bertsch et al., 2008; Lange et al., 2007) is used to remove a varying baseline from the signal. More refined methods use signal decomposition and reconstruction methods, such as Fourier transform and continuous wavelet transform (CWT) (Du et al., 2006; Fredriksson et al., 2009; Tautenhahn et al., 2008) in order to remove noise and baseline contributions from the signal and simultaneously find peaks.

2.3 Peak detection

Often the process of peak detection is decoupled from the actual preprocessing of the data. XCMS (Smith et al., 2006), for example, uses a Gaussian second derivative peak model with a fixed kernel width and signal-to-noise threshold to find peaks along the chromatographic domain of each ion bin. Other methods extend this approach to use a multi-scale continuous wavelet transform using such a kernel over various widths, tracking the response of the transformed signal in order to locate peak apex positions in scale-space before estimating the true peak widths based on the kernel scale with maximum response (Fredriksson et al., 2009; Tautenhahn et al., 2008). However, these methods usually allow only a small number of co-eluting peaks in different mass-bins, since they were initially designed to work with LC-MS data mainly, where only one parent ion and a limited number of accompanying adduct ions are expected. In GC-MS, electron-ionization creates rich fragmentation mass spectra, which pose additional challenges to deconvolution of co-eluting ions and subsequent association to peak groups. Even though its source code is not publicly available, the method used by AMDIS (Stein, 1999) has seen wide practical application and is well accepted as a reference by the metabolomics and analytical chemistry communities.

2.4 Alignment

The alignment problem in metabolomics and proteomics stems from the analytical methods used. These produce sampled sensor readings acquired over time in fixed or programmed intervals, usually called chromatograms. The sensor readings can be one- or multidimensional. In the first case, detectors like ultra violet and visible light absorbance detectors (UV/VIS) or flame ionization detectors (FID) measure the signal response as one-dimensional features, e.g. as the absorbance spectrum or electrical potential, respectively. Multi-dimensional detectors like mass spectrometers record a large number of features simultaneously, e.g. mass and ion count. The task is then to find corresponding and non-corresponding features between different sample acquisitions. This *correspondence problem* is a term used by Åberg et al. (2009) which describes the actual purpose of alignment, namely to find *true* correspondences between related analytical signals over a number of sample acquisitions. For GC-MS- and LC-MS-based data, a number of different methods have been developed, some of which are described in more detail by Castillo, Gopalacharyulu, Yetukuri & Orešič (2011) and Åberg et al. (2009). Here, we will concentrate on those methods that have been reported to be applicable to GC-MS. In principle, alignment algorithms can be classified into two main categories: peak- and signal-based methods. Methods of the first type start with a defined set of peaks, which are present in most or all samples that are to be aligned before determining the best correspondences of the peaks between samples in order to then derive a time correction function. Krebs et al. (2006) locate *landmark* peaks in the TIC and then select pairs of those peaks with a high correlation between their mass spectra in order to fit an interpolating spline between a reference chromatogram and the to-be-aligned one. The method of Robinson et al. (2007) is inspired by multiple sequence alignment algorithms and uses dynamic programming to progressively align peak lists without requiring an explicit reference chromatogram. Other methods, like that of Chae et al. (2008) perform piecewise, block-oriented matching of peaks, either on the TIC, on selected masses, or on the complete mass spectra. Time correction is applied after the peak assignments between the reference chromatogram and the others have been calculated. Signal-based methods include recent variants of correlation optimized warping (Smilde & Horvatovich, 2008), parametric time warping (Christin et al., 2010) and dynamic time warping (Christin et al., 2010; Clifford et al., 2009; Hoffmann & Stoye, 2009; Prince & Marcotte, 2006) and usually consider the complete chromatogram for comparison. However, attempts are made to reduce the computational burden associated with a complete pairwise comparison of mass spectra by partitioning the chromatograms into similar regions (Hoffmann & Stoye, 2009), or by selecting a representative subset of mass traces (Christin et al., 2010). Another distinction in alignment algorithms is the requirement of an explicit reference for alignment. Some methods apply clustering techniques to select one chromatogram that is most similar to all others (Hoffmann & Stoye, 2009; Smilde & Horvatovich, 2008), while other methods choose such a reference based on the number of features contained in a chromatogram (Lange et al., 2007) or by manual user choice (Chae et al., 2008; Clifford et al., 2009). For high-throughput applications, alignments should be fast to calculate and reference selection should be automatic. Thus, a sampling method for time correction has recently been reported by Pluskal et al. (2010) for LC-MS. A comparison of these methods is given in the same publication.

2.5 Statistical evaluation

After peaks have been located and integrated for all samples, and their correspondence has been established, peak report tables can be generated, containing peak information for each sample and peak, with associated corrected retention times and peak areas. Additionally, peaks may have been putatively identified by searching against a database, such as the GMD (Hummel et al., 2007) or the NIST mass-spectral database (Babushok et al., 2007).

These peak tables can then be analyzed with further methods, in order to detect e.g. systematic differences between different sample groups. Prior to such an analysis, the peak areas need to be normalized. This is usually done by using a spiked-in compound which is not expected to occur naturally as a reference. The normalization compound is supposed to have the same concentration in all samples. The compound's peak area can then be used to normalize all peak areas of a sample with respect to it (Doebbe et al., 2010).

Different experimental designs allow to analyze correlations of metabolite levels for the same subjects under different conditions (paired), or within and between groups of subjects. For simple paired settings, multiple t-tests with corrections for multiple testing can be applied (Berk et al., 2011), while for comparisons between groups of subjects, Fisher's F-Statistic (Pierce et al., 2006) and various analysis of variance (ANOVA), principal component analysis (PCA) and partial least squares (PLS) methods are applied (Kastenmüller et al., 2011; Wiklund et al., 2008; Xia et al., 2011).

2.6 Evaluation of hypothesis

Finally, after peak areas have been normalized and differences have been found between sample groups, the actual results need to be put into context and be interpreted in their biological context. This task is usually not handled by the frameworks described in this chapter. Many web-based analysis tools allow to put the data into a larger context, by providing name- or id-based mapping of the experimentally determined metabolite concentrations onto biochemical pathways like MetaboAnalyst (Xia & Wishart, 2011), MetabolomeExpress (Carroll et al., 2010), or MeltDB (Neuweger et al., 2008). The latter allows association of the metabolomics data with other results for the same subjects under study or with results from other "omics" experiments on the same target subjects, but this is beyond the scope of the frameworks presented herein.

3. Frameworks for GC-MS analysis

A number of Open Source frameworks have been developed for LC-MS based proteomics frameworks like OpenMS (Bertsch et al., 2008), ProteoWizard (Kessner et al., 2008), and most notably the TransProteomicPipeline (Deutsch et al., 2010). Even though many of the steps required for proteomics apply similarily to metabolomics applications, there are still some essential differences due to the different analytical setups and technologies (e.g. matrix assisted laser desorption ionization mass spectrometry, MALDI-MS) used in the two fields. XCMS (Smith et al., 2006) was among the first frameworks to offer support for data preprocessing in LC-MS based metabolomics. Later, MZmine2 (Pluskal et al., 2010) offered an alternative with a user-friendly interface and easy extendability. Lately, Scheltema et al. (2011) published their PeakML format and mzMatch framework also for LC-MS applications. As of now, there seem to be only a few frameworks available for GC-MS based metabolomics that offer similar methods, namely PyMS (Callaghan et al., 2010; Isaac et al., 2009) and Maltcms/ChromA (Hoffmann & Stoye, 2009; *Maltcms*, 2011) . These will be presented in more detail in this section. A compact overview of the Open Source frameworks discussed herein is given in Table 1. A detailed feature comparison can be found in Table 2.

3.1 XCMS

XCMS (Smith et al., 2006) is a very mature framework and has seen constant development during the last five years. It is mainly designed for LC-MS applications, however its binning, peak finding and alignment are also applicable to GC-MS data. XCMS is implemented in the *GNU R* programming language, the de-facto standard for Open Source statistics. Since *GNU R* is an interpreted scripting language, it is easy to write custom scripts that realize additional functionality of the typical GC-MS workflow described above. XCMS is part of the Bioconductor package collection, which offers many computational methods for various "omics" technologies. Further statistical methods are available from *GNU R*.

XCMS supports input in NetCDF, mzXML, mzData and, more recently, mzML format. This allows XCMS to be used with virtually any chromatography-mass spectrometry data, since vendor software supports conversion to at least one of those formats. XCMS uses the *xcmsRaw* object as its primary tabular data structure for each binned data file. The *xcmsSet* object is then used to represent peaks and peak groups and is used by its peak alignment and *diffreport* features.

The peak finding methods in XCMS are quite different from each other. For data with normal or low mass resolution and accuracy, the matched filter peak finder (Smith et al., 2006) is usually sensitive enough. It uses a Gaussian peak template function with user defined width and signal-to-noise critera to locate peaks on individual binned extracted ion current

(EIC) traces over the complete time range of the binned chromatogram. The other method, CentWave (Tautenhahn et al., 2008) is based on a continuous wavelet transform on areas of interest within the raw data matrix. Both peak finding methods report peak boundaries and integrated areas for raw data and for the data reconstructed from the peak finder's signal response values.

Initially designed for LC-MS, XCMS does not have a method to group co-eluting peaks into peak groups, as is a requirement in GC-MS methods using electron ionization. However, CAMERA (Tautenhahn et al., 2007) shows how XCMS can be used as a basis in order to create a derived application, in this case for ion annotation between samples.

Peak alignment in XCMS is performed using local LOESS regression between peak groups with very similar m/z and retention time behaviour and good support within each sample group. This allows a simultaneous alignment and retention time correction of all peaks. The other available method is based on the Obi-Warp dynamic time warping (Prince & Marcotte, 2006) algorithm and is capable of correcting large non-linear retention time distortions. It uses the peak set with the highest number of features as alignment reference, which is comparable to the approach used by Lange et al. (2007). However, it is much more computationally demanding then the LOESS-based alignment.

XCMS's *diffreport* generates a summary report of significant analyte differences between two sample sets. It uses Welch's two-sample t-statistic to calculate p-values for each analyte group. ANOVA may be used for more than two sample sets.

A number of different visualizations are also available, both for raw and processed data. These include TIC plots, EIC plots, analyte group plots for grouped features, and chromatogram (rt, m/z, intensity) surface plots.

XCMS can use GNU R's Rmpi infrastructure to execute arbitary function calls, such as profile generation and peak finding, in parallel on a local cluster of computers.

3.2 PyMS

PyMS (Callaghan et al., 2010; Isaac et al., 2009) is a programming framework for GC-MS metabolomics based on the *Python* programming language. It can therefore use a large number of scientific libraries which are accessible via the SciPy and NumPy packages (*SciPy*, 2011). Since Python is a scripting language, it allows to do rapid prototyping, comparable to GNU R. However, Python's syntax may be more familiar for programmers with a background in object-oriented programming languages.

The downloadable version of PyMS currently only supports NetCDF among the more recent open data exchange formats. Nonetheless, it is the only framework in this comparison with support for the JCAMP GC-MS file format.

PyMS provides dedicated data structures for chromatograms, allowing efficient access to EICs, mass spectra, and peak data.

In order to find peaks, PyMS also builds a rectangular profile matrix with the dimensions time, m/z and intensity. Through the use of slightly shifted binning boundaries, they reduce the chance of false assignments of ion signals to neighboring bins, when binning is performed with unit precision (bin width of 1 m/z). PyMS offers the moving average and the Savitzky-Golay (Savitzky & Golay, 1964) filters for signal smoothing of EICs within the profile matrix. Baseline correction can be performed by the top-hat filter (Lange et al., 2007). The actual peak finding is based on the method described by Biller & Biemann (1974) and involves the matching of local peak maxima co-eluting within a defined window. Peaks are integrated for all co-eluting masses, starting from a peak apex to both sides and ending if the increase in area falls below a given threshold.

Peak alignment in PyMS is realized by the method introduced by Robinson et al. (2007). It is related to progressive multiple sequence alignment methods and is based on a generic dynamic programming algorithm for peak lists. It proceeds by first aligning peak lists within sample groups, before aligning the aligned peak lists of different groups, until all groups have been aligned.

Visualizations of chromatogram TICs, EICs, peaks and mass spectra are available and are displayed to the user in an interactive plot panel.

For high-throughput applications, PyMS can be used together with MPI to parallelize tasks within a local cluster of computers.

3.3 Maltcms

The framework *Maltcms* allows to set up and configure individual processing components for various types of computational analyses of metabolomics data. The framework is implemented in *JAVA* and is modular using the service provider pattern for maximal decoupling of interface and implementation, so that it can be extended in functionality at runtime.

Maltcms can read data from files in NetCDF, mzXML, mzData or mzML format. It uses a pipeline paradigm to model the typical preprocessing workflow in metabolomics, where each processing step can define dependencies on previous steps. This allows automatic pipeline validation and ensures that a user can not define an invalid pipeline. The workflow itself is serialized to XML format, keeping track of all resources created during pipeline execution. Using a custom post-processor, users can define which results of the pipeline should be archived.

Maltcms uses a generalization of the ANDI-MS data schema internally and a data provider interface with corresponding implementations to perform the mapping from any proprietary data format to an internal data object model. This allows efficient access to individual mass spectra and other data available in the raw-data files. Additionally, developers need no special knowledge of any supported file format, since all data can be accessed generically. Results from previous processing steps are referenced in the data model to allow both shadowing of data, e.g. creating a processing result variable with the same name as an already existing variable, and aggregation of processing results. Thus, all previous processing results are transparently accessible for downstream elements of a processing pipeline, unless they have been shadowed.

Primary storage of processing results is performed on a per-chromatogram basis in the binary NetCDF file format. Since metabolomics experiments create large amounts of data, a focus is put on efficient data structures, data access, and scalability of the framework.

Embedding Maltcms in existing workflows or interfacing with other software is also possible, as alignments, peak-lists and other feature data can be exported as comma separated value files or in specific xml-based formats, which are well-defined by custom schemas.

To exploit the potential of modern multi-core CPUs and distributed computing networks, Maltcms supports multi-threaded execution on a local machine or within a grid of connected computers using an OpenGrid infrastructure (e.g. Oracle Grid Engine or Globus Toolkit (Foster, 2005)) or a manually connected network of machines via remote method invocation (RMI).

The framework is accompanied by many libraries for different purposes, such as the *JFreeChart* library for 2D-plotting or, for BLAS compatible linear algebra, math and statistics implementations, the *Colt* and *commons-math* libraries. Building upon the base library *Cross*, which defines the commonly available interfaces and default implementations, Maltcms provides the domain dependent data structures and specializations for processing of chromatographic data.

Name	Version	Analytical method	Software license	Programming language
XCMS	1.26.1^a	LC-MS/GC-MS	GNU GPL v2	*GNU R* 2.13/C++
PyMS	r371	GC-MS	GNU GPL v2	*Python* 2.5
Maltcms/ChromA	1.1	GC-MS	GNU L-GPL v3	*JAVA* 6

Table 1. Overview of available Open Source software frameworks for GC-MS based metabolomics. a: Part of Bioconductor 2.8

Feature (GC-MS pipeline)	XCMS	PyMS	ChromA
Data formats	A, B, C, D	A, E	A, B, C, D
Signal preprocessing	MM	SG, TH	MA, MM, TH
Peak detection	MF, CWT	BB	MAX
Multiple peak alignment	LOESS, DTW	PROGDP	DTW, CLIQUE
Visualization	TIC, EIC, SURF	TIC, EIC	TIC, EIC, SURF
DB search	no (LC-MS only)	no	MSP
Normalization	no	no	RP, EV
Statistical evaluation	TT	no	FT

Table 2. Feature comparison of Open Source software frameworks for preprocessing of GC-MS based metabolomics data. Keys to abbreviations: **Data formats** A: NetCDF, B: mzXML, C: mzData, D: mzML, E: JCAMP GC-MS. **Signal preprocessing** MM: moving median, SG: Savitzky-Golay filter, TH: top-hat filter, MA: moving average. **Peak detection** MF: matched Gaussian filter, CWT: continuous wavelet transform, BB: Biller-Biemann, MAX: TIC local maxima. **Multiple peak alignment** LOESS: LOESS regression, DTW: dynamic time warping, PROGDP: progressive using dynamic programming, CLIQUE: progressive clique-based. **Visualization** (of unaligned and aligned data) TIC: plots of total ion chromatogram/peaks, EIC: plots of extracted ion chromatograms/peaks, SURF: surface plots of profile matrix (rt x m/z x I). **DB search** MSP: msp-format, compatible with AMDIS and GMD format. **Normalization** RP: reference peak area, EV: external value, e.g. dry weight. **Statistical evaluation** TT: groupwise t-test, multiple testing correction, FT: F-test, between group vs. within group variance

3.3.1 ChromA

ChromA is a configuration of Maltcms that includes preprocessing, in the form of mass binning, time-scale alignment and annotation of signal peaks found within the data, as well as visualizations of unaligned and aligned data from GC-MS and LC-MS experiments. The user may supply mandatory alignment anchors as CSV files to the pipeline and a database location for tentative metabolite identification. Further downstream processing can be performed either on the retention time-corrected chromatograms in NetCDF format, or on the corresponding peak tables in either CSV format or XML format.

Peaks can either be imported from other tools, by providing them in CSV format to ChromA, giving at least the scan index of each peak in a file per row. Alternatively, ChromA has a fast peak finder that locates peaks based on derivatives of the smoothed and baseline-corrected TIC, using a moving average filter followed by top-hat filter baseline-substraction, with a predefined minimum peak-width. Peak alignment is based on a star-wise or tree-based application of an enhanced variant of pairwise dynamic time warping (DTW) (Hoffmann & Stoye, 2009). To reduce both runtime and space requirements, conserved signals throughout the data are identified, constraining the search space of DTW to a precomputed closed polygon. The alignment anchors can be augmented or overwritten by user-defined anchors, such as previously identified compounds, characteristic mass or MS/MS identifications. Then, the candidates are paired by means of a bidirectional best-hits (BBH) criterion, which can compare different aspects of the candidates for similarity. Paired anchors are extended to *k*-cliques with configurable *k*, which help to determine the conservation or absence of signals across measurements, especially with respect to replicate groups. Tentative identification of peaks against a database using their mass spectra is possible using the MetaboliteDB module. This module provides access to mass-spectral databases in msp-compatible format, for example the Golm Metabolite Database or the NIST EI-MS database.

ChromA visualizes alignment results including paired anchors in birds-eye view or as a simultaneous overlay plot of the TIC. Additionally, absolute and relative differential charts are provided, which allow easy spotting of quantitative differences.

Peak tables are exported in CSV format, including peak apex positions, area under curve, peak intensity and possibly tentative database identifications. Additionally, information about the matched and aligned peak groups is saved in CSV format.

4. Frameworks for GCxGC-MS analysis

The automatic and routine analysis of comprehensive GCxGC-MS data is yet to be established. GCxGC-MS couples a second chromatographic column to the first one, thereby achieving a much higher peak capacity and thus a better separation of closely co-eluting analytes (Castillo, Mattila, Miettinen, Orešič & Hyötyläinen, 2011). Usually, for a one-hour run, the raw data file size exceeds a few Gigabytes. Quite a number of algorithms have been published on alignment of peaks in such four-dimensional (first column retention time, second column retention time, mass, and intensity values) data (Kim et al., 2011; Oh et al., 2008; Pierce et al., 2005; Vial et al., 2009; Zhang, 2010), however only a few methods are available for a more complete typical preprocessing workflow. A compact overview of the available frameworks, their licenses and programming languages is given in Table 3. Table 4 gives a more detailed feature matrix of these frameworks. The remainder of this section gives a concise overview of the frameworks Guineu (Castillo, Mattila, Miettinen, Orešič & Hyötyläinen, 2011) and ChromA4D (*Maltcms*, 2011).

Name	Version	Supported methods	Software license	Programming language
Guineu	0.8.2	GCxGC-MS (LC-MS)	GNU GPL v2	*JAVA* 6
Maltcms/ChromA4D	1.1	GCxGC-MS	GNU L-GPL v3	*JAVA* 6

Table 3. Feature comparison of Open Source software frameworks for GCxGC-MS based metabolomics

4.1 Guineu

Guineu is a recently published graphical user interface and application for the comparative analysis of GCxGC-MS data (Castillo, Mattila, Miettinen, Orešič & Hyötyläinen, 2011). It currently reads LECO ChromaTOF software's peak list output after smoothing, baseline correction, peak finding, deconvolution, database search and retention index (RI) calculation have been performed within ChromaTOF.

The peak lists are aligned pairwise using the score alignment algorithm, which requires user-defined retention time windows for both separation dimensions. Additionally, the one-dimensional retention index (RI) of each peak is used within the score calculation. Finally,

Feature (GCxGC-MS pipeline)	Guineu	ChromA4D
Data formats	G	A,H
Signal preprocessing	no	MA, MM, TH, CV
Peak detection	no	MAX-SRG
Multiple peak alignment	SCORE	CLIQUE
Visualization	STATS	STATS, TIC, EIC, TIC2D
DB search	GMD, PUBCHEM, KEGG	MSP (GMD)
Normalization	RP	RP, EV
Statistical evaluation	CV, FLT, TT, PCA, CDA, SP, ANOVA	FT

Table 4. Feature comparison of Open Source software frameworks for preprocessing of GCxGC-MS based metabolomics data. Key to abbreviations: **Data formats** A: NetCDF, G: ChromaTOF peak lists, H: CSV peak lists. **Signal preprocessing** MA: moving average, MM: moving median, TH: top-hat filter, CV: coefficient of variation threshold. **Peak detection** MAX-SRG: TIC local maxima, seeded region growing based on ms similarity. **Multiple peak alignment** SCORE: parallel iterative score-based, CLIQUE: progressive clique-based.**Visualization** (of unaligned and aligned data) TIC: plots of total ion chromatogram/peaks, EIC: plots of extracted ion chromatograms/peaks, SURF: surface plots of profile matrix (rt x m/z x I), STATS: visualization of statistical values. **DB search** GMD: Golm metabolite database webservice, PUBCHEM: pubchem database webservice, KEGG: kegg metabolite database, MSP: msp-format, compatible with AMDIS and GMD format. **Normalization** RP: reference peak area, EV: external value, e.g. dry weight. **Statistical evaluation** CV: coefficient of variation, FLT: fold-test, TT: groupwise t-test, PCA: principal components analysis, CDA: curvilinear distance analysis, SP: Sammon's projection, ANOVA: analysis of variance, FT: F-test, between group vs. within group variance.

a threshold for mass spectral similarity is needed in order to create *putative* peak groups. Additional peak lists are added incrementally to an already aligned *path*, based on the individual peaks' score against those peaks that are already contained within the path.

Guineu provides different filters to remove peaks by name, group occurrence count, or other features from the ChromaTOF peak table. In order to identify compound classes, the Golm metabolite database (GMD) substructure search is used. Peak areas can be extracted from ChromaTOF using the TIC, or using extracted, informative or unique masses. Peak area normalization is available relative to multiple user-defined standard compounds.

After peak list processing, Guineu produces an output table containing information for all aligned peaks, containing information on the original analyte annotation as given by ChromaTOF, peak areas, average retention times in both dimensions together with the average RI and further chemical information on the functional group and substructure prediction as given by the GMD. It is also possible to link the peak data to KEGG and Pubchem via the CAS annotation, if it is available for the reported analyte.

For statistical analysis of the peak data, Guineu provides fold change- and t-tests, principal component analysis (PCA), analysis of variance (ANOVA) and other methods.

Guineu's statistical analysis methods provide different plots of the data sets, e.g. for showing the principal components of variation within the data sets after analysis with PCA.

4.2 ChromA4D

For the comparison of comprehensive two-dimensional gas chromatography-mass spectrometry (GCxGC-MS) data, ChromA4D accepts NetCDF files as input. Additionally, the user needs to provide the total runtime on the second orthogonal column (modulation time) to calculate the second retention dimension information from the raw data files. For tentative metabolite identification, the location of a database can be given by the user. ChromA4D reports the located peaks, their respective integrated TIC areas, their best matching corresponding peaks in other chromatograms, as well as a tentative identification for each peak. Furthermore, all peaks are exported together with their mass spectra to MSP format, which allows for downstream processing and re-analysis with AMDIS and other tools. The exported MSP files may be used to define a custom database of reference spectra for subsequent analyses.

Peak areas are found by a modified seeded region growing algorithm. All local maxima of the TIC representation that exceed a threshold are selected as initial seeds. Then, the peak area is determined by using the distance of the seed mass spectrum to all neighbor mass spectra as a measure of the peak's coherence. The area is extended until the distance exceeds a given threshold. No information about the expected peak shape is needed. The peak integration is based on the sum of TICs of the peak area. An identification of the area's average or apex mass spectrum or the seed mass spectrum is again possible using the MetaboliteDB module.

To represent the similarities and differences between different chromatograms, bidirectional best hits are used to find co-occurring peaks. These are located by using a distance that exponentially penalizes differences in the first and second retention times of the peaks to be compared. To avoid a full computation of all pairs of peaks, only those peaks within a defined window of retention times based on the standard deviation of the exponential time penalty function are evaluated.

ChromA4D's visualizations represent aligned chromatograms as color overlay images, similar to those used in differential proteomics. This allows a direct visual comparison of signals present in one sample, but not present in another sample.

ChromA4D creates peak report tables in CSV format, which include peak apex positions in both chromatographic dimensions, area under curve, peak intensity and possibly tentative database identifications. Additionally, information about the matched and aligned peak groups is saved in CSV format.

5. Application examples

The following examples for GC-MS and GCxGC-MS are based on the Maltcms framework, using the ChromA and ChromA4D configurations described in the previous sections. In order to run them, the recent version of *Maltcms* needs to be downloaded and unzipped to a local folder on a computer. Additionally, Maltcms requires a *JAVA* runtime environment version 6 or newer to be installed. If these requirements are met, one needs to start a command prompt and change to the folder containing the unzipped Maltcms.

5.1 An example workflow for GC-MS

The experiment used to illustrate an example workflow for one-dimensional GC-MS consists of two samples of standard compounds, which contain mainly sugars, amino acids, other organic acids and nucleosides, measured after manual (MD) and after automatic derivatization (AD) with the derivatization protocol and substances given below. Group AD consists of a sample of n-alkanes standard and two replicates of mix1, namely mix1-1 and mix1-2. We will show how ChromA can be used to find and integrate peaks, as well as compare and align the peaks between the samples, and finally how the alignment results can be used for quality control.

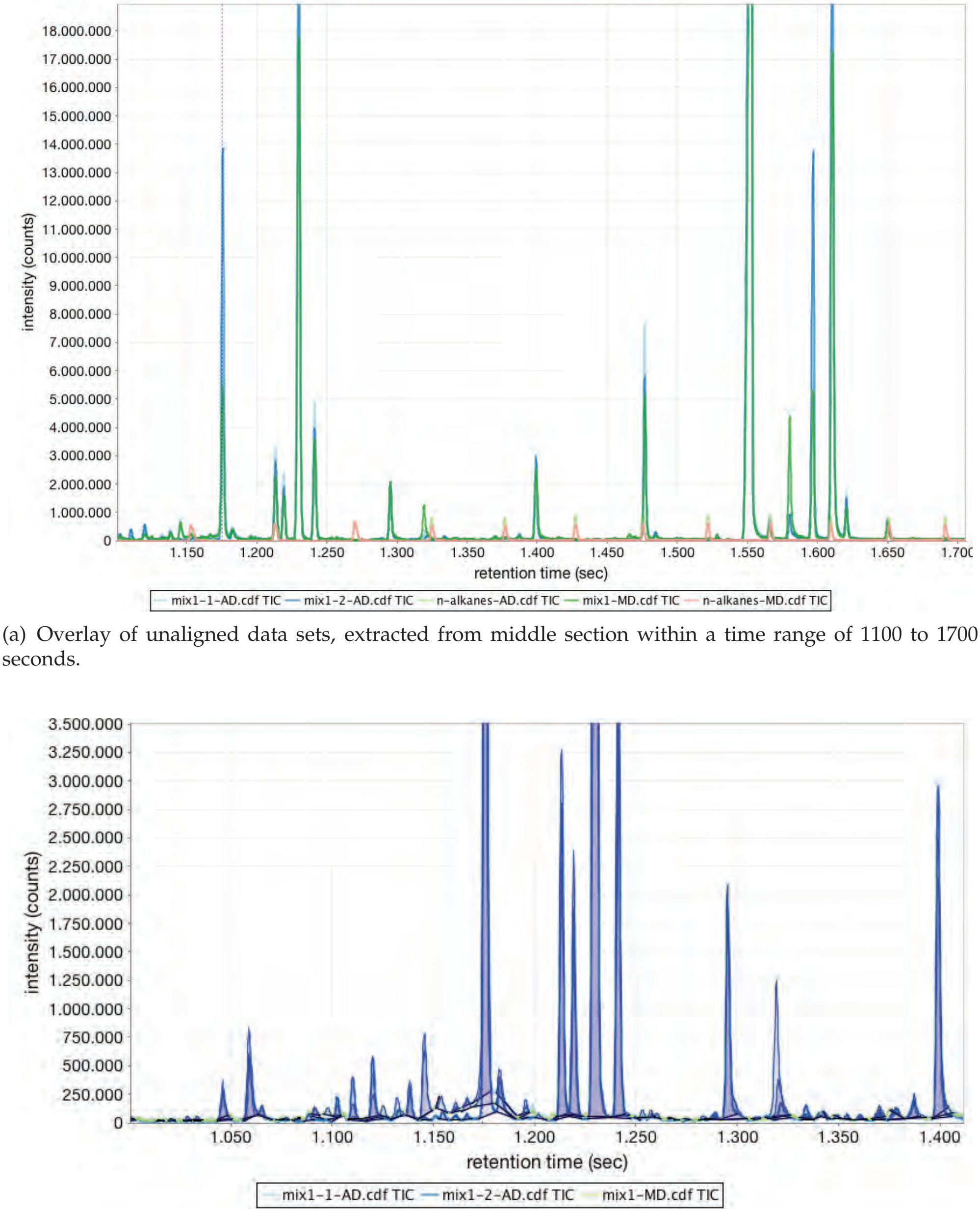
5.1.1 Sample preparation

20 μL of each sample were incubated with 60 μL methoxylamine hydrochloride (Sigma Aldrich) in pyridine (20 mg/ml) for 90 min at 60°C before 100 μL of N-Methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) (Macherey & Nagel) were added for 60 min at 37°C.

5.1.2 Acquisition and data processing

The samples were acquired on an Agilent GC 7890N with MSD 5975C triple axis detector. An Agilent HP5ms column with a length of 30 m, a diameter of 0.25 mm, and a film thickness of 0.25 μm (Agilent, Santa Clara CA, USA) was used for the gas-chromatographic separation, followed by a deactivated restriction capillary with 50 cm length and a diameter of 0.18 mm. Per sample, 1 μL was injected onto the column in pulsed splitless mode (30 psi for 2 min). The flow rate was set to 1.5 mL/min of Helium. The linear temperature ramp started at 50 °C for 2 min until it reached its maximum of 325 °C at a rate of 10 °C/min. The raw data were exported to NetCDF format using the Agilent ChemStation software v.B.04.01 (Agilent, Santa Clara CA, USA) with default parameters and without additional preprocessing applied.

A sample containing n-alkanes was measured as an external standard for manual (MD) and automatic derivatization (AD) in order to be able to later determine retention indices for



(a) Overlay of unaligned data sets, extracted from middle section within a time range of 1100 to 1700 seconds.

(b) Overlay with highlighted peak areas (without n-alkanes) after peak finding and integration. Zoomed in to provide more detail.

Fig. 2. TIC overlay plots of the raw GC-MS data sets.

the other samples. The acquired data were exported to ANDI-MS (NetCDF) format before ChromA was applied. The default ChromA pipeline `chroma.properties` was run from the unzipped Maltcms directory with the following command (issued on a single line of input):

```
> java -Xmx1G -jar maltcms.jar -i ../data/ -o ../output/ -f *.CDF  \
 -c cfg/chroma.properties
```

`-i` points to the directory containing the input data, `-o` points to the directory where output should be placed, `-f` can be a comma separated list of filenames or, as in this case, a wildcard expression, matching all files in the input directory having a file name ending with `.CDF`. The final argument indicated by `-c` is the path to the configuration file used for definition of the pipeline and its commands. An overlay of the raw TICs of the samples is depicted in Figure 2(a). The default ChromA pipeline configuration creates a profile matrix with nominal mass bin width. Then, the TIC peaks are located separately within each sample data file and are integrated (Figure 2(b)). The peak apex mass spectra are then used in the next step in order to build a multiple peak alignment between all peaks of all samples by finding large cliques, or clusters of peaks exhibiting similar retention time behaviour and having highly similar mass spectra. This coarse alignment could already be used to calculate a polynomial fit, correcting retention time shift for all peaks. However, the ChromA pipeline uses the peak clusters in order to constrain a dynamic time warping (DTW) alignment in the next step, which is calculated between all pairs of samples. The resulting distances are used to determine the reference sample with the lowest sum of distances to all remaining samples. Those are then aligned to the reference using the warp map obtained from the pairwise DTW calculations. The pairwise DTW distances can easily be used for a hierarchical cluster analysis. Similar samples should be grouped into the same cluster, while dissimilar samples should be grouped into different clusters. Figure 3 shows the results of applying a complete linkage clustering algorithm provided by *GNU R* to the pairwise distance matrix. It is clearly visible that the samples are grouped correctly, without incorporation of any external group assignment. Thus, this method can be used for quality control of multiple sample acquisitions, when the clustering results are compared against a pre-defined number of sample groups.

5.2 An example workflow for GCxGC-MS

The instructional samples presented in this section were preprocessed according to the protocol given by Doebbe et al. (2010). The description of the protocol has been adapted from that reference where necessary.

5.2.1 Sample preparation

The samples were incubated with 100 μl methoxylamine hydrochloride (Sigma Aldrich) in pyridine (20 mg/ml) for 90 min at 37°C while stirring. N-Methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) (Macherey & Nagel) was then added and incubated for another 30 min at 37°C with constant stirring.

5.2.2 Acquisition and data processing

The sample acquisition was performed on a LECO Pegasus 4D TOF-MS (LECO, St. Joseph, MI, USA). The Pegasus 4D system was equipped with an Agilent 6890 gas chromatograph

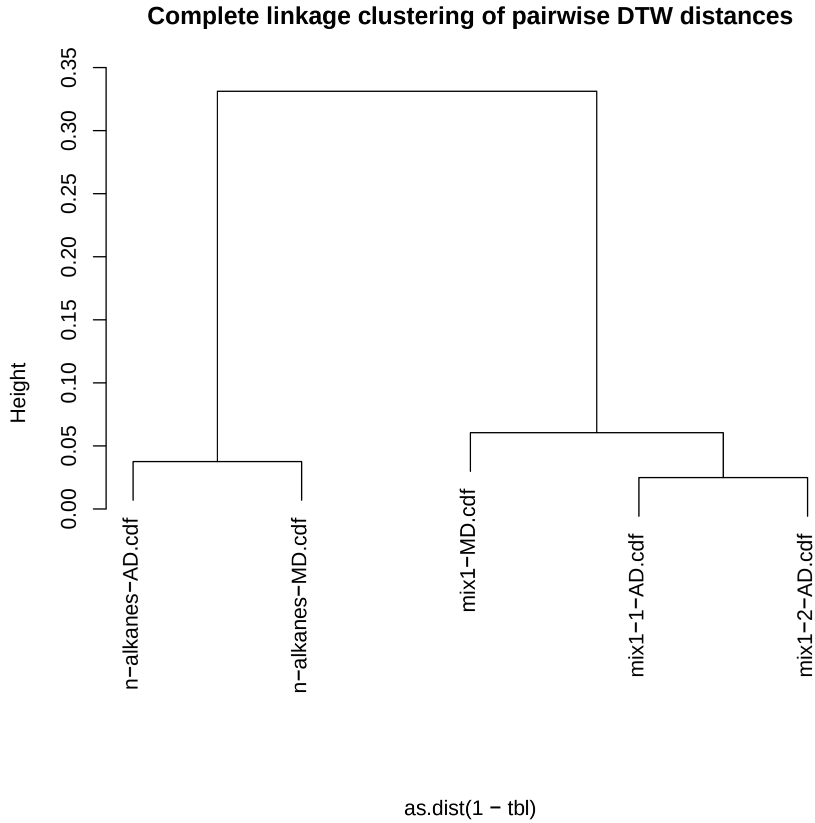
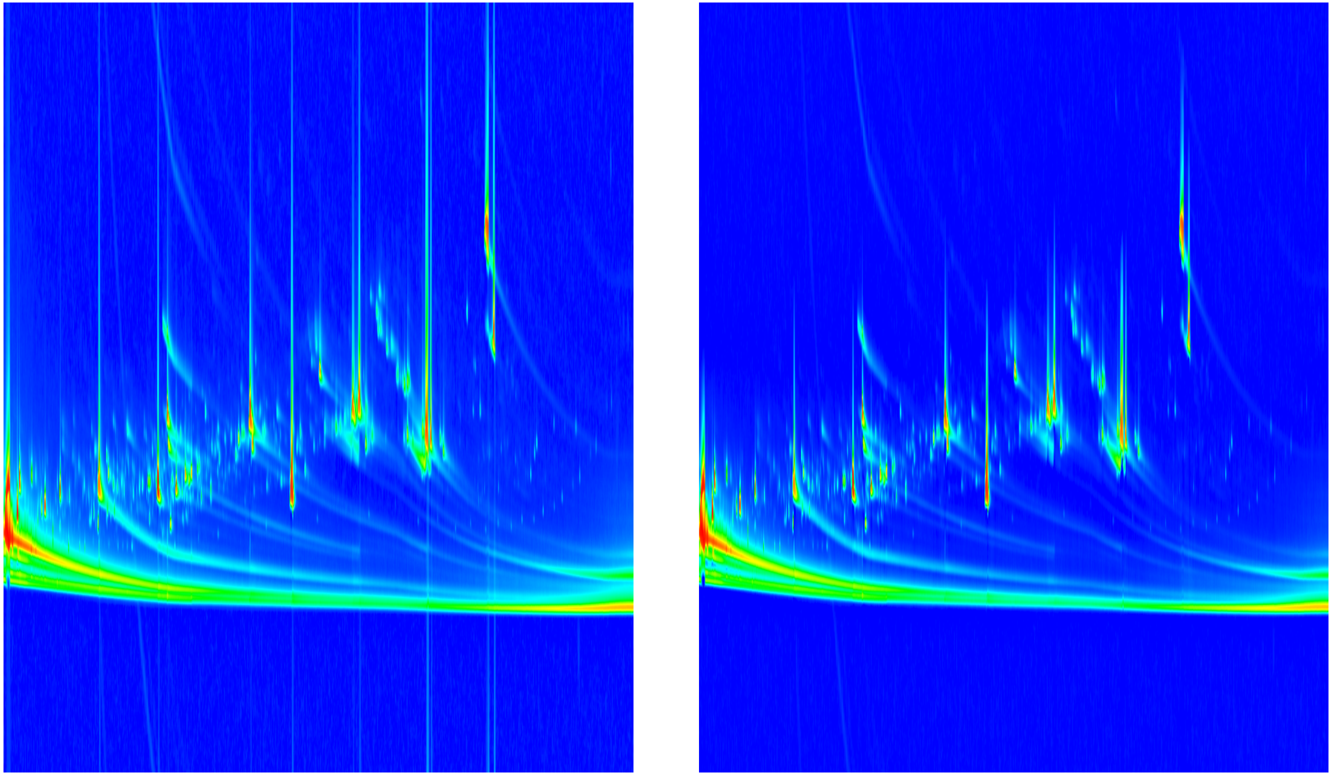


Fig. 3. Clustering of GC-MS samples based on pairwise DTW similarities transformed to distances. The samples are clearly separated into two clusters, one containing the n-alkane standard samples, the other one containing the mix1 samples.

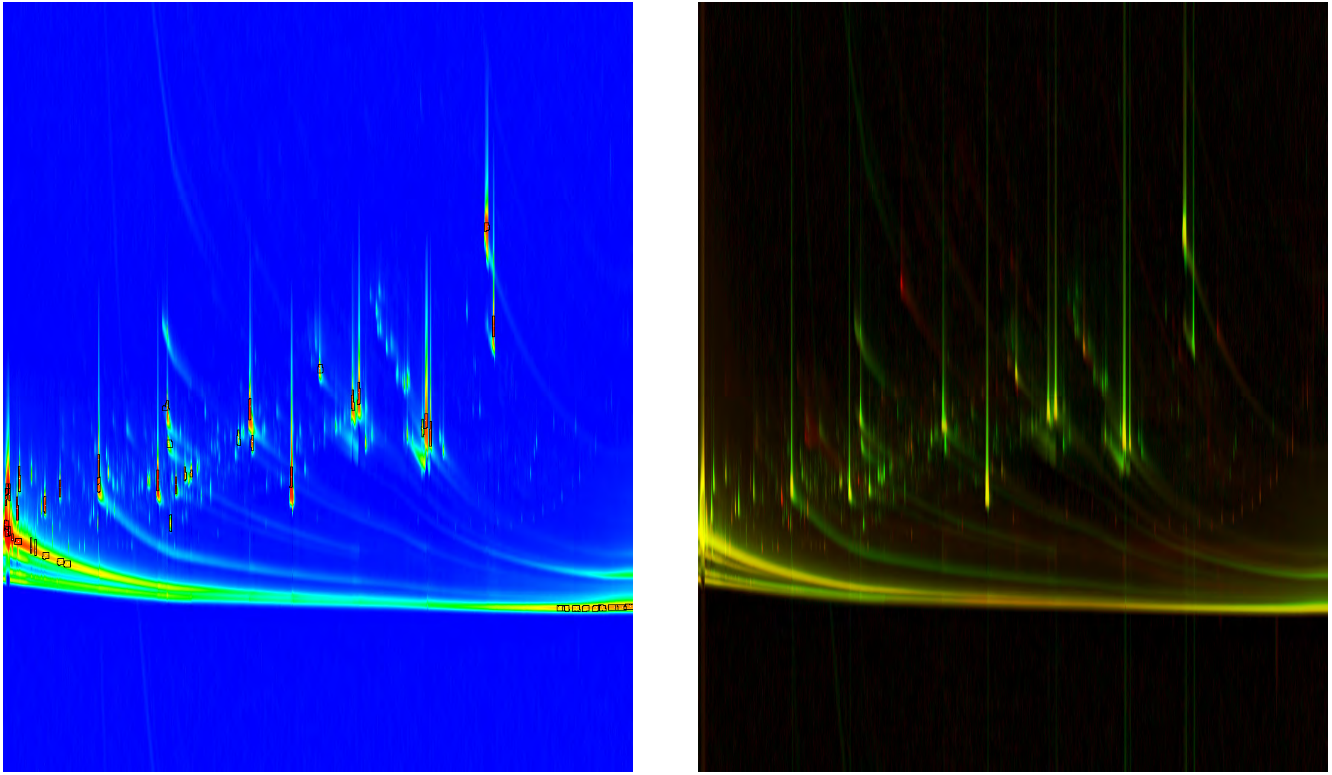
(Agilent, Santa Clara, CA, USA). The inlet temperature was set to 275°C. An Rtx-5ms (Restek, Bellefonte, PA, USA) capillary column was used with a length of 30 m, 0.25 mm diameter and 0.25 μm film thickness as the primary column. The secondary column was a BPX-50 (SGE, Ringwood, Victoria, Australia) capillary column with a length of 2 m, a diameter of 0.1 mm and 0.1 μm film thickness. The temperature program of the primary oven was set to the following conditions: 70°C for 2 min, 4°C/min to 180°C, 2°C/min to 230°C, 4°C/min to 325°C hold 3 min. This program resulted in a total runtime of about 70 min for each sample. The secondary oven was programmed with an offset of 15°C to the primary oven temperature. The thermal modulator was set 30°C relative to the primary oven and to a modulation time of 5 seconds with a hot pulse time of 0.4 seconds. The mass spectrometer ion source temperature was set to 200°C and the ionization was performed at -70eV. The detector voltage was set to 1600V and the stored mass range was 50-750 m/z with an acquisition rate of 200 spectra/second.



(a) 2D-TIC plot before filters were applied. Long tailing peaks are visible within the vertical dimension. Additionally, high frequency noise is present in the raw exported data, which is barely visible at this resolution.

(b) 2D-TIC plot after application of a moving median filter with window size 3 for smoothing of high-frequency noise and successive application of a top hat filter with a window size of 301 for baseline removal in order to reduce false positive peak finding results.

Fig. 4. Visualizations of Standard-Mix1-1 before and after signal filtering with the ChromA4D processing pipeline.



(a) 2D-TIC plot of Standard-Mix1-1 after peak finding and integration with seeded region growing based on the cosine mass spectral similarity with a fusion threshold of 0.99. Peak areas were limited to contain at most 100 points.

(b) Differential plot of the two Standard-Mix1 samples after DTW alignment based on vertical TIC slices. Yellow color indicates similar amounts of total ion intensity in both samples. Green shows a surplus in Standard-Mix1-1, while red shows a surplus in Standard-Mix1-2.

Fig. 5. Visualizations of Standard-Mix1-1 after peak finding and of Standard-Mix1-1 and Standard-Mix1-2 after alignment with DTW.

The raw acquired samples in LECO's proprietary ELU format were exported to NetCDF format using the LECO ChromaTOF® software v.4.22 (LECO, St. Joseph, MI, USA). Initial attempts to export the full, raw data failed with a crash beyond a NetCDF file size of 4GBytes. Thus, we resampled the data with ChromaTOF to 100 Hz (resampling factor 2) and exported with automatic signal smoothing and baseline offset correction value of 1 which resulted in file sizes around 3GBytes per sample. The samples presented in this section are named "Standard-Mix1-1" and "Standard-Mix1-2" and were measured on different days (Nov. 29th, 2008 and Dec. 12th, 2008).

The default ChromA4D pipeline for peak finding was called from within the unzipped Maltcms directory (issued on a single line of input):

```
> java -Xmx2G -jar maltcms.jar -i ../data/ -o ../output/ \
 -f *.cdf -c cfg/4Dpeakfinding.properties
```

The pipeline first preprocesses the data by applying a median filter followed by a top hat filter in order to remove high- and low-frequency noise contributions (Figures 4(a) and 4(b)). ChromA4D then uses a variant of seeded region growing in order to extend peak seeds, which are found as local maxima of the 2D-TIC. These initial seeds are then extended until the mass spectral similarity of the seed and the next evaluated candidate drops below a user-defined threshold, or until the peak area reaches its maximum, pre-defined size (Figure 5(a)). After peak area integration, the pipeline clusters peaks between samples based on their mass spectral similarity and retention time behaviour in both dimensions to form peak cliques (not shown) as multiple peak alignments, which are then exported into CSV format for further downstream processing. Another possible application shown in Figure 5(b) is the visualization of pairwise GCxGC-MS alignments using DTW on the vertical 2D-TIC slices, which can be useful for qualitative comparisons.

6. Summary and outlook

The present state of Open Source frameworks for metabolomics is very diverse. A number of tools have seen steady development and improvement over the last years, such as XCMS, MZmine, and PyMS, while others are still being developed, such as mzMatch, Guineu, and Maltcms. There is currently no framework available that covers every aspect of metabolomics data preprocessing. Most of the frameworks concentrate on one or a few analytical technologies with the largest distinction being between GC-MS and LC-MS. GCxGC-MS raw data processing is currently only handled by Maltcms' ChromA4D pipeline, while Guineu processes peak lists exported from LECO's ChromaTOF software and offers statistical methods for sample comparison together with a user-friendly graphical interface.

We showed two instructive examples on setting up and running the basic processing pipelines ChromA and ChromA4D for GC-MS and GCxGC-MS raw data. The general structure of these pipelines would be slightly different for each of the Open Source frameworks presented in this chapter, however, the basic concepts behind the processing steps are the same for all tools. Since metabolomics is an evolving field of research, no framework captures all possible use-cases, but it will be interesting to see which frameworks will be flexible and extendable enough to be adapted to new requirements in the near future.

In order to combine experiments from multiple "omics" experiments, another level of abstraction on top of local or web-service based tools for data processing, fusion, and integration of metabolomics experiments is a necessary future requirement. Generic workflow systems like Taverna (Hull et al., 2006) or Conveyor (Linke et al., 2011) offer integration of such resources, augmented with graphical editors for *point-and-click* user interaction. However, due to their generic nature these systems are far away from being as user-friendly as applications designed for a specific data analysis task and require some expert knowledge when assembling task-specific processing graphs.

One point that requires further attention is the definition and controlled evolution of peak data formats for metabolomics, along with other formats for easier exchange of secondary data between applications and frameworks. A first step in this direction has been taken by Scheltema et al. (2011) by defining the PeakML format. However, it is important that such formats are curated and evolved, possibly by a larger non-profit organization like the HUPO within its proteomics standards initiative *HUPO PSI*. Primary data is already acessible in a variety of different, defined formats, the most recent addition being mzML (Martens et al., 2010) which is curated by the PSI. Such standardization attempts can however only be successful and gain the required momentum if also the manufacturers of analytical machinery support the formats with their proprietary software within a short time frame after the specification and see a benefit in offering such functionality due to the expressed demand of scientists working in the field as in case of NetCDF, mzData, or mzML.

7. Acknowledgements

We would like to thank Manuela Meyer at the Center for Biotechnology (CeBiTec), Bielefeld University, Germany, for the preparation of the GC-MS samples used in this work. We furthermore thank Denise Schöfbeck and Rainer Schumacher, Center for Analytical Chemistry, IFA Tulln, University of Natural Resources and Life Sciences, Vienna, Austria, for measuring those samples and for kindly providing us with the datasets. Furthermore, we would like to express our gratitude to Anja Döbbe and Olaf Kruse, Algae BioTech Group at the CeBiTec, for providing the GCxGC-MS samples and datasets. We finally thank our students Mathias Wilhelm for his work on ChromA4D and Kai-Bernd Stadermann for his work on the RMI-based remote execution framework.

8. References

Åberg, K., Alm, E. & Torgrip, R. (2009). The correspondence problem for metabonomics datasets, *Analytical and Bioanalytical Chemistry* 394(1): 151–162.

Ahmad, I., Suits, F., Hoekman, B., Swertz, M. A., Byelas, H., Dijkstra, M., Hooft, R., Katsubo, D., van Breukelen, B., Bischoff, R. & Horvatovich, P. (2011). A high-throughput processing service for retention time alignment of complex proteomics and metabolomics LC-MS data, *Bioinformatics* 27(8): 1176–1178.

Babushok, V. I., Linstrom, P. J., Reed, J. J., Zenkevich, I. G., Brown, R. L., Mallard, W. G. & Stein, S. E. (2007). Development of a database of gas chromatographic retention properties of organic compounds., *Journal of Chromatography A* 1157(1-2): 414–421.

Berk, M., Ebbels, T. & Montana, G. (2011). A statistical framework for biomarker discovery in metabolomic time course data, *Bioinformatics* 27(14): 1979–1985.

Bertsch, A., Hildebrandt, A., Hussong, R. & Zerck, A. (2008). OpenMS - An open-source software framework for mass spectrometry., *BMC Bioinformatics* 9(1): 163.

Biller, J. E. & Biemann, K. (1974). Reconstructed Mass Spectra, A Novel Approach for the Utilization of Gas Chromatograph—Mass Spectrometer Data, *Analytical Letters* 7(7): 515–528.

Callaghan, S., De Souza, D., Tull, D., Roessner, U., Bacic, A., McConville, M. & Likić, V. (2010). Application and comparative study of PyMS Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data, *2nd Australasian Symposium on Metabolomics*, Melbourne 2010.

Carroll, A. J., Badger, M. R. & Harvey Millar, A. (2010). The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets, *BMC Bioinformatics* 11(1): 376.

Castillo, S., Gopalacharyulu, P., Yetukuri, L. & Orešič, M. (2011). Algorithms and tools for the preprocessing of LC–MS metabolomics data, *Chemometrics and Intelligent Laboratory Systems* 108(1): 23–32.

Castillo, S., Mattila, I., Miettinen, J., Orešič, M. & Hyötyläinen, T. (2011). Data Analysis Tool for Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry, *Analytical Chemistry* 83(8): 3058–3067.

Chae, M., Reis, R. & Thaden, J. J. (2008). An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks, *BMC Bioinformatics* 9(Suppl 9): S15.

Christin, C., Hoefsloot, H. C. J., Smilde, A. K., Suits, F., Bischoff, R. & Horvatovich, P. L. (2010). Time Alignment Algorithms Based on Selected Mass Traces for Complex LC-MS Data, *Journal of Proteome Research* 9(3): 1483–1495.

Clifford, D., Stone, G., Montoliu, I., Rezzi, S., Martin, F.-P., Guy, P., Bruce, S. & Kochhar, S. (2009). Alignment Using Variable Penalty Dynamic Time Warping, *Analytical Chemistry* 81(3): 1000–1007.

Deutsch, E. (2008). mzML: a single, unifying data format for mass spectrometer output., *Proteomics* 8(14): 2776–2777.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I. & Aebersold, R. (2010). A guided tour of the Trans-Proteomic Pipeline, *Proteomics* 10(6): 1150–1159.

Doebbe, A., Keck, M., Russa, M. L., Mussgnug, J. H., Hankamer, B., Tekce, E., Niehaus, K. & Kruse, O. (2010). The Interplay of Proton, Electron, and Metabolite Supply for Photosynthetic H2 Production in Chlamydomonas reinhardtii, *Journal of Biological Chemistry* 285(39): 30247–30260.

Du, P., Kibbe, W. A. & Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, *Bioinformatics* 22(17): 2059–2065.

Foster, I. T. (2005). Globus Toolkit Version 4: Software for Service-Oriented Systems., *in* H. Jin, D. A. Reed & W. Jiang (eds), *IFIP International Conference on Network and Parallel Computing*, Springer, pp. 2–13.

Fredriksson, M. J., Petersson, P., Axelsson, B.-O. & Bylund, D. (2009). An automatic peak finding method for LC-MS data using Gaussian second derivative filtering., *Journal of Separation Science* 32(22): 3906–3918.

GNU R (2011).
URL: *http://www.r-project.org/*

Hoffmann, N. & Stoye, J. (2009). ChromA: signal-based retention time alignment for chromatography-mass spectrometry data, *Bioinformatics* 25(16): 2080–2081.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P. & Oinn, T. (2006). Taverna: a tool for building and running workflows of services, *Nucleic Acids Research* 34(suppl 2): W729–W732.

Hummel, J., Selbig, J., Walther, D. & Kopka, J. (2007). The Golm Metabolome Database: a database for GC-MS based metabolite profiling, *in* J. Nielsen & M. Jewett (eds), *Metabolomics*, Springer Berlin / Heidelberg, pp. 75–95.

HUPO PSI (2011).
URL: *http://www.psidev.info/*

Isaac, A., Lee, L., Keen, W., Erwin, T., Wang, Q., De Souza, D., Roessner, U., Pyke, J., Kotagiri, R., Wettenhall, R., McConville, M., Bacic, A. & Likić, V. (2009). PyMS: A Python toolkit for processing of gas chromatography-mass spectrometry data, *Bioinformatics Australia Conference*, Melbourne 2009.

JAVA (2011).
URL: *http://www.java.com/en/*

Kankainen, M., Gopalacharyulu, P., Holm, L. & Orešič, M. (2011). MPEA–metabolite pathway enrichment analysis, *Bioinformatics* 27(13): 1878–1879.

Kastenmüller, G., Römisch-Margl, W., Wägele, B., Altmaier, E. & Suhre, K. (2011). metaP-Server: A Web-Based Metabolomics Data Analysis Tool, *Journal of Biomedicine and Biotechnology* 2011: 1–8.

Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development., *Bioinformatics* 24(21): 2534–2536.

Kim, S., Fang, A., Wang, B., Jeong, J. & Zhang, X. (2011). An Optimal Peak Alignment For Comprehensive Two-Dimensional Gas Chromatography Mass Spectrometry Using Mixture Similarity Measure, *Bioinformatics* 27(12): 1660–1666.

Krebs, M. D., Tingley, R. D., Zeskind, J. E., Holmboe, M. E., Kang, J.-M. & Davis, C. E. (2006). Alignment of gas chromatography-mass spectrometry data by landmark selection from complex chemical mixtures, *Chemometrics and Intelligent Laboratory Systems* 81(1): 74–81.

Lange, E., Gropl, C., Schulz-Trieglaff, O., Huber, C. & Reinert, K. (2007). A geometric approach for the alignment of liquid chromatography mass spectrometry data, *Bioinformatics* 23(13): i273–i281.

Linke, B., Giegerich, R. & Goesmann, A. (2011). Conveyor: a workflow engine for bioinformatic analyses, *Bioinformatics* 27(7): 903–911.

Lommen, A. (2009). MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing, *Analytical Chemistry* 81(8): 3079–3086.

Maltcms (2011).
URL: *http://maltcms.sourceforge.net*

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Rompp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P. A. & Deutsch, E. W. (2010). mzML–a Community Standard for Mass Spectrometry Data, *Molecular and Cellular Proteomics* 10(1): R110.000133–R110.000133.

Matthews, L. (2000). ASTM Protocols for Analytical Data Interchange, 5(5): 60–61.

Miura, D., Tsuji, Y., Takahashi, K., Wariishi, H. & Saito, K. (2010). A strategy for the determination of the elemental composition by fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios., *Technical Report 13*, Innovation Center for Medical Redox Navigation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 12-8582, Japan.

Neumann, S. & Böcker, S. (2010). Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules, *Analytical and Bioanalytical Chemistry* 398(7-8): 2779–2788.

Neuweger, H., Albaum, S. P., Niehaus, K., Stoye, J. & Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data, *Bioinformatics* 24(23): 2726–2732.

Neuweger, H., Persicke, M., Albaum, S. P., Bekel, T., Dondrup, M., Hüser, A. T., Winnebald, J., Schneider, J., Kalinowski, J. & Goesmann, A. (2009). Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example., *BMC Systems Biology* 3: 82.

Oh, C., Huang, X., Regnier, F. E., Buck, C. & Zhang, X. (2008). Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm, *Journal of Chromatography A* 1179(2): 205–215.

Oliver, S. G., Paton, N. W. & Taylor, C. F. (2004). A common open representation of mass spectrometry data and its application to proteomics research, *Nature Biotechnology* 22(11): 1459–1466. 10.1038/nbt1031.

Orchard, S., Hermjakob, H., Taylor, C. F., Potthast, F., Jones, P., Zhu, W., Julian, R. K. & Apweiler, R. (2005). Second proteomics standards initiative spring workshop., *Expert review of proteomics*, EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. pp. 287–289.

Pierce, K. M., Hoggard, J. C., Hope, J. L., Rainey, P. M., Hoofnagle, A. N., Jack, R. M., Wright, B. W. & Synovec, R. E. (2006). Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts, *Analytical Chemistry* 78(14): 5068–5075.

Pierce, K. M., Wood, L. F., Wright, B. W. & Synovec, R. E. (2005). A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data, *Analytical Chemistry* 77(23): 7735–7743.

Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, *BMC Bioinformatics* 11(1): 395.

Prince, J. & Marcotte, E. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping, *Analytical Chemistry* 78(17): 6140–6152.

Python (2008).
URL: *http://www.python.org/download/releases/2.5.2/*

Rew, R. & Davis, G. (1990). NetCDF: an interface for scientific data access, *Computer Graphics and Applications, IEEE* 10(4): 76–82.

Robinson, M. D., De Souza, D. P., Saunders, E. C., Mcconville, M. J., Speed, T. P. & Likić, V. A. (2007). A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments, *BMC Bioinformatics* 8(1): 419.

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures., *Analytical Chemistry* 36(8): 1627–1639.

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A. & Breitling, R. (2011). PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis, *Analytical Chemistry* 83(7): 2786–2793.

SciPy (2011).
URL: *http://www.scipy.org/*

Smilde, A. K. & Horvatovich, P. L. (2008). Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms, *Analytical Chemistry* 80(18): 7012–7021.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, *Analytical Chemistry* 78(3): 779–787.

Stein, S. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, *Journal of the American Society for Mass Spectrometry* 10(8): 770–781.

Tautenhahn, R., Böttcher, C. & Neumann, S. (2007). Annotation of LC/ESI-MS Mass Signals, *in* S. Hochreiter & R. Wagner (eds), *Bioinformatics Research and Development*, Springer Berlin / Heidelberg, pp. 371–380. 10.1007/978-3-540-71233-6_29.

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

Tohge, T. & Fernie, A. R. (2009). Web-based resources for mass-spectrometry-based metabolomics: A user's guide, *Phytochemistry* 70(4): 450–456.

Vial, J., Noçairi, H., Sassiat, P., Mallipatu, S., Cognon, G., Thiébaut, D., Teillet, B. & Rutledge, D. N. (2009). Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts, *Journal of Chromatography A* 1216(14): 2866–2872.

Wang, S. Y., Ho, T. J., Kuo, C. H. & Tseng, Y. J. (2010). Chromaligner: a web server for chromatogram alignment, *Bioinformatics* 26(18): 2338–2339.

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T. & Trygg, J. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models, *Analytical Chemistry* 80(1): 115–122.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J. & Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome, *Nucleic Acids Research* 37(Database): D603–D610.

Xia, J., Sinelnikov, I. V. & Wishart, D. S. (2011). MetATT: a web-based metabolomics tool for analyzing time-series and two-factor data sets, *Bioinformatics* 27(17): 2455–2456.

Xia, J. & Wishart, D. S. (2010a). MetPA: a web-based metabolomics tool for pathway analysis and visualization, *Bioinformatics* 26(18): 2342–2344.

Xia, J. & Wishart, D. S. (2010b). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data, *Nucleic Acids Research* 38(suppl 2): W71–W77.

Xia, J. & Wishart, D. S. (2011). Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst, *Nature Protocols* 6(6): 743–760.

XML (2008).
URL: *http://www.w3.org/TR/REC-xml/*

Zhang, X. (2010). DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics, *Analytical Chemistry* 82(12): 5069–5081.